DOCUMENT RESUME

ED 428 574 FL 025 773

AUTHOR	Coombe,. Christine A., Ed.
TITLE	Current Trends in English Language Testing. Conference Proceedings for CTELT 1997 and 1998, Vol. 1.
INSTITUTION	TESOL Arabia.
PUB DATE	1998-10-00
NOTE	194p.
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Cloze Procedure; College Second Language Programs; Comparative Analysis; Computer Assisted Testing; Copyrights; Distance Education; Educational Trends; *English for Academic Purposes; *English (Second Language); Foreign Countries; Higher Education; Internet; *Language Proficiency; *Language Tests; Learning Motivation; Listening Comprehension Tests; Oral Language; Program Descriptions; Reading Tests; Second Language Instruction; Student Developed Materials; Teacher Role; Test Construction; Test Use; *Testing Programs; Trend Analysis; Verbal Tests; Vocabulary
IDENTIFIERS	International English Language Testing System; Oman; Preliminary Test of English as a Foreign Language; Saudi Arabia; Test of English as a Foreign Language; United Arab Emirates

ABSTRACT

Papers from the 1997 and 1998 Current Trends in English Language Testing (CTELT) conferences include: "Computer-Based Language Testing: The Call of the Internet" (G. Fulcher); "Uses of the PET (Preliminary English Test) at Sultan Qaboos University" (R. Taylor); "Issues in Foreign and Second Language Academic Listening Assessment" (C. Coombe, J. Kinney, C. Canning); "Saudi Development and Training's Five Star Proficiency Test Project" (J. Pollard); "Test Writing for UAE Distance Learning Students" (L. Barlow, C. Canning); "Student Created Tests as Motivation to Learning" (P. Cozens); "Student Errors--To Use or Not To Use? That Is the Question" (J. Eadie, A. Abdel-Fattah, H. Guefrachi); "The Impact of High-Stakes Testing on Teaching and Learning" (D. Wall); "Myths of Testing and the Icons of PET, TOEFL and IELTS" (G. Tennent); "Copyright Infringement: What Are the Legal Rights of Educators as Test Writers?" (C. Canning); "C-Testing: A Theory Yet To Be Proved" (N. McBeath); "Why Teacher, That's Your Job" (P. Cozens); "Computer Generated Visuals and Reading Texts" (C. Canning, L. Barlow, C. Kawar); and "How Different Examination Boards Present Vocabulary" (K. Aldred). (MSE)

TESOL ARABIA
Testing Special Interest Group

CURRENT TRENDS IN ENGLISH LANGUAGE TESTING

Conference Proceedings for CTELT 1997 and 1998

Vol. I

Editor:
Christine A. Coombe

Al-Ain, United Arab Emirates

October 1998

FROM THE EDITOR

This volume presents a collection of papers delivered at the 1997 and 1998 Current Trends in English Language Testing Conferences. The papers address a variety of issues of direct relevance to the EFL/ESL teacher and tester.

I feel that we have an excellent representation of the comprehensive range of presentation content and style which make the CTELT Conference so valuable for its participants.

Several people have assisted at various stages in the production of this volume and this project could not have been completed without their support. To the following people I express my sincere thanks and gratitude: The United Arab Emirates University General Requirements Unit, for their continued financial assistance and support of CTELT; the TESOL Arabia Executive Council for their financial assistance and R.S. Pillai for his careful typing and formatting of this manuscript. Special thanks go to the sixteen authors who contributed papers and made this volume possible. Their papers represent an important contribution to the teaching and testing of EFL in the Arab World.

1997 CTELT CONFERENCE

CONTENTS

1998 CTELT CONFERENCE

CONTENTS

Computer-Based Language Testing: The Call of the Internet

GLENN FULCHER
University of Surrey

Abstract

This paper presents a brief overview of concerns in computer based test delivery. It does not attempt to be exhaustive in its review, but to select and discuss key issues and concepts, exemplified by representative examples. Against the background of the general development of computer based tests (CBTs) and computer adaptive tests (CATs), the advantages and disadvantages of using the internet for test delivery are explored. Some aspects of internet delivery are highlighted as problematic, either from a technical or measurement perspective, but it is argued that with continued research, the use of this medium will introduce a revolution in test delivery and (at a later stage) test design.

1. Introduction

Since the first book on computerised language testing was produced in 1970 (Holtzman, 1970), the number of software products on the market has grown almost as quickly as the number of journal articles on the subject. The two concerns of the discussion in the literature is, for the most part, related to what technology can offer the test designer or developer, and the equivalence of paper-and-pencil and computer delivered tests. The former focus reflects primarily an interest in *technology as technology*, and the way in which it can speed up the work of language test design and development. This is not surprising, given the rapid changes in computer technology, and the ever-increasing range and availability of software. Only 9 years ago, Bunderson, Inouye and Olsen, (1989: 367–268) wrote:

> "The computer revolution has been marked by the growth in power and sophistication of computing resources. The computing power of yesterday's mainframes is routinely surpassed by today's supermicros. Yesterday's ENIAC computer, which filled an entire room, was less powerful than the current generation of microcomputers, which fit on a desktop."

It is only when placing this next to another quotation from the same paper (Bunderson et al. 1989: 371) that we realise how fast technology is in fact changing:

> "The memory capacity of most modern delivery systems is evolving rapidly. Most microcomputer workstations now have 1/2 to 2 megabytes of random access memory. Future workstations will use even larger amounts of random access memory. The early, expensive, mass-storage devices are being replaced by inexpensive, high-density, magnetic and opto-electronic devices. Hard-disk storage exceeding 100 megabytes per workstation is becoming more common."

These quotations make it clear that any discussion of hardware, memory size, and processing capacity will inevitably be out of date within a few years, if not months, of being written. We must simply assume that the constant development, availability and use of larger and faster computers will be the norm for the foreseeable future.

The latter focus on test equivalence, reflects the fact that many computer based tests (CBTs) grow out of paper-and-pencil tests which already exist. The paper-and-pencil version usually pre-dates the CBT by some years, and subsequent forms constructed and monitored to be interpretable in the light of established norms. For this reason, it is critical that equivalence be demonstrated across the delivery medium.

2. Background: The Use of Computers In Testing

2.1 The Lure of Technology

In a recent review of the role of technology in language testing, Burstein, Frase, Ginther and Grant (1996), isolated eight areas in which computers are now used in language testing. These may be summarised as:

- Test design: the exchange of written and graphic materials between test designers who may be working in different locations.

- Test construction: including item trials, the main function of the computer is envisaged to be the exchange of written and graphic materials, as well as items, between those involved in item writing and revision. In this phase, it is also important to have the memory capacity to store information on items, including any material related to the prompts, including audio and, we might now add, video.

- Item tryout: this is probably what we would normally refer to as pre-testing, in which items are delivered in their near-final format, responses are stored on the computer, and item level statistics calculated and stored in a database with the test items.

- Test item delivery: the delivery of actual tests from databases, including the collection and storage of responses. Appropriate technologies for test taker identification should also be considered in this phase of the process.

- Item management: storing and updating item information.

- Item scoring and transforming item responses into test scores.

- Item analysis and interpretation: relating the score to some general interpretative scheme for the score.

- Score reporting: delivering scores and related information.

In this review, there is a clear focus on available technology, and how technology can be used to make the construction and delivery of more traditional tests easier and quicker. Although the authors lament the fact that language testing has not made use of the multimedia capabilities of the computer in the way that instructional programs have done (ibid., 245), this review remains essentially a consideration of the technical aspects of the use of computers in language testing.

Although any test that is currently delivered using paper-and-pencil can also be delivered by computer, the most important development of the last decade has been computer adaptive testing, in which the computer branches to certain sub-tests (branching routines) or selects the next test item (adaptive routines) depending upon the response pattern of the individual test taker. The latter approach has been made possible by the extensive use of Item Response Theory, and the development of algorithms that drive the test program. Bunderson et al. (1989: 381) describe this as the second generation of computerized testing. One of the most advanced CAT systems is MicroCAT (ASC, 1994). Consisting of a number of subsystems (test development, delivery, and test evaluation), MicroCAT is able to deliver adaptive tests on stand-alone computers or local area networks using fixed branching adaptive strategies and item response theory models. Item calibration can even be conducted on-line, so that adaptivity continually improves in accuracy.

2.2 Computer Adaptive Testing and Beyond

Computer adaptive tests constructed and delivered with systems like MicroCAT have a number of major advantages. Firstly, all items and item level information is contained in an item bank on the local machine or network. As the selection of the next item or sub-test is dependent upon the responses of the test-taker, no test-taker takes exactly the same test as any other, assuming that the item bank is reasonably large. Secondly, the number of items that the test-taker is required to attempt is reduced, as the computer will terminate the test once an assessment of the test-taker's ability level has been estimated within pre-set error parameters. Not only does this save time and resources in terms of test delivery and the amount of time needed to administer tests, it also provides instant results and reporting.

Of particular importance in the emergence of CATs is the use of item banking facilities. In principle, the use of calibrated items from a bank allows clearer interpretation of score meaning. Achieving this should not, however, be seen as an easy process. Calibrating test items to a scale is ultimately norm-referencing, and establishing criterion-referenced meaning at various points on the scale requires the careful consideration of establishing standards. The latter is judgemental, and often depends on the values of the institution using the tests. The second problem associated with the item bank is one of sampling. It is frequently assumed that items are written to adequately reflect the domain of interest, and the implementation of a CAT means that only a small proportion of the items are selected for any individual test. It is therefore appropriate that the issue of content validity of a CAT is problematized, so that test developers consider whether, and to what degree, the CAT should be forced to include a representative sample of items in the test, even if they are not needed in order to place a student on the ability scale.

At the moment, a CAT could not be implemented on the internet. Although the internet, and the World Wide Web (more conventionally only the "web" or "WWW") in particular, is a global information distribution network that would easily allow the delivery of tests anywhere in the world, its potential has not been realised. The interactivity that is currently available on the web is provided by programs stored on the server (a computer that stores information freely available on the internet) written in PERL script (Practical Extraction and Report Language) or Java. At the present time, programs to run computer adaptive tests do not exist in either of these computer languages. It seems to me that there are two basic reasons for this. Firstly, large scale high-stakes test delivery over the web is unlikely until many of the security problems associated with the transfer of information over the internet have been solved, and secondly, the speed of change in computer programming for the web, including the rapid development and expansion of hypertext mark up language (html) in which web pages are written, makes it quite possible that a development project of this nature would be out of date before it had made very much progress. Nevertheless, there are plans to develop a web based CAT for the delivery of the very low stakes European System for Diagnostic Language Testing (DIALANG Project).[1]

The third and fourth generation of computerized testing, as described by Bunderson et al. (1989), whilst visionary, are nevertheless still some way in the future. The third generation of computerised testing is the continuous assessment of learning and the projection of learning trajectories from the current ability level of the student to another ability level at some point in the future. However, the assumption is that it is possible to calculate trajectories in language learning in a meaningful way. Given what we currently know about language acquisition, including U-shaped and discontinuous learning (Larsen-Freeman and Long, 1991: 105–107; Perkins et al., 1996), and taking into account the multitude of variables that affect language learning, it seems unlikely that the progress of individuals can be meaningfully predicted very far into the future. Research in the development of multidimensional Item Response Theory models (Mislevy and Wilson, 1992) may allow the tracking of a number of variables that increases the reliability of future prediction, but even a multidimensional CAT is likely to be based on the assumption that learning is linear.

This makes it more unlikely that we will see the development of the predicted fourth generation of assessment (Bunderson et al. 1989: 398–402), in which artificial intelligence will be brought to bear on continuous measurement in order to provide advice on learning patterns and content for the learner, related to the current estimated stage of learning. This fourth generation of tests would have all the properties of the third generation, but would be linked to expert second language acquisition systems. The field of second language acquisition is currently not able to provide such an expert system, and even if a model of language development could be generally

[1] The Dialang Project is the development of a new diagnostic language testing system for the official languages of the European Union, plus a small number of other languages. A description of the project is available at:
http://www.jyu.fi/DIALANG/index.htm

agreed, calibrating the test to the theoretical model would be a major project that would occupy researchers for many years.

Progress is clearly being made, and the use of computers in language testing is giving new insights and opportunities. Nevertheless, many of the developments that we would ultimately expect of an intelligent computer based testing procedure will not be realised for many years to come.

2.3 Testing on the Internet: Opportunities and Limitations

More traditional "computerized testing", the delivery of a test by computer that could just as easily be delivered by paper-and-pencil, is therefore still a very real option for language testers. The speed of marking and decision taking, and the resulting reduction in costs associated with the administration of tests, combine to recommend computerized testing to us when it is available and practical. The delivery of these traditional CBTs on the internet is of particular interest, for a variety of reasons.

Firstly, the only software needed to take the tests is a standard World Wide Web (WWW) browser. These can be loaded onto any type of computer, making the test delivery system truly platform independent. The only requirements relate to hardware (the need for a modem), and a reasonably fast processor to download the information from the host server providing the test. With most CATs it is essential that specialist software be resident on the machine of the test taker, and although immediate results are available, these results cannot be sent back across the internet and stored on the server of the test provider. We therefore have significant gains in efficiency in internet delivered tests where certification is involved.

The second important advantage of the internet as a means of delivery is that the tests can be delivered to any machine linked to the internet, at any time convenient to the provider and the client. One example of this is at the University of Surrey, where individual students may be asked to take a distance placement test if there is any reason to suspect that the course for which the student has applied may not be appropriate for them on the basis of the student's current ability. It is not in the interests of a student to travel overseas to attend a course only to discover that it is far too easy or difficult for them. Where there is concern, tests can be delivered at a distance prior to any commitments being entered into. This saves much anxiety on the part of the student, and could result in significant savings both for them and the University in administrative costs.

Another situation in which a test delivered over the internet might be extremely useful would be where a group intends to travel from one country (or institution) to another, for a course. In these cases, the group travels as a whole, or usually does not go. Distance testing provides crucial data that enables institutions to take decisions regarding the appropriateness of the course for the group, and identify any members of the group who may need additional help or tuition. This, in turn, may also aid in the process of budgeting for such activities. It is certainly the case in this example that potential disasters can be avoided before individuals and groups have entered into major commitments and incurred significant costs.

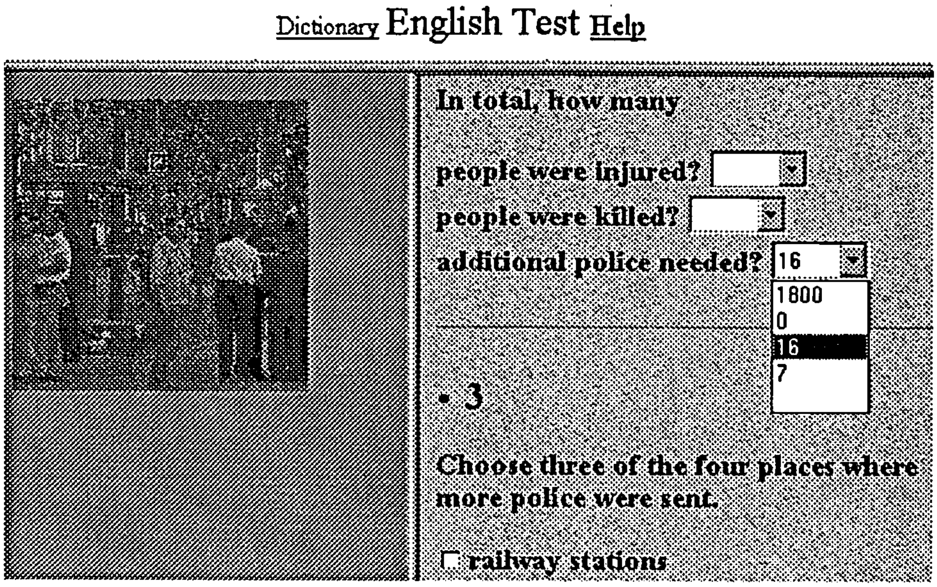
The WWW also provides advantages in the flexibility of test design. It is quite feasible, for example, to use the frames facility of modern browsers such as Netscape 2.0 or higher to divide the computer screen into windows or "frames", each of which contains a content "page". Prompts may be set up on a series of frames and sent to a particular frame, and can incorporate text, images, audio, and video, where computer links are reliable and quick. In fact, the flexibility of html in designing web pages makes it possible to design a range of novel task types through the imaginative combination of multimedia in a frames environment (Fulcher, 1996). This is demonstrated in figure 1, which shows a screen from a prototype web based listening test. The top of the screen contains a help function that allows the test taker to browse through explanations of the key terms used in the rubrics of the test. Also available is a link to a dictionary; this is not stored on the university server in the United Kingdom, but on another server in the United States.

Here lies a further advantage of web based tests: in principle, links can be established to information, help facilities, databases, or libraries, to deliver the kind of indirect performance test frequently recommended for placement purposes in academic programmes (see Robinson and Ross, 1996). Tests need no longer be self-contained, water-tight units, but involve the use of information from the outside world, to any degree the test designer wishes to incorporate it. This potential can be used to increase the "authenticity" of some testing activities, but from a measurement perspective it raises many questions that still need to be investigated. Test developers should be aware that systematic construct irrelevant variance (bias) may be introduced into the assessment procedure, and care should be taken to isolate and quantify any such effects.

In figure 1 the left hand frame of the screen is updated with pages that contain audio and video files. These can be played by the test takers using a set of menu options controlled by the right mouse button. This could just as easily be text, graphics, or some other form of illustration. The right hand frame contains the test items page and the buttons that update the pages in the left hand window. It is necessary to scroll down the right hand window to view each of the test items in turn, but the test taker can return to any previous items if required, until the test "submit" button is finally pressed.

In computer based testing on the internet, innovation is clearly possible where there is flexibility over the format and content of the prompt. However, it is not as easy to be as innovative in the area of item type. Most internet browsers support multiple choice, multi-choice, pull-down menu and constructed response item types, and combinations of these. For example, multiple pull-down menus can provide matching or sequencing items. Constructed response items may be of two types: limited constructed response where a word or short phrase is required, and which is automatically scored against a template, and extended constructed response, which must be e-mailed to human raters for scoring. In this respect, little has changed since Alderson (1986) found it difficult to design innovative item types for computer based language tests.

In summary, we are currently in a situation where innovation and flexibility is possible with prompts and tasks, but items must be of four basic types, or combinations of these. But as it is primarily with prompts and tasks that innovation will be concerned with in the immediate future, this is not seen to be a major disadvantage of internet delivery.



Dictionary English Test Help

In total, how many

people were injured?

people were killed?

additional police needed? 16

1800
0
16
7

• 3

Choose three of the four places where more police were sent.

□ railway stations

Fig. 1. A Screen Capture from a Prototype Listening Test

3. Equity in Computer-Based Testing

3.1 Equivalence of Forms

Computer based tests are subject to the same standards of reliability and validity that would be expected of any test. However, in the case of computer based tests, certain critical issues of equity have been raised. The classic statement of these concerns may be summarised from the Guidelines for Computer Based Tests and Interpretations (APA, 1986):

- Test developers should demonstrate that paper-and-pencil and computer based versions of a test are equivalent forms.

- The two forms of the test should rank order test-takers in approximately the same way.

- The means and standard deviations of the two forms should be approximately the same.

The main reason for the implementation of what could be called “minimal standards” is simply that most computerised tests are paper-and-pencil tests that have been translated into the new medium for the ease of scoring and reporting that the technology provides. The three equity issues are therefore related directly to the medium, and focus the attention of test developers and users to two key issues: the change in medium may alter the construct underlying the existing test, or alter the scale. If either of these were to happen, it would not be possible to interpret the scores in the same way as is possible for the paper-and-pencil form of the test. The scores will *mean* something different.

A great deal of useful work has been conducted into the equivalence of paper-and-pencil with computer based forms of tests. Bunderson et al. (1989) provide an overview of studies of test equivalence to the end of the last decade, which shows that 3 studies had shown a higher score on computer based tests, 13 had shown higher scores on the paper-and-pencil test, and 11 had shown no difference in scores across forms. They concluded that lack of familiarity with computers could be assumed to be a major factor in achieving lower scores on computer based tests, but that scores on paper-and-pencil tests were lower for younger test-takers who had not been familiarised with the method of completing the answer booklet. The issue of familiarity is not new in language testing. It has always been accepted that test-takers should be familiar with the item types and mode of test delivery in advance of taking a test to avoid introducing these factors as confounding variables in score interpretation. Further, it has recently been suggested by Russell and Haney (1997) that it is increasingly becoming unfair to test writing ability by paper-and-pencil in situations where learners have become used to composing directly on word processors. In this case, one would expect the computer delivered version of a writing test to result in higher scores simply because of familiarity. Yet the higher score would more accurately reflect the ability to be tested. Familiarity with test format and item types is a factor that effects all testing, and is not specific to computerised testing, however much it has become associated with it in recent literature.

The most recent meta-analysis of equivalence of paper-and-pencil with computer based forms of tests has been conducted by Mead and Drasgow (1993). This study is concerned with the method of delivery, paper-and-pencil or computer, as most other studies have been, but also includes two other variables: conventional vs. adaptive tests and power vs. speeded tests. We have already discussed CATs to some extent, but it does seem reasonable to consider the possibility that a CAT would not yield the same scores as a paper-and-pencil version, or a computer based non-adaptive version. The fact that a CAT presents different items in different orders to different people, means that no two test takers are likely to undergo the same experience. It is certainly the case that a comparison of total-score correct cannot be made between forms of the test, so any attempt to look at the issue of equated forms must be done in terms of a latent scale. It also seems likely that the speededness of the test may affect scores across forms. On the computer screen, the fact that scrolling has to be operated by the test taker could result in a score reduction on the computer based form. Mead and Drasgow (1993) report that in the meta-analysis of 28 studies, the computer tests were slightly harder than their paper-and-pencil counterparts, but that the only variable to significantly effect scores was speededness. Whilst correlations between tests was on average .91 for different administration modes, the cross mode correlation for speeded tests was .72 on average. One possible explanation for this finding is the differences in motor skills required of paper-and-pencil compared to computer based response techniques, when working under severe time limits. However, when the test is a timed power test, there is no significant influence of medium on test scores. Mead and Drasgow (1993: 456) conservatively state:

> "Our conclusion – that a computerized version of a timed power test can be constructed to measure the same trait as a corresponding paper-and-pencil form – should not be taken for granted for any computerized test."

This should be a warning that future computer based tests should be submitted to the same rigorous scrutiny as previous computerized tests, especially if they are speeded.

3.2 Further Equity Issues

Although the issue of equivalence has dominated the discussion of computer based testing, this is not the only equity concern, and as we have indicated, may have been part of the reason for lack of innovation in the format of computer based tests. Other equity issues relate directly to:

- Previous experience of using computers. Factors in this category include the familiarity of test takers with the computer itself, the frequency with which a computer is used (if at all), and familiarity with the manipulation of a mouse with two buttons. If the test is delivered over the internet using a standard browser such as Netscape or Internet Explorer, familiarity with the WWW (perhaps also including frequency of e-mail use) should be taken into account.

- Attitudes of test takers to computers, the software being used (the degree of user-friendliness, for example), and the WWW in general also needs to be investigated. It is at least feasible to suggest that negative attitudes to the medium of delivery could effect test scores, and this may be more likely among those with little or no experience of using computers or the internet.

- Finally, the background of the test taker may be relevant to the validity of the test score. Background factors worthy of investigation would seem to be age, gender, geographical location, and level of education or subject specialism in the case of applicants for university places.

The largest move to computer based testing is likely to take place with the introduction of the computer based TOEFL in 1998. Educational Testing Service (ETS) has conducted a significant amount of research on the TOEFL takers access to and experience with computers, in their attempt to design a computer based TOEFL that is minimally dependent upon previous experience. In the case of ETS this has involved the development of a tutorial package that test takers do prior to taking the TOEFL.

Taylor, Jamieson, Eignor and Kirsch (1997) investigated computer familiarity in the TOEFL test taking population and its effect on test scores, including gender and geographical location as variables. Using analysis of covariance (ANCOVA), Taylor et al. argue that significant statistical relationships between some of the sub-tests of the TOEFL and computer familiarity were so small as to be of "no practical importance". The same picture emerged for gender and geographical location. In total, the scores of around 20% of TOEFL test takers may be effected by the computer based form, although it is assumed that the tutorial package will further minimize this. Despite these claims, it would appear that 20% of the test taking population is a significant number, and for them the change in scores may have some practical importance. The assumption that a front-end tutorial package will mitigate the effects of the change in medium should at least be the subject of research, prior to the launch of the new TOEFL.

4. One Small-Scale Study

4.1 Questions

In order to exemplify some of the problems that should be acknowledged and researched, we will refer to one small scale study conducted at the University of Surrey during 1997. In this project, a test used for internal placement was computerised and placed on the WWW for delivery. It was necessary to answer a number of basic questions before the test could be used as a basis for decision making. The questions it is appropriate to ask are: is the computer based form sufficiently reliable for its purpose? How well are the two forms of the test correlated? Are scores infected by systematic construct irrelevant variance from any particular source?

Fifty seven students attending pre-sessional courses in 1997 were asked to take both the paper-and-pencil and the computerised versions of the test. These were identical in every respect, other than the medium of delivery. The CBT was taken in a standard university computing laboratory using the Netscape 3.0 internet browser. The test is a timed-power test, and was designed to take 45 minutes in total. The computing laboratory was invigilated by two course tutors, who helped students log onto the test, and ensured that each test taker submitted the answers for scoring on completion of the test.

After taking the computer based test, data was collected from each student on:

— Computer familiarity, consisting of:

- Frequency of computer use
- Familiarity with the mouse
- Familiarity with the WWW
- Frequency of e-mail use

— Attitudes towards taking tests on the internet, including:

- Preference for paper-and-pencil format
- Preference for internet format
- Student estimated likelihood of getting a higher grade on one or the other test

— Background of the test taker, including:

- Age
- Gender
- Primary Language Background
- Subject area specialism

The data were analysed to compare the test scores for the sample of students across the two test formats. Scores on the computer based test were investigated using ANCOVA for the effect of computer familiarity, attitudes, or background of the test taker. In this process the paper-and-pencil based test was used as a covariate to take ability into account. It should, however, be stressed that with a sample of only 57 students, it is impossible to investigate complicated interaction effects of factors that may threaten the interpretability of test scores; such sample sizes only allow the investigation of main effects.

4.2 Discussion

Test reliability was estimated using the one parameter IRT method, using Rascal software (ASC, 1994). Reliability for the paper and pencil test was estimated at .91, whilst the reliability of the computer based test was .93 with an average standard error of .31.

The two forms of the test were highly correlated, as can be seen in table 1, below. However, a significant correlation of .82 only translates into 67% shared variance. Thus, 33% of test variance is unique to one test or the other. A possible explanation for this will be offered below.

Table 1. Correlation Between the Two Forms

Correlations

		TEST	CTEST
Pearson Correlation	TEST	1.000	.824**
	CTEST	.824**	1.000
Sig. (2-tailed)	TEST	.	.000
	CTEST	.000	.
N	TEST	57	57
	CTEST	57	57

** Correlation is significant at the 0.01 level (2-tailed).

In figure 2, a line graph representing the scores of the 57 students on the two tests is presented. The lines represent observed scores, rather than the possible range of scores that could be expected from any further experiment, and so should be treated with some caution. However, the similar patterns across scores indicates that whilst there is clearly some method effect that reduces the correlation between the two forms, it should in principle be possible to use the computer based form to arrive at placement decisions.

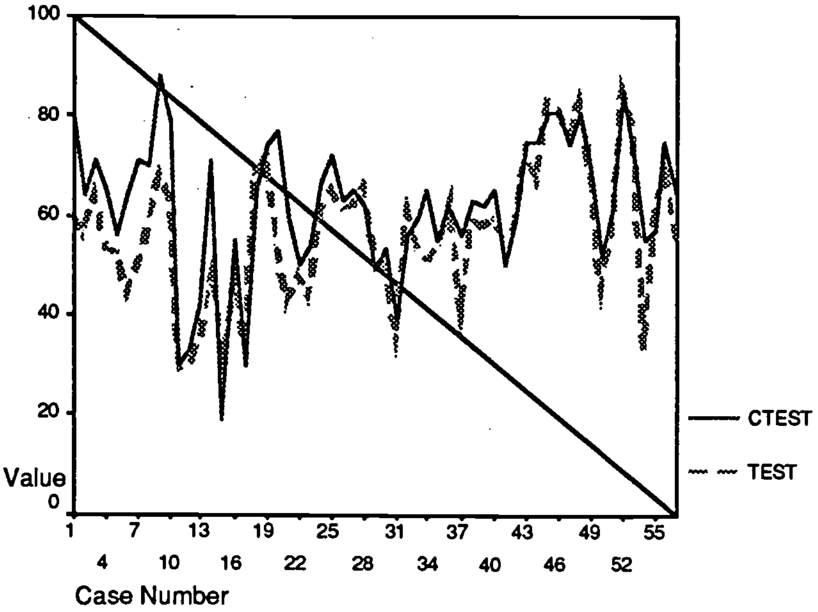


Figure 2: Line Graph of Tests on Two Forms for 57 Students

The areas of computer familiarity, attitudes towards taking tests on the internet, and background of the test taker, were also investigated. The independent variables in each case were measured on a 5-point Likert scale, except for polar responses (for example "which test did you prefer?") and questions relating to age. No significant results were found, apart from one: the first language background of the test taker. This finding cannot be attributed to variation in the ability of the sample by country, as there was no such difference between students with different first languages on the paper-and-pencil test. Nor can it be put down to lack of familiarity of computers and the internet among some students, as they had not recorded significantly different experience or exposure to computing on the questionnaires. On closer investigation, it appeared the systematic variance in scores could be attributed to lower scores on the CBT among students whose first language was not Indo-European, primarily those from Japan and South Korea. In this sub-section of the population there appears to be an interaction effect between L1 and taking a test on the computer, in specific relation to the fact that *the test* (a familiar activity) is being taken *on a computer* (a familiar machine), which are bring brought together for the first time in an unfamiliar way. If this hypothesis turns out to be correct, it is essential to investigate the kinds of pre-test taking activities that would reduce this effect, and how much of the treatment is necessary to eliminate it. This is precisely the kind of problem that has not been investigated by the TOEFL researchers, in their work described above.

5. Conclusions

This paper has attempted a brief overview of the use of the use of computers in language test delivery, and presented evidence to suggest that the use of the internet (WWW) is a very real option that should be considered by test developers. The advantages of an internet test have been listed, but a number of aspects have also been problematized. The latter include the current inability to develop adaptive tests for internet delivery, and the threat that under certain circumstances for some sub-groups of the test taking population, bias may be a worry.

Whilst it is certain that a great deal of research needs to be conducted into CBT, it is also the case that innovative experimentation with what is possible using the WWW as a medium of delivery may lead to changes in language testing that will be as profound as the introduction of the first CATs.

References

Alderson, J. C. (1986) *Innovation in Language Testing: Can the Micro-computer Help*? University of Lancaster: Special Report No 1, Language Testing Update.

APA.1986. Guidelines for Computer Based Tests and Interpretations. Washington D.C.: American Psychological Association.

ASC.1994. MicroCAT Testing System. Minnesota: Assessment Systems Corporation.

Bunderson, C. V., Inouye, D. I. And Olsen, J. B. 1989. "The Four Generations of Computerized Educational Measurement. In Linn, R.L. (Ed) Educational Measurement (3rd edition). Washington, D.C.: American Council on Education. 367–407.

Burstein, J., Frase, L. T., Ginther, A. and Grant, L. 1996. "Technologies for Language Assessment". Annual Review of Applied Linguistics, 16, 240–260.

Fulcher, G. 1996. "Taking the Web to Task: the Internet as a Medium for Test Delivery." Paper presented at the Language Testing Forum, Lancaster University, 22 - 24 November.

Holtzman, W. H. (Ed.) 1970. Computer-assisted instruction, testing and guidance. New York: Harper Row.

Larsen-Freeman, D. and Long, M. 1991. *An Introduction to Second Language Acquisition Research*. London: Longman.

Mead, A. D. and Drasgow, F. 1993. "Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis". *Psychological Bulletin*, 114, 3, 449 – 458.

Mislevy, R. J. and Wilson, M. 1992. *Marginal Maximum Likelihood Estimation for a Psychometric Model of Discontinuous Development.* Unpublished report, Princeton, N.J.: Educational Testing Service.

Perkins, K., Brutten, S. R., and Gass, S. M. 1996. "An investigation of patterns of discontinuous learning: Implications for ESL measurement." *Language Testing* 13, 1, 63–82.

Robinson, P. and Ross, S. 1996. "The Development of Task-Based Assessment in English for Academic Purposes Programs." *Applied Linguistics* 17, 4, 455 – 476.

Russell, M. and Haney, W. 1997. "Testing Writing on Computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives* 5, 3. [Online] http://olam.ed.asu.edu/epaa/v5n3.html

Taylor, C., Jamieson, J., Eignor, D., and Kirsch, I., 1997. "Measuring the Effects of Computer Familiarity on Computer-based Language Tasks" Paper presented at the Language Testing Research Colloquium, Orlando, Florida, 8 March.

Uses of the PET (Preliminary English Test) at Sultan Qaboos University

RICHARD LORING TAYLOR
Sultan Qaboos University
Muscat, Oman

I. Introduction

The Language Centre at Sultan Qaboos University used the PET (Preliminary English Test) for a period of five years, from 1991 to 1996. During this time the PET was used for a variety of purposes. The initial purpose was as an entry level proficiency test, to determine the level of students entering the university. The second purpose was as an exit level proficiency test, to determine the level of students after they had completed the Language Centre program. The third purpose was as a streaming test, to stream the students into programs of various lengths and intensity. The fourth purpose was as an exit level achievement test, to determine whether students had reached a level sufficient from them to proceed to credit based courses in their major fields of study.

All students entering the Language Centre took the PET as an entry level proficiency test and as a streaming test. Only two groups took the test as an exit level proficiency test—1993 students in the College of Commerce and 1993 English Education specialists in the College of Education. Students took the PET as an exit level achievement test according to the requirements of the various colleges. In 1996, the PET was replaced as an exit level achievement test by an IELTS-type test purposely designed by UCLES. In 1997 the PET was replaced as a streaming test by an in-house designed test.

II. Characteristics of the PET

The PET is one of a "suite" of tests produced by UCLES (University of Cambridge Local Examinations Syndicate). The latest brochure describing the PET is dated January 1996. For a comparison between the PET and other English Language Examinations, see Susan Davies & Richard West, The Longman Guide to English Language Examinations (Pitman, 1981; Longman, 1989). For evaluative commentary on the PET see Keith Morrow, of the Bell Educational Trust, in J. Charles Alderson & Karl J. Krahnke, Reviews of English Language Proficiency Tests (TESOL, Washington, D. C. 1987).

According to the PET brochure, the PET was introduced in the late 1970's as a PASS/FAIL instrument equivalent to IELTS BAND 3.5 in response to (a) a demand for an examination at a level lower than that of the First Certificate and (b) the Council of Europe Threshold Level (1975) as defined by van Ek and Trim. The PET assumes 375 contact hours of study. The PET was revised in 1988/89 and again in 1992/93. As a point of comparison, in the "suite" of UCLES tests, the Certificate of Proficiency (IELTS BAND 6.8) is the standard test used in Britain to determine whether a student is prepared to enter University. According to Davies and West, "Successful performance in the test should equip students with the language abilities which would enable them to enjoy a normal social life in an English-speaking country in line with the recommendations of the Council of Europe's Threshold Level" (P.114).

Regarding the PET lexicon, the Handbook (P.17) makes the following observations: "Candidates should also know the more specialized lexis appropriate to their personal requirements, for example, nationalities, hobbies, likes and dislikes. It should be noted that the lexis in the authentic material contained in PET is not modified unless the original wording would constitute an unfair test of students at PET level. As a result several words not shown in the lexicon appear in every version of the PET examination. However, no answers to questions in PET depend solely on a knowledge of the meaning of any word outside the lexicon. Thus, though appearing in some of the authentic material, words not contained in the list will be largely irrelevant to answering questions on that text. This lexicon is for UCLES use only."

To summarize (and re-phrase) the above points, the test was designed as a pass/fail test to determine whether candidates could "enjoy a normal social life" in a non-academic native speaking environment. To preserve authenticity, one may assume that this is an environment of some specific group of native speakers, as of British English. This description differs from the context at Sultan Qaboos University, which is a non-native speaking academic environment in which a variety of dialects of English may be spoken. The definition of a "normal social life" appears linked to the lexicon, which emphasizes such "personal requirements" as hobbies. Since the candidates are expected to be entering a native speaking environment, the question of cultural bias in the definition of such "personal requirements" is not expected to arise. The text is based upon a "core lexicon," which the candidate is expected to know. However, the details of this core lexicon are not specified, and the lexicon is known only to the makers of the test. In addition, every version of the test will include in its "authentic" material lexical items outside this core lexicon. One assumes that, since these fall outside the lexical domain being tested, the candidate will not be familiar with them unless they fall under the domain of his "personal requirements." However, knowledge of these items will not be required to understand the passage. It is not made clear whether candidates must know ALL the items inside the core lexicon in order to infer from the context the meaning of the items outside the core.

An affiliation between the PET and the Council of Europe "Threshold Level" is specified in both the brochure supplied by UCLES and in secondary sources. However, the precise nature of this affiliation is difficult to determine. In the PET brochure, the point is made that the test is not necessarily linked to any specific program of study. (In other words, PET can function as an independent proficiency test.) On the other hand, it is stated in the brochure that "up to 50%" of people taking the test take "some sort of program" leading into the test, including the Threshold Program or the Look Ahead Project. If candidates are taking a program leading into the test, it could function as an achievement test based on the program in question. Apparently students within the "threshold level" program would fall somewhere within this 50%. It appears that PET is trying to have it both ways to define itself as both an achievement test and a proficiency test.

Whereas the PET is vague about the content it examines, the Council of Europe program is almost compulsively precise. Not only does the C of E program define the lexicon down to the last word; it also includes a "grammatical inventory of T-level English, some methodological implications of waystage and threshold level" (P.v), as well as "language learning objectives in a European unit/credit system," specifying T-level according to such language parameters as "situations, social roles, psychological roles, types of language activities, language functions, behavioral specifications, topics, notions, language forms and degree of skill."

In the chapter entitled "Language Activities," the point is made that the T-level is defined differently for the different language skills. For speaking and listening, the T-level is defined as a level where the learning is able to transfer grammatical structures or lexical items to new contexts or situations. On the other hand, for reading and writing, the learner is not expected to have reached this transfer level. Van Ek observes: "The objective for writing at T-level is extremely limited. It is assumed that for this skill the actual needs of the majority of the members of the target group do not go beyond the ability to write letters of one particular type and to fill in certain forms. This means, in fact, that no general ability to write is required but only a strictly limited formulaic manner of expression" (P.24). Regarding the skill of reading, Van Ek states: "The objective for reading is also narrowly restricted. At T-level the learners will be able to read road signs, public notices, menu items, and simple brochures sent in return for letters written by the learners themselves" (P.25). The descriptions for reading and writing at the T-level appear to correspond to the descriptors for IELTS BANDS I & II.

Although the skill levels required for reading and writing at the T-level may be limited and formulaic, the skill levels required for listening and speaking are comprehensive. Under "social roles"one finds "stranger/ stranger, friend/friend, private person/official"; under "psychological roles" one finds "neutrality, equality, sympathy and antipathy"; under "place" one finds terrace, open air swimming pool, camping site, indoor market, canteen, youth hostel, lost property office, surgery, chemist, public lavatory, sauna, language institute, art gallery, night-club, ferry" etc. Under "topics" one finds "personal identification, types of accommodation, amenities, flora and fauna, professions, conditions of work, employment prospects, intellectual and artistic pursuits, aversions, invitations, sensory perception, hygiene, insurance, academic qualifications, auto repairs" etc. Under "language functions" one finds "inquiring whether an offer or invitation has been accepted or declined, inquiring whether someone has remembered or forgotten something or someone, inquiring whether others are obliged to do something or have permission to do something, expressing disappointment, gratitude, preference, intention, surprise, fear, sympathy, granting forgiveness, expressing disapproval, regret or indifference" etc. Under "behavioral specifications" one is expected to perform such tasks as "explain what forms of art one is interested in, describe taxation regulations, describe employment conditions and unemployment benefits, characterize countries according to their prospects for tourism" etc. One must also comprehend such "notions" as "differences between existence/nonexistence, presence/ absence, availability/nonavailability, anaphoric/ nonanaphoric deixis, instrumental/benefactive relations, disjunction/conjunction" etc. One must also comprehend and express "frequency, continuity, intermittence, permanence, temporariness, priority, ownership, reference without time focus, repetitiousness, uniqueness" etc.

Not only must be able to one to express all these ideas; one must be able to transfer these skills and notions in order to achieve communicative competence in a variety of original and unfamiliar contexts or situations. Apparently the only speaking or listening task one is not expected to perform at the T-level is to comprehend a list of the tasks one is expected to perform at this level. The skills levels required for speaking and listening must correspond to at least IELTS BAND VI. The discrepancy between the skill levels required of the different language skills appears designed to reflect that of a native speaker, who may achieve sophisticated communication in listening and speaking long before he develops equivalent skills in reading and writing. But although the native speaker model appears implicit in the skills hierarchy, van Ek insists that the program represented or recommended was designed for adult learners. A program which could achieve such a range of skill levels would have to utilize some sort of highly specialized audio/lingual approach.

Regarding the relationship(s) between the PET and the T-level, the following concerns appear significant. First, since the term T-level was designed by and for the Council of Europe, if the PET were based on this criterion, the expected target audience for this test would be individuals from the European community. Many of the concepts, such as "sauna" and "youth hostel," support this point. Second, the descriptors for the English speaking Union's nine BAND scale appear to assume the same or equivalent level of competence for each of the four language skills. In the PET, the four language skills are tested and graded; grades are issued in the form of BANDS, for the total test and for each of the different skills sections. The level tested appears to be equivalent across the four skills. If this is true, the PET cannot be considered representative of the T-level. A third point has to do with test weighting. In a test representing the T-level, speaking and listening should carry much more weight than listening and reading. However, in the PET, at least in the most recent version, all four skills carry the same weight. (In the version reviewed by Morrow (Alderson P.20), the point distribution is as follows: Reading and Writing: 40 marks; Listening: 15 marks; Speaking: 25 marks.) A fourth point consists of a combination of the second and third points. If a candidate's reading skill is in the range defined by the T-level, he might not even comprehend the instructions or the options for the listening component of the test. In the PET the skills are not tested in a pure form. Reading and writing receive an emphasis in the PET that is not representative of the T-level program. A final point concerns the role of testing, as defined for the T-level. According to van Ek and Alexander (P.250), "grading should closely approximate the situational context in which the communicative competence is required, and testing objectives should correspond to teaching or curriculum objectives." Van Ek describes the goal of the T-level program as "defining the learning objectives with such precision that they can fit within an integrated unit/credit system applicable across the European community." In other words, the test should represent a curriculum which, in turn, should be defined as precisely as possible. But in the case of the PET, neither the curriculum nor the test content is specified. In other words, any relationship between the PET and the T-level program must be considered tenuous at best.

The PET has been reviewed by Keith Morrow (Alderson, P.21). Following are some of his observations: "Results are reported as overall Pass or Fail. However, it is not clear whether candidates must pass all three sections or if compensation is permitted. No technical manual or information about scoring method is publicly available. The official specification and description of these examinations is vague. As always with examinations from this source, it is necessary to infer the rationale and the approach adopted from sample papers. No explicit specification of the content is provided beyond very general statements in the information booklet. In the absence of an explicit statement of target competence we are left to guess what any individual will have achieved in 350 hours. In addition, UCLES' terminology is very loose. Page 1 of the "Information for Centres" states, "Preliminary English Test is a test of achievement. In the absence of a specification of test content, this use of 'achievement' is idiosyncratic."

Morrow makes the following points: The information booklet, as well as the test itself, by simultaneously hinting at and disclaiming connection with any specific program of study, blurs the distinction between an achievement test and a proficiency test. Little information which might be used in evaluating the test is publicly available, including evaluation criteria or reliability statistics or studies. Since results are issued only in the form of broad bands rather than precise scores, the value of validity studies must be limited. The published literature indicates that the PET was designed as a PASS/FAIL proficiency test. For a PASS/FAIL test to achieve optimum reliability, all or most of the questions should be set as close as possible to the break point between passing and failing. In contrast, such a broad spectrum test as the TOEFL is considered to have equal reliability across its entire range. Generally, proficiency tests tend to be broad band tests.

III. Modifications to the PET and Their Consequences

The initial function of the PET was as an entry level proficiency test, given in order to determine the level of the students according to some international norm. Sultan Qaboos University Language Centre results based on years of testing indicate that around 20% of entering students are able to pass a test set at the IELTS BAND 3.5 level.

The next function of the PET at Sultan Qaboos University was as a streaming test. In order for the PET to function as a streaming test, the following changes in the instrument were agreed upon between UCLES and the Language Centre. First, the FAIL group was divided into two BANDS—I & II; the pass group was also divided into two BANDS—III & IV. It was not published whether the test was re-written to include additional questions at each of these new break points or whether these BANDS simply represented arbitrarily chosen scores on the test as set at the original level. The Language Centre repeatedly re-adjusted the break point between BANDS I & II. While such adjustments may have been academically justified, they tend to increase one's doubts that the test was actually re-written to include questions set at the break points. If the test did not include additional questions at the break points, then the domain of optimum reliability would have been between BANDS II and III (the original PASS/FAIL level). Although the results of the streaming test varied from year to year and college to college, the figures were approximately BAND IV—5%; BAND III—15%; BAND II—40%; BAND I—40%.

The second change was to apply the IELTS descriptors to the BANDS produced by the divisions described above. Since the descriptors could have implications for the curriculum and course goals, this step was significant. The third change was to change the name of the PET from "Preliminary English Test" to "Placement English Test." The name change reflected the altered function of the test. A fourth change was to drop the speaking component of the test and re-distribute the points for the speaking component among the other components of the test. Apparently, the reason this was done was that it was felt that qualified evaluators for this component were not available. It should be pointed out that, not only would such changes have altered whatever reliability the test may originally have had; dropping the speaking component would sever whatever connection the test may once have had with the "Threshold Level" program, for which speaking was evidently considered the primary skill.

The methodological implications of such a change in skill priority could be substantial. If speaking is not included in the PET test, and the PET test becomes an achievement test at the exit level, speaking could receive minimal attention in a curriculum leading up to such an exit requirement. However, the Threshold system, on which the PET was supposedly based (however loosely) clearly assumes that speaking is the most important skill.

In general, a streaming test is an example of a norm referenced test, since students are placed in groups based on their performance relative to that of their peers. Regarding such tests, Bachman (P.211) observes: "Because of the differences in score interpretation, Criterion Referenced and Norm Referenced tests are typically used in different decision contexts. Norm Referenced tests scores are most useful in situations in which comparative decisions, such as the selection of individuals for a program are to be made." Bachman goes on the observe: "In order for such tests to be interpretable, they (the scores) must be clearly distinct from each other. Thus tests that are intended to provide NR score interpretations are designed and developed to maximize inter-individual score difference, or score variance."

Bachman goes on to explain that, for such purposes as streaming, it is crucial for the testing entity to take into account the standard error of measurement of the test in order to prevent decisions being taken on the basis of variations in score which are statistically insignificant or which do not measure real variations in language ability. The standard error of measurement of a test is determined by a formula based on the actual standard deviation of the population being tested and the reliability of the test, as determined independently. Unfortunately, in the case of the PET, published reliability figures are not available. Furthermore, the changes made in the test would have significantly altered its reliability, had it been known. It could be, therefore, that on the basis of small variations in score within a test domain of limited reliability, students were divided into groups supposedly corresponding to qualitative differences in skill level equivalent to over 200 hours of language instruction.

The next function for which the Language Centre used the PET was as an achievement test at the exit level. Students were required to achieve BAND IV on this test before being allowed to proceed with the academic programs. In addition, students had to pass each BAND level before being allowed to proceed to the next BAND level. This program was begun in the College of Commerce before being adapted by the other colleges in the university.

This function of the test corresponded to a shift from a norm referenced test to a criterion referenced test. Regarding this type of test, Bachman (Pp.74-75) makes the following observations: "CR tests are designed to enable the test user to interpret a test score with reference to a criterion level of ability in a domain of content. An example would be the case in which students are evaluated in terms of their relative degree of mastery of course content, rather than with respect to their relative ranking in the class. Thus all students who master the course content might receive an 'A,' irrespective of how many students achieve this grade. The primary concern in developing a CR test are that it adequately represent the criterion ability level or sample the content domain... The two primary distinctions between NR and CR tests are (1) in their design, construction and development; and (2) in the scales they yield and the interpretation of these scales. NR tests are designed and developed to maximize distinctions among individual test takers, which means that the items or parts of such tests will be selected according to how well they discriminate individuals who do well on the test as a whole from those who do poorly. CR tests, on the other hand, are designed to be representative of specified levels of ability or domains of content."

Since Bachman makes it clear that NR tests and CR tests differ in design, construction, development, scale and interpretation of results, it would appear that the use of the same test as both a CR and an NR instrument would constitute unorthodox testing practice.

The reason for Morrow's concern for the "looseness" of the UCLES terminology should now be clear. According to Bachman, a proficiency test, an achievement test and a streaming test should be three completely different types of tests, with different design criteria and interpretation of results. For an achievement test, detailed specification of course content (as is done in the T-level program) is important. For a proficiency test, equal reliability across the domain of the test is important. For a streaming test, determination of the standard error of measurement is important. The PET pretends to be all three types of tests without meeting the precise criteria of any one type.

When the BAND IV requirement was first introduced in the College of Commerce and Economics (in Spring 1994), students experienced considerable difficult in achieving this requirement. In fact, 59 out of 60 students failed to reach BAND IV after a semester of study at the BAND III level. Several measures were undertaken to deal with this problem. First sections of the reading and writing sections were re-written to make the test more "culturally relevant." By this time, very little remained of the original PET. Second, a teacher with computer expertise loaded all available versions of the test into the computer and produced a lexicon. He then produced computer based exercises to help the students master this lexicon. These measures prove so successful that virtually all the students were able to achieve the specified requirement of improving one BAND level per semester of study (24 contact hours per week).

Through these measures the PET was transformed into a genuine achievement test, in the sense that the content and/or skill criteria became known in advance of the test. The curriculum at this stage became dominated by the test. However, by forcing the test off the fence, so to speak, question could be raised whether the instrument continued to function in any way a proficiency test, that is as a random external measure of student ability. Before these measures were taken, it could reasonably be described as a proficiency test.

One other group of students took the test on a proficiency basis. In January 1995, 1993 English Education Specialists took the PET three semesters after they had entered the university. Since this testing was not required as part of their curriculum, students took no special measures to prepare for the test. The results, which are analyzed in detail in another study, confirm the figures from the first group of Economics students to be tested. The general conclusion is that it takes the average student three times as long to achieve BAND IV on a proficiency basis than on an achievement basis. Students, teachers and administrators tend to talk about BAND IV as if were a single entity. It may be necessary to distinguish between proficiency BAND IV and achievement BAND IV.

IV. Concurrent Validity or Predictive Value of the Pet

A study was conducted on 90 English Education Specialists, who entered the university in the fall semester, 1993. These students were given the PET upon entry, and were divided into groups, as follows: Group A, BANDS III & IV; Groups B & C. BAND II; Groups C & D, BAND I. During the academic year 1994-95, these students were given a different proficiency test, the CELT (Comprehensive English Language Test) at the beginning, middle and end of the year. In the middle of the year, these students were re-tested with the PET. For this academic year, grades in all English courses were tabulated. There were four English courses in each semester, or a total of eight courses throughout the year—literature 1 & 2, writing and grammar 1&2, reading comprehension 1&2, phonetics/linguistics. Various comparative studies on the grades and test scores were run.

a. Group Studies According to Anova

The mean grades for all courses were calculated for all courses on a group by group basis. Although there were differences in the mean grades, the result of the ANOVA was that no group was statistically distinct from any of the other groups to the .05 level. As far as mean grades were concerned, all five groups constituted a single statistical population. It is of course possible that different skill levels between the groups were obliterated by teachers grading "on the curve." This would constitute a problem in absolute vs. relative grading standards.

For each of the three sittings (pre-test, mid-test and post-test) of the CELT, ANOVAS were run to determine whether the groups constituted a single statistical population. At the pre-test, no group was statistically distinct from any other group at the .05 level. However, at the mid-test and post-test, statistically significant differences between the groups began to emerge. These differences could be attributed to differences in the teaching methods of the teachers in the English courses.

Although these tests were sensitive enough to pick up differences in teaching methods during the course of the year, they did not show differences in the groupings as determined initially by the PET.

b. Pet/grade Pearson Correlation Studies

A Pearson correlation was run between the PET post test score (December, 1944) and the overall mean grades for all courses during the academic year 1993-94. Since the PET was taken in the middle of the academic year, this test could be regarded as a concurrent validity test for the PET. The Pearson Correlation Coefficient, r, for this test was .3707. The coefficient of determination (r 2 x 100) was 13.74%. This indicates that whatever was measured by the PET contributed 13.74% to the overall student grades. The correlation between the PET scores and the mean course grades can be described as "very low positive correlation." The level of significance of the above figures was .001, which can be considered highly significant. In other words, this measure of correlation can be considered reliable.

According to its accompanying literature, the PET measures various language skills in a global or integrative fashion. As administered by the SQU Language Centre, the scores from this test are released in the form of BANDS for the skills of writing, reading comprehension and listening comprehension. One would assume that certain of these skills would be more important than others in achieving success in specific types of academic courses. In order to verify this hypothesis, BAND scores on the components of the PET were correlated with grades in apparently related academic courses. The results of this study were as follows:

The first test run was between the PET scores in reading comprehension and composite grades in literature courses. The Pearson Correlation Coefficient for this test was .2886. The coefficient of determination was 5.21%, indicating that the skill which was measured by the reading comprehension component of the PET contributed 5.21% toward the average literature grade. One would have thought that ability in reading comprehension should have made a greater contribution to the grade in literature courses. Either something is wrong with this component of the PET, or students are not required to read or at least are not evaluated by this activity in most literature courses. There may be some truth to both these alternatives.

The correlation between the writing component of the PET and the composite literature grades was slightly higher, at .2886, with a high level of significance at .01. The coefficient of determination was 8.33%, indicating that the skill of writing, as measured by the PET contributed 3.21% more to the grades in literature than did reading comprehension. This would make some sense since students would be evaluated by their writing in most examination contexts. However, the level of correlation is still low. This fact needs to be explained since, for the writing component of the test, the PET grades depend more on the grader than on the test itself. Evidently, the PET grades must be evaluating the skills of writing differently from the way literature graders evaluate the same skill.

The correlation coefficient between the writing component of the PET and the overall mean grades in language development courses was .3233, with a significance level of .01. The coefficient of determination, at 10.45%, was 2.31% higher than the equivalent figures for literature courses. This difference might be explained by the greater subjectivity in the grading of literature courses.

The final test was between the listening comprehension component of the PET and the grades in the Phonetics course, which was a practical course in distinguishing and producing the sounds of English, both individually and within words and sentences. One would have thought that the skill most relevant to success in the tasks required for such a course would be listening comprehension. However, the correlation coefficient for this test was a mere .0899. The coefficient of determination, at 0.81%, suggests that the factors which were measured by the listening comprehension component of the PET contributed less than 1% to the grade in this course. A possible explanation for this discrepancy may be that the accent of the speaker used in the listening component of the PET may have been unfamiliar to most of the testees.

c. CELT/PET/Grade Correlations

Mean grades and CELT scores were tabulated for students who had obtained various BAND levels on the second sitting of the PET. For students in BAND IV, the mean CELT score was 47.6; the mean grade was 2.28. For students in BAND III, the mean CELT score was 47.15; the mean grade was 2.36. For the single student in BAND II, the CELT score was 43.7; the grade was 2.38. The mean grades and CELT scores for students in BANDS II, III and IV appear virtually identical although no attempt was made to determine the significance of these differences.

For students who obtained BAND IV in the PET, the range of CELT scores was from 41.7 to 54.3; for students who obtained BAND III in the PET, the range of CELT scores was from 29.3 to 58.3. For students who obtained BAND IV in the PET, the range of mean grades was from 1.69 to 2.50; for students who obtained BAND III in the PET, the range of mean grades was from 1.21 to 3.34. For both grades and CELT scores, the range of students who scored BAND IV within the PET fitted almost within the middle of the range of students who scored BAND III.

Since the PET in this case functioned as a proficiency test, the difference between BANDS III and IV should have corresponded to around 500 contact hours of language study. Yet, according to mean grades and CELT scores and ranges of mean grades and CELT scores, it seems to have been largely a matter of chance whether a student obtained BAND III or BAND IV on the PET.

V. Conclusions

According to ANOVA, the differences between the groups generated by the PET as a streaming test are not statistically significant. According to PET/Grade Pearson Correlation studies, overall correlation between PET scores and grades is very low, and correlation between components of the PET and grades in corresponding courses ranges from very low to no correlation at all. CELT/PET/grade studies do not identify significant differences between students who obtained different BAND levels on the PET. These low correlation figures could have arisen from the alterations which had been made to the PET or to the inappropriateness of the instrument for some of the uses to which it was put. When the PET is used as an achievement test, students can improve one BAND level in 200 contact hours of study; however, when the PET is used as a proficiency test, it apparently takes students at least three times as long to improve by one BAND level. More detailed studies of the PET are difficult due to the way in which grades are presented and due to the absence of published reliability and validity data.

Bibliography

Alderson, J. Charles & Krahnke, Karl J. Reviews of English Language Proficiency Tests (TESOL, Washington D. C., 1987).

Bachman, Lyle F. Fundamental Considerations in Language Testing (OUP, Oxford, 1990).

Brown, H. Douglas. Principles of Language Learning and Teaching (Prentice Hall, Englewood Cliffs, 1980).

Carroll, Brendan J. Testing Communicative Performance (Pergamon, Oxford, 1980).

Davies, Susan & West, Richard. The Longman Guide to English Language Examinations. (This book was first published by Pitman in 1981; it was revised and re-published by Longman in 1989.)

Hammerly, Hector. Fluency and Accuracy: Towards Balance in Language Teaching & Language Testing, (Multilingual Matters, Clevedon, England, 1991).

Harris, David P. & Palmer, Leslie. Comprehensive English Language Test (CELT) (two versions), (McGraw Hill, N. Y., 1986).

Harris, David P. Testing English as a Second Language (McGraw-Hill, NY, 1969, new edition 1988).

Heaton, J. B. Writing English Language Tests (Longman, London, 1975, new edition 1988).

Henning, Grant. A Guide to Language Testing (Newbury, Cambridge, MA., 1987).

Hinkle, Dennis, Wiersma, William & Jurs, Stephen. Applied Statistics for the Behavioral Sciences (Houghton Mifflin, Boston, 1988).

Krashen, Stephen D. Principles and Practice in Second Language Acquisition (Prentice Hall, Hemel Hempstead, 1987).

Mackay, R. (ed.). English for Specific Purposes (University of Michigan Microfilms, Ann Arbor, 1986).

Madsen, Harold S. Techniques in Testing (OUP, Oxford, 1983).

Mcall, Robert. Fundamental Statistics for Behavioral Sciences (Harcourt Brace Jovanovich, NY, 1986).

Morrow, Keith (Bell Educational Trust). "Review of PET," Reviews of English Language Proficiency Tests (TESOL, Washington D. C., 1987).

Oxford, Rebecca. "Review of CELT," Reviews of English Language Proficiency Tests (TESOL, Washington D. C., 1987).

Van Ek, Jan Ate. Threshold Level English (Pergamon Press, Oxford, copyright 1975, new edition 1980).

Issues in Foreign and Second Language Academic Listening Assessment

CHRISTINE COOMBE, JON KINNEY AND
CHRISTINE CANNING
United Arab Emirates University

1.0 Introduction

Unlike foreign and second language listening in a conversational setting, a major part of F/SL listening in a university context involves lecture comprehension or listening to extended monologue. Many observers (Richards 1983; Dunkel 1991; Chastain 1988; Flowerdew 1994; Rost 1990) have noted there are differences between general listening and academic listening. In some cases, these differences have important implications for testing. Overall, test designers face many similar challenges in the development of either general or academic listening assessment instruments.

The purpose of this paper is four-fold. First, we will summarize the differences which exist between academic and general listening. Secondly, we will review the current state of academic listening assessment practices. Thirdly, several important issues which must be taken into consideration when designing academic listening assessment instruments will be identified and discussed. Finally, we will propose a formal schema that can be used by practitioners to evaluate the academic listening tests they currently produce and use.

2.0 General vs. Academic Listening

Researchers (Richards 1983; Dunkel 1991; Flowerdew 1995) point out that academic listening comprehension has its own unique characteristics which distinguish it from conversational listening.

Richards (1983) provided a comprehensive taxonomy of aural skills involved in conversational discourse or general listening. These microskills include clustering; recognizing redundancy; comprehending reduced forms; comprehending other performance variables such as hesitations, pauses, false starts, and corrections; understanding colloquial language (ie. idioms, slang, shared cultural knowledge); processing speech at different rates of delivery, processing prosodic features such as stress, rhythm, and intonation patterns; understanding and using rules of conversational interaction such as negotiation, clarification, turn-taking, topic nomination, maintenance, and termination.

Richards (1983) observes that academic listening requires higher level skills in addition to those needed in general listening comprehension. Recently, Flowerdew (1995) has identified several skills in addition to general listening skills which a student must employ in order to listen effectively in an academic milieu. These skills include:

- activating specialized background knowledge
- distinguishing relevant information from irrelevant
- negotiating meaning given limited opportunities to interact with the speaker
- concentrating and comprehending for long periods of time
- integrating the incoming lecture with related information derived from reading assignments, textbook information, handouts, and OHP or black/whiteboard lecture notes
- taking notes

3.0 A Change In Listening Assessment Practices

Although a great deal of research has been conducted on the listening skill, listening comprehension assessment remains a rather dusty corner in the world of professional testing (Douglas 1988; Thompson 1991). In the past, the testing emphasis was placed on a students' ability to discriminate phonemes, to recognize stress and intonation patterns, and to record through a written product (usually in an objective format) what had been heard. While such testing perhaps assessed general listening skills, especially those associated with bottom-up processing, the higher-order academic listening skills were not addressed.

3.1 A shift in emphases

More recently, however, emphasis in testing listening skills has shifted toward contextualized tests of listening comprehension where communication of meaning provides the focus rather than structural understanding (Rost 1990). This shift corresponds directly and as a result of a corresponding movement away from the belief that listening comprehension is a one-way bottom-up process towards the theory that comprehension requires a more complex combination of top-down and bottom-up processing. In short, listening is not just the consecutive processing of sub-skills. Rather it is a combination of a number of different levels of processing to which no fixed order can be attributed (Lewkowicz 1991).

3.2 Indirect vs. Direct Testing

In terms of testing, two competing traditions are discussed in the literature, indirect and direct testing of the listening skill.

3.2.1 Indirect Tests

Indirect tests assess language performance indirectly by predicting performance in certain language use situations (Henning 1987). The discourse and tasks in a indirect test tend to focus on bottom up micro-skills. Testers characterize indirect tests as being less natural, more contrived. Indirect tests are norm-referenced, standardized, and often used to compare proficiency levels of individuals in a large population. Examples of indirect tests of listening would be the Michigan Test of Aural Comprehension as well as Section A of the TOEFL Listening Test.

3.2.2 Direct Tests

Direct tests of listening comprehension, on the other hand, measure language use in more realistic and less contrived, communicative, situations. The emphasis in a direct test of listening comprehension is to assess the proficiency of realistic task performance. Direct tests emphasize macro and top-down listening skills. A direct test reinforces the principle that both teaching and learning should focus on what students really need to know. (Hansen & Jensen 1994:242). A recent example of a direct test of academic listening comprehension is the Test of Listening for Academic Purposes (T-LAP).

The distinction between indirect and direct testing raises particularly important issues which must be considered by the academic listening test designer. While indirect tests are reliable, they appear to suffer in terms of content validity and authenticity. Direct tests on the other hand, are more authentic and valid, but their reliability is sometimes questionable. The challenge for the academic listening test designer is to measure academic listening directly using authentic discourse as stimuli while at the same time not compromising on testing standards.

3.3 Cornerstones of Good Testing Practice

From a testing perspective, all tests, including academic listening tests, should possess certain characteristics. These four characteristics, validity, reliability, practicality, and washback are known as the "cornerstones" of good testing practice. Tests need to be valid, ie they should assess what has been taught, how it has been taught. Scores obtained should be reliable or consistent. For example, reliable tests produce similar scores with the same student over repeated administrations. Practicality is an especially important issue for classroom teachers. Tests should be easy to administer and score. Finally, tests should provide washback or feedback into the curriculum, the course materials, student proficiency level, and a teacher's own teaching. In short, everyone involved should learn from testing experiences (Alderson, Clapham & Wall 1996).

In practice, it can be very difficult to uphold the cornerstones of good testing practice while at the same time retaining authentic features of actual communicative discourse as they occur in natural, non-testing environments. This is particular true in the assessment of general or academic listening comprehension.

4.0 Important Issues in the Assessment of Academic Listening Comprehension

The assessment of academic listening comprehension presents several challenges to ELT professionals. Because listening comprehension in any context is unobservable in and of itself, assessment of comprehension must rely on instrumentation in which testers design task items and responses that approximate real listening situations. Authenticity in instrumentation can be difficult to achieve. Furthermore, listening instruments almost always contain some degree of skill contamination. In academic listening, skills other than listening (reading or writing) are almost always required to successfully complete the assessment task. Another set of challenges arise from the student population being tested. Tests that assess academic listening skills must take into account the cultural background of the students as well as the schema or background knowledge they bring to the test. Testers should also be careful not to offend local sensibilities. Finally, test developers need to control for variables in the presentation of the listening passages they present to students in academic tests.

4.1 Instrumentation

In order to assess listening skills testers rely on testing instruments which are designed to simulate real listening events and appropriate responses to those events. Instrumentation problems arise when the listening text, task, or medium differ too greatly from what people really listen to, what they really do when they listen, or in what package or form the oral message is heard.

As much as possible, listening test instruments should display authenticity of text, task, medium. In academic listening tests, texts should mirror actual language as it is used by teachers and students in academic contexts. For example, extended monologues or formalized interactions seem more realistic in academic assessment than a set of de-contextualized and unrelated short exchanges. Texts should display features of oral language rather than just being oral readings of written language. Exam tasks should simulate how students actually respond in effective ways in real listening situations. Notetaking tasks, for example, would seem to display more task authenticity than say, multiple choice questions. Finally, the media used to present the text or script should be as authentic as possible. Audio tapes appear to be less authentic in an academic context than live readers or video presentations that show people speaking.

Instrumentation should demonstrate authenticity of text, task, and medium. This is especially important given the inauthentic or contrived nature of the testing process. Although a testing situation can never fully simulate an authentic context, attempts should be made in order to make components of the instrumentation as real to life as possible.

4.2 Skill Contamination

Test designers should also be sensitive to the phenomenon of skill contamination which threatens the validity of many tests of academic listening comprehension (Buck 1988). In a perfect world, reading or writing would not interfere with the assessment of the listening skill. In reality, however, the successful completion of most if not all listening tests requires competence in other language skill areas, particularly reading and writing.

Listening tests can be contaminated in terms of either the task or the test procedure. Task contamination refers to instances where successful completion of the task depends on students reading or writing ability in the L2. Procedural contamination, on the other hand, refers to the way in which exam procedures are explained to students. Directions given in the native language would be less contaminated, for example, than directions given in the target language. The issue of skill contamination suggests that written directions in L1 would be superior to other options.

Some degree of contamination is not necessarily disadvantageous. In fact, testers should be willing to sacrifice test purity in some instances "in order to maximize task authenticity and provide a positive washback effect on teaching" (Lewkowicz 1991:26)

4.3 Cultural Sensitivity

Although the specific cultural background of students studying in heterogeneously grouped classes in target or ESL communities rarely influence test design, in an EFL context cultural issues need to be addressed. Many ELT professionals develop assessment instruments for students with a shared cultural background that is not shared with the teacher or test designer. In such cases, problems can arise through the selection of culturally inappropriate testing topics, situations, or formats. Tests in such contexts should reflect a sensitivity to students' cultural background, customs, and values.

4.4 Background Knowledge

An undisputed fact in F/SL education is that background knowledge or schemata have a profound influence on L2 student listening and reading comprehension. (Anderson 1985; Carrell 1983; Bernhardt 1986; Carrell, Devine, & Eskey 1988). Furthermore, there is every reason to suggest that both top-down processing and schemata play as important a role in listening as in reading (Long 1989; Buck 1992). In academic settings background knowledge is particularly important. Students must relate their background knowledge to the message of the incoming lecture in the absence of opportunities to negotiate meaning or benefit from contextual cues.

Testers should be careful to control for background knowledge. They might do so in one of three ways. First, test designers can ensure students have an equal amount of background knowledge by writing listening tests that exploit specific course materials. Second, they can provide students with the requisite background knowledge during testing via advanced organizers or question prompts. Third, test writers can use topics they believe to be entirely unfamiliar to the student population. In any case, an attempt to standardize the presence or absence of background knowledge should be made in any academic listening text that purports to be a valid indicator of comprehension.

4.5 Presentation Factors

Practitioners should control for various presentation factors in academic listening tests that can influence student scores and invalidate results. These variables are either media variables, speaker variables, or procedural variables.

Media variables are those that come about as a result of the media used to deliver the message. Examples of commonly used media in academic listening assessment include audio tapes, video tapes, and live readers. Differences in the media used influence the degree of comprehension exhibited by students. Students, for example, comprehend less when paralinguistic cues are unavailable (Kellerman 1992).

Speaker variables represent a second type of presentation factor. Speaker variables arise when speakers having different accents, styles, speech rates, or different degrees of professional or social status deliver a message. A number of speaker variables have been identified which significantly affect listening comprehension in L2 students. Such variables include rate of speech (Griffiths 1990), perceived expertness of the speaker (Markham 1988), gender of speaker (Markham 1988; Coombe, Kinney, & Canning 1997), dialect of the speaker (Tauroza 1997), and pauses (Blau 1991).

Procedural presentation factors are those in which non-standardization of procedures results in unreliable test scores. Common examples of procedural factors include the number of text repetitions given, length of time given to complete the task, undue speaker stress or intonation used to cue correct answers, policies regarding false starts, disfluencies, and misspeaks in live reader situations.

Practitioners should attempt to standardize presentation factors when possible. Teachers should also be aware that certain factors in their current practice might interfere with student comprehension in some instances. When different speakers are used with different groups of students, for example, speaker interaction could influence test results in a variety of ways.

5.0 How Effective are Your Academic Listening Tests?

The issues cited above suggest a schema or evaluative tool which can be used by EF/SL testing and teaching practitioners to assess the suitability and effectiveness of their academic listening tests. We propose practitioners utilize a scheme to compare and evaluate their assessment instruments (See Figure 1). Designers should ask themselves the following questions as they review their product:

Instrumentation	This category refers to the content of the test itself including the listening script, the task, and how it is delivered. • How well does the listening test approximate the target setting of an academic listening context, such as a live lecture? • How well does the listening text reflect a realistic portrayal of an academic talk or speech? Are academic spoken discourse markers, redundancies, emphases, asides, jokes, colloquialisms, instances of idiomatic language, etc. absent or present? • To what extent does the task require students to adopt realistic listening roles? • To what extent does the task require students to interact with the text?
Skill Contamination	This category describes the extent to which the test measures pure listening comprehension as opposed to other language skills. • What other langauge skills are needed to successfully complete the task? • If other skills are required, to what extent do these skill requirements outweigh listening skill requirements? • Do the learners have the skills required to properly process and understand the test directions? • Are directions given in L1 or L2?
Cultural Sensitivity	This category describes the extent to which the test reflects a sensitivity to students' shared cultural background, customs, and values. • Is the content of the text objectionable in any way given religious, political, or social orientations in the local community? • To what extent does test content reflect local beliefs, customs, and traditions?

Background Knowledge	This category describes the extent to which successful completion of the assessment task relies on a standardized background knowledge. • To what extent do the text and task allow students to access background knowledge? • To what extent has the test writer controlled for student background knowledge?
Presentation Factors	This category describes the extent to which test designers have controlled for presentation factors like speaker, media, and procedural variables. • What media (audio, video, live reader) is being utilized to deliver the listening text? • Are listening tests consistently presented via this medium? • To what extent have the designers controlled for individual speaker variables? • To what extent are administration procedures articulated and applied consistently? • In large-scale testing situations, to what extent have standards for live reading been calibrated?

6.0 Implications for the Practitioner

The valid and reliable testing of academic listening comprehension is a complex process. A number of issues specific to listening as well as to listening in an academic setting need to be addressed by the practitioner.

Practitioners should be aware that there are significant differences between academic and general listening as well as similarities. Academic listeners are required to listen to extended monologue with limited opportunities to interact with the speaker. Furthermore, they must utilize higher order skills in addition to general skills in order to comprehend effectively.

Practitioners should also make sure that the cornerstones of good testing guide their academic listening test design and administration, bearing in mind that there is often a trade-off between validity and reliability in the testing of listening comprehension.

Practitioners should also evaluate their present assessment instruments in terms of the challenges addressed in this article. They should examine the authenticity of the text, task, and medium used in their current tests. They should note the degree of skill contamination in their tests of academic listening. Practitioners also need to consider whether or not their listening tests reflect a sensitivity to students' cultural background. The extent to which success on the test relies on shared background knowledge in the student population also merits consideration. Finally, practitioners should control for presentation factors including speaker variables like speech rate, accent, gender, pausing because these factors influence comprehension scores. If live readers are used to test students, policies regarding procedural presentation factors like number of repetitions, length of time given to complete the task, policies regarding false starts, misspeaks and disfluencies should be established beforehand and adhered to.

Some of the challenges posed in testing academic listening seem formidable, but they can be met in a way that adheres not only to good testing practice but also to the theory and research related to academic listening. By considering some of the issues mentioned in this article, practitioners can make their tests more effective and make the process of testing academic listening a more rewarding experience for their students.

References

Alderson, J., C. Clapham & D. Wall. 1996. *Language Test Construction and Evaluation.* Cambridge, UK: OUP

Anderson, A. & T. Lynch. 1988. *Listening.* Oxford, UK: OUP.

Anderson, J.R. 1985. *Cognitive Psychology and its Implications.* 2nd ed. New York: Freeman.

Bernhardt, E. 1984. Toward an information processing perspective in foreign language reading. *Modern Language Journal,* 68:322-31.

Blau, E. 1990. The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly* 24 (4):746-753.

Buck, G. 1988. Testing listening comprehension in Japanese university entrance examinations. *JALT Journal,* 10 (1&2): 15-42.

Carrell, P.J. 1983. Background knowledge in second language comprehension. *Language Learning and Communication* 2 (1): 25-34.

Carrell, P.J. 1984. Evidence of a formal schema in second language comprehension. *Language Learning,* 34: 87-112.

Carrell, P., J. Devine, & D. Eskey. 1988. (eds.) *Interactive Approaches to Reading.* Cambridge: CUP.

Chastain, K. 1988. *Developing Second Language Skills.* Chicago: Rand McNally.

Coombe, C., J. Kinney, & C. Canning. 1997. Listening comprehension and the Gulf Arab student: A research initiative. *New Directions in Listening.* Cairo: AUC Press.

Douglas, D. 1988. Testing listening comprehension in the context of the ACTFL proficiency guidelines. *Studies in Second Language Acquisition,* 10, 245-261.

Dunkel, P. & J. Davis. 1994. "The effects of rhetorical signaling cues on the recall of English lecture information by speakers of English as a native or second language." In *Academic listening: Research perspectives*, J. Flowerdew (ed.), 55-74. Cambridge, UK: CUP.

Dunkel, P. 1986. Developing listening fluency in L2: Theoretical principles and pedagogical considerations. *Modern Language Journal*, 70, 99-106.

Dunkel P. 1991. Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly*, 25 (3), 431-457.

Flowerdew, J. 1994. "Research of relevance to second language lecture comprehension—an overview." In *Academic Listening: Research Perspectives*, J. Flowerdew (ed.), 55-74. Cambridge, UK: CUP.

Griffiths, 1991. Pausological research in an L2 context: A rationale and review of selected studies. *Applied Linguistics*, 12 (4):345-364.

Hansen, C. & C. Jensen. 1994. Evaluating lecture comprehension. In *Academic Listening: Research Perspectives*, J. Flowerdew (ed.), 241-268. Cambridge, UK: CUP.

Henning, G. 1987. *A guide to language testing*. Cambridge: Newbury House.

Kellerman, S. 1992. I see what you mean: The role of kinesic behavior in listening and implications for foreign and second language learning. *Applied Linguistics*, 13 (3):239-258.

Lew Kowicz, J. (1991). Testing Listening Comprehension: A New Approach? Hong Kong Papers in Linguistics and Language Teaching, Vol. 14: 25-31.

Long, D. R. 1990. What you don't know can't help you: An exploratory study of background knowledge and second language listening comprehension. *Studies in Second Language Acquisition* 12 (1): 65-80.

Markham, P. 1988. Gender differences and the perceived expertness of the speaker as factors in ESL listening recall. *TESOL Quarterly* 22:397-406.

Richards, J. 1983. Listening comprehension: Approach, design procedure. *TESOL Quarterly*, 17 (2), 219-39.

Rost, M. 1990. *Listening in language learning*. London: Longman.

Tauroza, S. & J. Luk (1997). Accent and Second Language Listening Comprehension. *RELC Journal*, 28 (1): 54-71.

Thompson, I. 1990. An investigation of the effects of text and tasks on listening comprehension: Some evidence from Russian. *Georgetown University Roundtable on Languages and Linguistics*, 294-305. Washington D.C.: Georgetown University Press.

Saudi Development and Training's Five Star Proficiency Test Project

JOHN POLLARD
Saudi Development and Training

This talk was offered 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*) tests.

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-stakes 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 (between 0.88 and 0.98) were obtained. 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 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 a first insight into validity. 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 expressions 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 emphasises 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 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 constructs of the domain. 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' contamination 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, 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 conversational 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.

APPENDIX I

KEY

- ♦ CONFIRMS UNDERSTANDING
- ◈ SEEKS CONFIRMATION
- ■ SEEKS CLARIFICATION
- ▩ INDICATES NEED FOR CLARIFICATION
- ◍ CONFIRMS OWN PREVIOUS TURN
- ✯ RE-FORMS OWN PREVIOUS TURN

♦ **CONFIRMS UNDERSTANDING** 1. Offers appropriate response. 2. Says "I understand" or = . 3. Says "Yeh", "Yeah", "u-uh", etc. 4. Agrees 5. Disagrees 6. Laughs (appropriately)	◈ **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.	▩ **IMPLIES 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.*

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.

CANDIDATE: Fahad Al-Radass	VERBAL INTERACTION (facilitating self-understanding)				VERBAL INTERACTION (facilitating other-understanding)	
TASK	1 CONFIRMS UNDERSTANDING	2 SEEKS CONFIRMATION	3 SEEKS CLARIFICATION	4 IMPLIES NEED FOR CLARIFICATION	5 CONFIRMS OWN PREVIOUS TURN	6 RE-FORMS OWN PREVIOUS TURN
4 Names	9	2	4	4	4	0
11 Numeracy	8	5	0	2	1	9
13 Al Harbis	8	1	0	0	4	3
15 Student Grades I	11	2	0	0	1	3
19 Student Grades II	6	2	0	5	2	1
23 Vehicles	3	4	6	3	1	0
24 Footballers	2	2	1	2	0	3
26 Kettle	1	0	0	0	0	0
28 Signs I	5	2	2	0	3	0
29 Fridge	8	3	2	5	1	0
50 Signs II	3	1	0	0	0	3
55 Kuwait City	5	1	1	4	3	0
62 Car ownership	8	1	0	0	4	4

References

Alderson J. C., Clapham C. and Wall D. (1995). *Language Test Construction and Evaluation*, CUP: 173-4.

Atkinson J. M. and Heritage J. (1984*). Structures of Social Action: Studies in Conversation Analysis*, CUP.

Barnwell D. (1989). *'Naïve' Native Speakers and Judgements of Oral Proficiency in Spanish.* LT: 6/2:152-63.

Chalhoub-Deville M. (1995) *A Contextualised Approach to Describing Oral Language Proficiency.* Language Testing 45/2:251-281.

Chaudron C. (1988) *Second Language Classrooms: Research on Teaching and Learning.* CUP.

Criper C. & Davies A. (1988) *Validation Project Report.* The British Council/Cambridge UCLES.

Eggins S. & Slade D. (1997) *Analysing Casual Conversation.* Cassell.

Ellis R. (1991) *The Interaction Hypothesis: A Critical Evaluation.*

In Sadtono, E. (Ed) *Language Acquisition and the Second Language Classroom.*

Færch C. & Kasper G. (1984) *Two Ways of Defining Communication Strategies.* Language Learning 34/1: 45-63.

Jacoby S. & Ochs E. (1995) *Co-construction: an Introduction.* In Jacoby & Ochs (eds) *Special Issue on Co-construction.* Research on Language and Social Interaction 28/3: 171-83.

Johnson K. E. (1995) *Understanding Communication in Second Language Classrooms.* Cambridge University Press.

Kramsch C. (1986) *From Language Proficiency to Interactional Competence.*

The Modern Language Joumal, 70 / iv

Lantolf, J. P. & Frawley, W. (1985). *Oral Proficiency Testing: A Critical Analysis.* The Modern Language Journal, 69/iv: 337-345.

Lantolf, J. P. & Frawley, W. (1988). *Understanding the Construct.* Studies in Second Language Acquisition, 10: 181-195.

Lazaraton A. (1992). *The Structural Organisation of a Language Interview: A Conversation Analytic Perspective.* System: 20: 373-86.

Lazaraton A. (1996). *Interlocutor Support in Oral Proficiency Interviews: The Case of CASE.* Language Testing 13/2: 151-172.

Long, M. H. (1983). *Native speaker / Non-native Conversation and the Negotiation of Comprehensible Input.* Applied Linguistics Speaker 4/2: 126-41.

Lumley T. (1993). *The Notion of Subskills in Reading Comprehension Tests: an EAP Example.* Language Testing 10/3: 211-234.

McNamara, T. (1996) *Measuring Second Language Performance.* Longman.

Messick S. 1994. *The Interplay of Evidence and Consequences in the Validation of Performance Assessments.* Educational Researcher 23/2: 13-23.

Morton J. & Wigglesworth G. 1994. *Evaluating Interviewer Input in Oral Interaction Tests.* Paper Presented at the Second Language Research Forum, Montreal, October.

Pica, T. & Long, M. H. (1986). *The Classroom and Linguistic Performance of Experienced vs ESL Teachers'* In R. Day (Ed.) *Talking to Learn.* Newbury House.

Pica, T. (1994). Review Article. Research on Negotiation: What Does It Reveal About Second-Language Learning Conditions, Processes and Outcomes ? Language Learning 44/3.

Pica, T. Lincoln-Porter, F., Pinanos, D., & Linnell, J. (1996). *Language Learners' Interaction: How Does It Address the Input, Output, and Feedback eeds of L2 Learners ?* TESOL Quarterly 30/1: 59-84.

Pica T, Holliday, L., Lewis, N., & Morgenthaler, L. (1989). *Comprehensible Output as an Outcome of Linguistic Demands on the Learner* Studies in Second Language Acquisition, 11: 63-90.

Pollard J. D. E. (1994). *Paper on Proficiency Testing.* IATEFL Testing Newsletter July, (1994).

Pollard J. D. E. & Underhill N (1996). *Developing and Researching Validity for a Computer-Resourced Proficiency Interview Test.'* Language Testing Update, 20: 9-52

Ross S. and Berwick R. (1990). *The Discourse of Accommodation in Oral Proficiency Examinations.* SSLA: 14:159-76.

Anthology Series SEAMEO Regional Language Centre (RELC), Singapore.

Ross, S. (1992) *Accommodative Questions in Oral Proficiency Interviews.* Language Testing 9/2: 173-186.

Ross, S. (1994). *Formulaic Speech in Language Proficiency Interviews.* Paper Presented at the Annual Conference of the American Association for Applied Linguists, Baltimore MD, March.

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974) *A Simplest Systematics for the Organisation of Turn-taking for Conversation*. Language 50/4.

Scheggloff E. A., Jefferson G. & Sacks H. (1977). *The Preference for Self-Correction in the Organisation of Repair in Conversation*. Language 53: 361-82.

Schegloff, E. A. (1995) *Discourse as an Interactional Achievement III: The Omnirelevance of Action*. In Jacoby S. & Ochs E. (eds) *Special Issue on Co-construction*. Research on Language and Social Interaction 28/3:185-211.

Shohamy E. (1983). *The Stability of Oral Proficiency Assessment on the Oral Interview Testing Procedures*. Language Learning 33/4: 527-40.

Shohamy E. (1988). *A Proposed Framework for Testing the Oral Language of Second / Foreign Language Learners*. SSLA 10/2: 165-79.

Shohamy E. (1993). *The Exercise of Power and Control in the Rhetorics of Testing*. In Huhta A., Sajavaara K. & Takala S. (eds) *Language Testing: New Openings*. University of Jyväskylä Institute for Educational Research, Jyväskylä, Finland: 23-38.

Underhill N. (1987) *Testing Spoken Language*. Cambridge University Press.

Underhill N. (1997) *Unpublished research report*.

Van Lier, L. (1989-a). *Reeling, Writhing, Drawling, Stretching, and Fainting in Coils: Oral Proficiency Interviews as Conversation*. TESOL Quarterly, 23/3.

Van Lier, L. (1989-b) *Puno: Teacher, School, and Language*. In Coleman, H (Ed.) Working with Language: A Multidisciplinary Consideration of Language used in Work Contexts. Mouton de Gruyter.

Wigglesworth G. (1993) *Exploring Bias Analysis as a Tool for Improving Rater Consistency in Assessing Oral Interaction*. Language Testing, 10/3: 305-335 Wigglesworth G (1994). *An Investigation of Planning Time and Proficiency Level on Oral Test Discourse*. LT 14/1: 84-106.

Young R. and Milanovic M. 1992. *Discourse Variation in Oral Proficiency Interviews*. SSLA: 14:403-24.

Young R. (1994). *Conversational Styles in Language Proficiency Interviews*. Language Learning 45/1: 3-42.

Zuengler J. (1993). *Encouraging Learners' Conversational Participation: The Effect of Content Knowledge*. Language Learning 43: 403-32.

Paper II:

The Influence of Assessor Training on Rater-as-Interlocutor Behaviour During a Computer-Resourced Oral Proficiency Interview-cum-Discussion (OPI/D) known as the "Five Star Test"

This considers the influences of the values, and the discourse behaviours of the NS Assessor as Interlocutor. These variables in assessments of oral proficiency interrelate in complex ways with features of test design such as tasks, prompts, topics, guidelines, and assessor training. This paper reviews some previous studies, and looks at measures taken in the design of the Five Star Test to improve discourse and interactive consistency. An exploratory study suggests that assessor training might be a more beneficial area of attention for test developers than recent research indicates—particularly when coupled with innovative features of test design.

Assessors of oral proficiency have been shown to carry preconceived, internalised, and perhaps prescriptive, notions of proficiency which operate on their judgements independently of band-descriptors and in spite of guidelines provided by examining and testing bodies. (Ludwig, 1984; van Lier, 1989; Barnwell, 1989; McNamara, 1990; Ross, 1992; Chalhoub-Deville, 1995). This may result in a failure to credit, or even a tendency to penalise learners for behaviours that constitute recognised models of second language performance (Canale & Swain, 1980; Lantolf & Frawley, 1988; Bachman, 1990; Alderson, Clapham & Wall, 1995; Bachman, & Palmer, 1982, 1984, 1997).

The problems for test reliability resulting from this are compounded in many OPIs (including "Five Star") where the assessor and interlocutor are the same person. The Steven Ross study cited above, for example, shows that proficiency ratings vary inversely with the amount of accommodation offered.

In the 1989 article, van Lier expressed the view that OPIs could be made more like real conversations. He did not give detailed indications of how this might be achieved, beyond citing an example of test design where more consideration than usual was given to the themes and topics of the tasks (van Lier 1989: 501-2), and pointing researchers in the direction of Conversational Analysis. Most of the research that has followed this lead has been based on tests that are already long-established. This is inevitable, given the nature of research funding. (Spolsky, 1995) However, it means that the first part of van Lier's proposition remains unexplored. This paper aims to take a very tentative step towards rectifying this, by basing an enquiry on a test that has been developed with a number of innovative features which are, to the best of my knowledge, currently unavailable to the proficiency instruments of major testing bodies such as ACTFL, IELTS and UCLES. These include aspects of Test design—including rating procedures, and with reference to the demands made on the Assessor-cum-Interlocutor.

Theme and Topic—with reference to the importance placed on local, cultural and personal saliencies.

Task Design—referring to the creation of "role-reversals", two-way information gaps, and the establishment of clear and unobtrusive performance criteria.

Assessor/Interlocutor Training—with an analytical glimpse of what can happen when there is none.

Test Design

One difficulty for OPI Assessor/Interlocutors appears to be their inability to relate knowledge of scale descriptors to actual performance. Band-descriptors are by nature generalisations. Many include expressions like 'when discussing a familiar event', 'in everyday conversation', etc. It is understandably difficult to compare real and particular instances of performance with such statements. Applying these descriptors whilst actually being engaged in the complex processes of interaction with the candidate further complicates the event. In the case of some OPIs (e.g. ACTFL OPI, Cambridge CASE) there is yet another requirement—namely, suppressing features of interaction which appear to be quite natural to non-test NS-NNS encounters—such as slowing down one's speech and supplying items that seem to be 'on the tip' of the candidate's tongue, or 'collaborative completion'. (Perret, 1990; Lazaraton, 1996: 154-5). Other tests require the assessor to refer to an interview format and/or evaluation criterion during the actual process of the OPI. There are recorded instances where this inauthenticates the interaction when compared with non-test exchanges. For example, the candidate "does some further topic talk on his name, . . . but all that he gets in response . . . are three weak agreement markers" because the interviewer is preoccupied at the time with his interview agenda. (Lazaraton, 1992: 378)

The philosophy underpinning the design of the Five Star test is that both the assessor and the candidate will behave more naturally, if the assessor is relieved of the burdens of monitoring the candidate and consciously deselecting responses, and if vague and obtrusive guidelines and evaluation interventions are removed from the event. Support for this can be found in work on Interlanguage (Selinker, 1972; Higgs & Clifford, 1982; see also Underhill, 1987 and Pollard, 1994). This is implemented through design features such as the use of pop-up on-screen text-boxes, which, after the initial assessor training, serve only as reminders of task requirements and evaluation criterion. In actual test use they rarely need to be accessed, and then only briefly, causing minimal distraction of the assessor from his / her engagement with the candidate. The auto-scoring mechanism which multi-functions as a task navigator is equally unobtrusive to the process and inconspicuous to the candidate.

These features can be seen in the first task, depicted below. At the bottom of the task screen, in very low profile, are three evaluation buttons which appear as arrows. To the left of each arrow is an icon (shown below as a square) 'pointing-and-clicking' which reveals pop-up scripted evaluation criteria, as shown in the speech bubbles. As mentioned above, these are used during assessor-training, and only in extremis during real tests.

In this way specific rather than global evaluation decisions are made; they are made on a local task-by-task basis rather than cumulatively; and they are made instantly rather than retrospectively. At the end of each individual test algorithm, (consisting of between 12 and 20 tasks) a histogram graph and set of band-descriptors for the cumulative performance across the constructs tested, are automatically generated.

There is no empirical evidence for this claim, other than an intra- and inter-rater reliability study referred to in Paper I—but reflective feedback suggests these features help to ease the burden on assessor or rater reliability.

Task

Each task on the Five Star Test has been designed with a specific criterion which, on the basis of pre-trialling, discriminates performance into one of three broad categories. (Pollard, 1994) The test is algorithmic and adaptive, so that all candidates begin at the same point, and then branch according to performance. The replication throughout the test of task types, supporting graphics, and screen configurations is employed to ease the burden on the assessor, reducing his/her need to access the on-screen pop-up guidance. This and the gradual introduction of multimedia features is also thought to diminish method effect due to lack of familiarity on the part of the candidate. For example sound from the computer is first encountered in simple instances in the early stages of a test pathway where there is a considerable amount of additional support.

One of the consistent criticisms of OPIs has been the lack of symmetry in the discourse they generate, which, at one extreme has been described as being more like an interrogation than a conversation (van Lier, 1989; Young & Milanovic, 1992; Zeungler, 1993; Young, 1996). Through the use of audio-recorded of L1 (Arabic) instructions, tasks can be set up where the information for proceeding with the test is given to the candidate rather than the assessor. For example, immediately following the initial task which is based around the registration of the candidate's details, the candidate is instructed (in L1) to gather similar information for on-screen boxes which are designated (First Name, Second Name, Family Name, Nationality, etc) in Arabic. These have to be entered in English by the assessor under the guidance of the candidate, effectively reversing the discourse initiative without requiring the candidate to use the computer. In other instances, task instructions are given in Arabic, and the candidate has to explain them to the assessor. This explanation forms a Speaking and Interaction task in its own right. Both of these are instances of what have been referred to in second language classroom contexts as two-way 'information gaps' and are designed to authenticate interaction through a real need to communicate. (Doughty & Pica, 1986; Nunan, 1988).

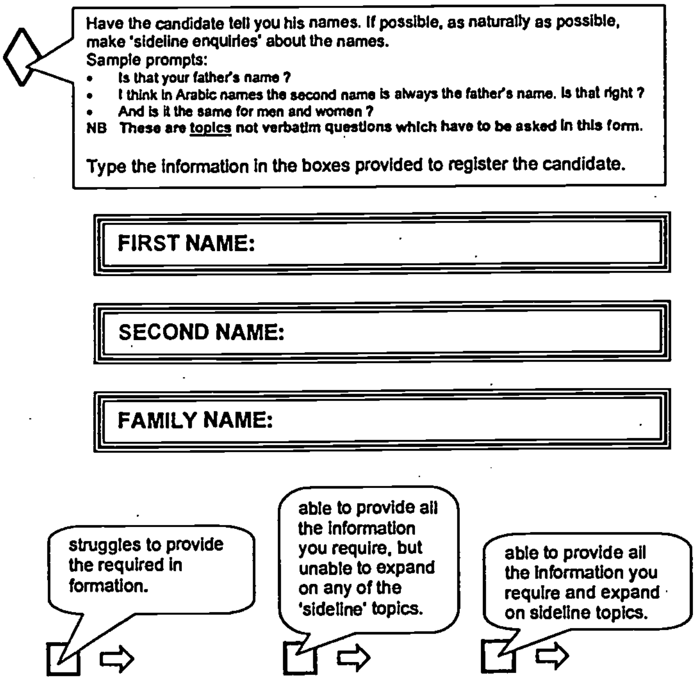
As mentioned above, task-by-task evaluation is one means by which the Five Star process seeks to alleviate the burden on the assessor. The first task, illustrated below, shows how this works. This task combines the registration procedure of obtaining personal details (names, approximate age, birthplace, etc) with specified 'sideline' discussions. As Lazaraton points out, in many OPIs the registration and introduction are assigned as uncredited 'warm-ups', 'losing', for evaluation purposes one of the most authentic phases of the event. (Lazaraton, 1992: 382)

Theme and Topic

There is research evidence that saliency of topic is a powerful influence on the discourse structure (Woken & Swales, 1989; Milanovic & Young, 1992; Zeungler, 1989, 1993; Young, 1996). In the example below, the 'embedded' task of having 'sideline' discussions is based on a wide range of name-related topics which were trialled for their accessibility to the test population in terms of the language sample they elicited, and the 'naturalness' with which they merged into the 'dominant' task—in this instance, registering biodata. An effort has been made to juxtapose themes of successive tasks which make for natural conversational progressions. For example, the discussion of names mentions locality (since family or 'tribal' names are regional); this leads into a discussion of birthplace and on to schooling experiences. This leads into post-school experiences, and then into travel experiences, etc.

In every task the attempt is to personalise the topic so that the candidate and not the assessor is the 'knower'. It has been shown that when topics are more equally shared between assessor and candidate—as in the case of academic subject specialisms—there can be an affective reaction bound up with self image. (Zeungler, 1989:238)

The following diagram gives an illustration of some of the developmental measures taken to moderate variables of Test, Task and Topic Design which have been empirically demonstrated to skew the discourse structure of OPIs.



Have the candidate tell you his names. If possible, as naturally as possible, make 'sideline enquiries' about the names.
Sample prompts:

- Is that your father's name ?
- I think in Arabic names the second name is always the father's name. Is that right ?
- And is it the same for men and women ?

NB These are <u>topics</u> not verbatim questions which have to be asked in this form.

Type the information in the boxes provided to register the candidate.

FIRST NAME:

SECOND NAME:

FAMILY NAME:

struggles to provide the required in formation.
able to provide all the information you require, but unable to expand on any of the 'sideline' topics.
able to provide all the information you require and expand on sideline topics.

Assessor / Interlocutor Training

A bigger (and I think more interesting) challenge, however, is defining Interlocutor Support, and ensuring that it is consistently offered. This is where Dr Chalhoub-Deville's interest in raters as dimensions of the proficiency both coincides with and departs from my own. Her concern is that raters bring with them their own, internalised evaluation criteria, which operate regardless of guidelines. This concern was also a part of the motivation for van Lier's 'warrant' for an enquiry into Conversational Analysis. The view he expressed in 1989 was that lack of detailed knowledge about the precise things that make proficient interactive performance was partly responsible for Assessors falling back on 'criterial linguistic features' (such as microlinguistic accuracies in terms of pronunciation and grammatical formatives) and disregarding instructions to base assessments on successful task completion.

What is the nature of this detailed knowledge, then, that might help us out of the difficulty? The tendency in research triggered by van Lier's seminal observations has been to identify tokens of contingency (from the conversational analysts) which shape the discourse structure of samples of audio-recorded OPI assessments. (Young & Milanovic, 1992; Young, 1996). Lazaraton (1992) used some video data, but again the focus was mainly on overall discourse structure, and the test in focus had design features which dominated the goal-orientation of the interviewer/assessor and skewed the process towards asymmetric contingency. The 1996 Lazaraton study looks more closely at the turn-by-turn construction with particular attention to interlocutor support, but does so primarily from linguistic perspectives of content and structural formatives. The Conversational Analysts focus much more attention on 'the omnirelevance of action' (Schegloff, 1995) and are more concerned than applied linguists with the paralinguistic features which attach to 'turn-constructional units'. In fact they examine how the linguistic, supralinguistic and paralinguistic interrelate, and recognise 'the sequential relevance to interaction for participants of eye gaze, facial expression, gesture, body deployment, pitch, intonation, vocal stress, orientation to objects in interactional space, laughter, overlap and its resolution, unfinished and suppressed syllables, and silence.' This reveals the limited scope of extant test research, and of studies which focus only on discourse structure and language—limitations which recent researchers have acknowledged (Young & Milanovic, 1992:422; Young, 1996: 37). Above everything else, it reveals the complex array of rater-as-interlocutor dependent variables.

With that in mind let me share with you some very raw and tentative information—I can't call it data. The background is as follows: Circumstances threw in my way something that I wasn't able to construct in a controlled research environment, though this may be possible in the future.

Managers working in a separate location from where the test was being developed had a requirement for some English proficiency assessments, and asked two new members of our ELT staff to administer the Five Star Test. At this stage the only version available to them was an early prototype version which did not include the assessor guidance and evaluation 'pop-ups'. The assessors were given time to flick through and familiarise themselves with the tasks, but were given no training in line with the design.

As a result, no-one had emphasised the importance of maintaining a friendly, non-judgemental mien when conducting the assessment, and it was later established that the teachers were unaware of supportive behaviours which have been studied in OPI, Counselling, and similar contexts. The test developers were powerless in the face of commercial pressures to do anything about this, but the teachers were willing to video some of their tests for later feedback, to 'ensure they were doing it right'. In contrast, and at the same time, other assessors were being trained in Riyadh where the principle development was taking place. The latter had, of course, been fully briefed in the intended approach, and some of their early test performances had also been videoed for developmental purposes. Between all the assessors, there were no gender differences and no significant age differences. All had at least three years' experience of living and working in the Gulf. The candidates assessed in the videos were also all-male, all were between the ages of 25-35, and an independent evaluation of the video data estimated that they covered comparable ranges of proficiency.

So, basically, the two pairs of Assessors were conducting themselves according to different briefings and guidelines: the one briefing led them to assume the role of 'tester' in the judgemental sense, and gave no indication that the interactions that took place around specific 'tasks' was a part of the assessment process; the other briefing stressed the importance of friendliness and supportiveness, introduced the assessors to some of the relevant literature, and allowed the pre-set criterion to steer the evaluation.

One of the behaviour variables referred to in the work on Conversation Analysis is 'eye-contact' or, as it is referred to in this field, 'gaze'. This behaviour is generally viewed in interactional terms as a 'display of recipiency' or 'co-participation' . (Atkinson & Heritage, 1984; Goodwin, 1984). On the basis of this, I decided to compare the frequency and duration of Assessor-initiated eye-contact between these assessors.

Method

Ten performances of the same task were identified for each of four assessors—the two who had not been briefed vis-a-vis interlocutor supportiveness, and two from those who had. The tasks were audio-recorded from the video and then transcribed. The transcriptions were printed and then manually 'marked-up' for instances and duration of assessor eye-contact with candidates.

Shared task-boundaries were identified (commencing with a request for the candidate's first name or application number and ending with the onset of the closing move before the assessor moved to the next task.) Importantly, the more discursive parts of this task—those prompted by the 'sideline enquiries' in the pop-up task instructions—were excluded from the analysis. This more 'socially' than 'functionally-oriented' part of the task was not implemented by the 'untrained' assessors, and its inclusion would have skewed the data.

Data

A table of eight column/cells was compiled for each assessor, all containing, from left to right:

1 the median of the total task duration, as recommended in Young & Milanovic (1992) and Young (1996) for instances where the data set is small and the range wide. (With such a small data set it could clearly be seen that this was a far better representation of the time each assessor typically spent on the task than the mean would have provided, as well as a better measure for inter-rater comparisons.)

2 the range, or longest and shortest time dedicated to this task in the samples.

3 the number of spoken turns (again, using the median).

4 the range of spoken turns.

5 the instances of assessor-to-candidate gaze (median).

6 the range of instances of assessor-to-candidate gaze.

7 the duration of assessor-to-candidate gaze across all samples, measured in seconds and hundredths of seconds. The means for each assessor across the ten tasks were used, having first omitted rapid 'glances' of less than three seconds duration, as these seemed to be fulfilling a different function than 'engagement and co-participation', and were singularly evident in one assessor.

8 the range of duration of assessor-to-candidate gaze across all samples, again with the exclusion mentioned in 5.

Results

The assessors who had been trained in accordance with the philosophy of the test were A and C. The only data which consistently varies between the trained and untrained assessors is the median figure for duration of assessor-to-candidate gaze, where more 'gaze-time' was given by the trained assessors.

Discussion

Due to the small sample size, these results are only an indication of what might be found through more rigorous enquiry. In addition to seeking more rigorous data and variable controls within the parameters employed here, further potential lies in examining whether, per comparable unit of talk, the putatively more affiliative/less judgemental approach by assessors elicits more interaction, as measured in comprehension checks, requests for clarification, etc. As an ethnographic exercise, groups from the target population could be asked to evaluate the relative styles of the assessors, as well as whether the candidates appear more or less reserved. It is probable that such research might be revealing in terms of cross-cultural paralinguistic behaviours, and might provide specific advice that could be given to trainee assessors.

As referred to in the Assessor A's pattern of eye contact differed from that of B, C and D, in that it included a number of rapid 'glances' which ranged from 00:29 to 03:04 seconds. These appeared to follow questions which were posed without any eye contact (while the assessor was looking at the computer screen). One could speculate that this behaviour emanates from an approach which views the questions as prompts for test performance rather than any desire to engage with the candidate or ask questions to find out information of shared interest. Without investigating it in the necessary detail, it appears that although the questions asked revolved around the task theme (names), they juxtaposed topics within that theme which bore little reference to the candidate's responses. This results in a series of questions and answers more characteristic of interviews in terms of 'asymmetric contingency' than the 'more conversation-like' pattern of reactive and mutual contingency. Interestingly, this candidate refers to the process as an interview when commencing the test. He also consistently fails to exploit the opportunities for 'ice-breaking' and bonding—particularly with candidates at the lower end of the range selected, and in one assessment, only made eye-contact twice with the candidate. On one occasion, a candidate felt it necessary to ask if he could know the assessor's name after the assessment had been completed.

On this basis perhaps we can surmise that two of the assessors in this study took on the role of judge but not the responsibilities of being a supportive interlocutor, and that in doing so they initiated and maintained shorter periods of eye-contact with their NNS-candidates. The NS-assessors who supportively engaged their NNS-candidates, and indulged in the collaborative construction of meaning before making an assessment, maintained longer periods of eye-contact with their NNS-candidates. Given the importance of 'gaze' in signalling commitment and participation that has been recorded in NS-NS conversation, it is likely that it plays a part in encouraging the co-constructional 'interactive' abilities of L2 learners in oral proficiency assessments. If this is so, developmental investment in assessor training and awareness-raising would contribute to both validity and reliability.

References

Alderson J. C., Clapham C and Wall D (1995). Language Test Construction and Evaluation, CUP: 173-4.

Atkinson J. M. & Heritage J. (1984). The Interaction of Talk with Nonvocal Activities. In Atkinson & Heritage (Eds) 1984: 223-4.

Bachman L. S. & Palmer A. S. (1982). The Construct Validation of Some Components of Communicative Proficiency. TESOL Quarterly, 16/4.

Bachman L. S. & Palmer A. S. (1984). Some Comments on the Terminology of Language Testing In Rivera, ed.: Communicative Competence Approaches to Language Proficiency Assessment: Research and Applications, Multilingual Matters.

Bachman L. S. (1990). Fundamental Principles in Language Testing. OUP.

Bachman L. S. & Palmer A. S. (1997). Language Testing in Practice. OUP.

Barnwell D. 1989. 'NaÔve' Native Speakers and Judgements of Oral Proficiency in Spanish. LT: 6/2:152-63.

Canale M. & Swain M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. Applied Linguistics 1/1.

Doughty C. & Pica T. (1986). "Information Gap"Tasks: Do They Facilitate Second Language Acquisition ? TESOL Quarterly 20/2: 305-325.

Chalhoub-Deville M. (1995) 'A Contextualised Approach to Describing Oral Language Proficiency'. Language Learning 45/2:251-281.

Goodwin C. (1984) Notes on Story Structure and the Organisation of Participation. In Atkinson & Heritage (Eds) 1984: 223-4.

Lantolf J. P. & Frawley W. (1988). 'Understanding the Construct' Studies in Second Language Acquisition, 10: 181-195.

Lazaraton A. (1992) A Conversation Analysis of Structure and Interaction in the Language Interview. PhD Dissertation, UCLA. Analysis of 20 ESL Course Placement Interviews.

Lazaraton, A. (1996) Interlocutor Support in Oral Proficiency Interviews: The Case of CASE. Language Testing 13/2: 151-172.

Ludwig J. (1984) Native Speaker Judgements of Second Language Learners' Efforts at Communication: A Review. Modern Language Journal, 66:274-83.

Nunan D. (1988). The Learner Centred Curriculum. CUP.

Perrett G. 1990. The Language Testing Interview: A Reappraisal. In De Jong J H A L & Stevenson D K (eds) Individualising the Assessment of Language Abilities Multilingual Matters, Clevedon Avon: 225-38.

Pollard J. D. E. (1994). Paper on Proficiency Testing. IATEF Testing Newsletter July, 1994.

Ross S. & Berwick (1990) The Discourse of Accommodation in Oral Proficiency Interviews. Seminar in Program Evaluation and Language Testing, Singapore. Anthology Series SEAMEO Regional Language Centre (RELC), Singapore.

Ross S. (1992) 'Accommodative Questions in Oral Proficiency Interviews.' Language Testing 9/2: 173-186.

Selinker L. (1972) Interlanguage Language Learning.

Shegloff E. A. (1995) Discourse as an Interactional Achievement III: The Omnirelevance of Action. In Jacoby & Ochs, Eds. Special Issue on Co-construction. Research on Language and Social Interaction 28/3: 171-83.

Spolsky B. (1995) Measured Words. Oxford University Press.

Underhill N. (1987) Testing Spoken Language. CUP.

Van Lier L. (1989). Reeling, Writhing, Drawling, Stretching, and Fainting in Coils: Oral Proficiency Interviews as Conversation. TESOL Quarterly, 23/3.

Van Lier L. (1996). 'Interaction in the Language Curriculum: Awareness, Autonomy and Authenticity '. Longman.

Woken M. & Swales J. M. (1989) Expertise and Authority in Native-nonnative Conversations: The Need for a Variable Account. In Gass, Madden, Preston & Selinker (Eds) Variation in Second Language Acquisition: Discourse and Pragmatics. Multilingual Matters.

Young R. 1996. Conversational Styles in Language Proficiency Interviews. Language Learning 45/1: 3-42.

Young R. & Milanovic M. (1992) Discourse Variation in Oral Proficiency Interviews. Studies in Second Language Acquisition 14:

Zuengler J. (1993). Encouraging Learners' Conversational Participation: The Effect of Content Knowledge. Language Learning 43: 403-32.

Test Writing for UAE Distance Learning Students

LISA BARLOW AND CHRISTINE CANNING
United Arab Emirates University

1.0 Introduction

This paper will address the *feasibility*, and or difficulties, involved in producing multiple versions of exams administered during a three-week period to students in a United Arab Emirates University Distance Learning Program for English. The challenges faced in the production of these exams included: first, how to equalize and maintain *reliability* of test items and texts for up to seven versions of exams. Second, how to modify, yet retain *validity*, when writing up to seven multiple versions of the same exam from a limited base of course objectives. Third, how to vary test items and text subject matter, yet safeguard *reliability* and *test security*, so that students who take the exams in the first few days, do not effectively disseminate format and possible variation of test items to students in centers with later test dates. Lastly, in keeping with University General Requirements Unit (UGRU) guidelines, test writers were challenged with how to produce culturally sensitive, but interesting and applicable, test material for the Gulf Arab students. In addition to the challenges discussed in the paper, a comparative-descriptive statistical analysis was completed to assess the overall success or failure of the final exam testing periods for the Fall 1996 English Levels 1,2 and 3 final exams. In the analysis, 312 student test scores were examined.

1.1 Introduction to the UAE Distance Learning Program

What is the Distance Learning Program (DLP) in the UAE? The program was created in 1982 by the UAE University Faculty of Education. Nine centers were designated at the following sites: Abu Dhabi, Dubai, Sharjah, Ajman, Fujairah, Um Al Quwaiin, RAK, Al Ain and Merfa. The Centers were created for Emirati elementary and secondary school teachers to upgrade their Associate of Arts Teaching certificate to a Bachelor of Arts in Education. The Merfa Center is the exclusive exception due to geographical limitations which prevents students living near the Qatar border and on the UAE owned islands in the Persian Gulf from attending university in Al Ain due to transportation purposes. Thus at the Merfa center, the students are regular 5-year university students.

It was not until Winter 1996 that the UGRU English Unit began its involvement with the DLP. It is important to note that UGRU involvement means faculty work with the program design, writing and implementing curriculum and test, yet faculty are not included in any administrative decision making.

The DLP in the UAE has an older student population (with the exception of the Merfa Center) who juggle work, home and family responsibilities. Thus, the UAE DLP entails all the problems that a regular Adult Education program would include. However, the incentive to achieve is high. Completion of the advanced degree of a BA through Distance Learning would entitle the student to not only a substantial salary increase, but also a rise in position and prestige in the workplace. In turn, these incentives, along with time restrictions for completion by the university, influence the student to participate in the program with a well-defined goal to su:ceed.

2.0 Theory of Distance Learning

There are many terms in use to denote distance education: correspondence study, independent study, external study, and distance teaching. There terms have been used to name particular ventures which have been set up over the past forty years, and often overlap. Distance education has been introduced in many countries for the purpose of meeting certain demands and of achieving certain goals. It can be a way of bridging a gap between the growing number of people who want or need education and the limited resources of conventional education.

The theory and definition of distance education is drawn from observation of the many new institutions which have sprung up in the last forty years. One of the first of these was the United Kingdom Open University which has served as a model for similar projects in many other countries.

Perraton (1981, p.13) defines distance education as an educational process in which a significant proportion of teaching is conducted by someone removed in space and/or time from the learner. This definition focuses on the most striking difference between conventional education and distance education, that is, the separation in space of the student from the teacher and the freedom in time which the student enjoys.

Dressel and Thompson (1973, in Wememeyer, 1977) define independent study as a self-directed pursuit of academic competence in as autonomous a manner as a student is able to exercise. For the student, the capacity to study independently is a virtue and a major goal of education including conventional education.

Wedemeyer (1977, Pp.2114-2121) notes that in the USA the use of the term independent study links distance education with developments in conventional education. He defines independent study as those arrangements in which teacher and learner carry out their essential tasks apart from one another, though they communicate in a variety of ways. He also emphasizes the role of the distance learner as well as mentions the various ways of communication by which the distance between learner and teacher is bridged in a course of study.

For the purpose of this study, distance learning implies a pedagogical method where oral teaching is limited and concentrated to a few intensive periods (16 hours) spread out over a 16-week semester. In between these periods, the student studies on his/her own at home, without the possibility of consulting teachers by phone or letter.

3.0 Current Trends in Worldwide Distance Learning Programs

In one form or another, DLPs are present in nearly all nations that have a regular university program. Perhaps the most common feature of DLPs in many countries which organize distance education, is that the majority of the students are teachers who wish to better their qualifications. Teachers have traditionally been the majority of learners at the Open University in England (Gratton and Robinson, 1975). In a report on DLPs in Australia, two-thirds of these students were teachers (Dahllof, 1976). In the UAE DLP 100% of the students are teachers, with the exception of the Merfa Center whose students are beginning university students studying full-time to become teachers.

If we are to adhere to the working definition of the UAE DLP as outlined above, perhaps the major difference between traditional forms of DLPs and the UAE DLP is the use of media or communications. In traditional DLPs post WWII, the British and American DLPs used radio, television and satellite to enhance DLPs. Later, the telephone became one of the most important aids in distance learning. In DLPs, the telephone is used for conferencing between student and teacher. Daniel and Turok (1975) write that for the telephone to be used effectively, it should be used for both actual teaching and for feedback to the students. Also, fixed consultations are set up between student and teacher to talk about the course, syllabus and assignments, and ask questions. Answering machines have also used so that students can ask questions outside regular working hours.

In the 1990s, the use of the telephone in DLPs has taken a back seat to the Internet. In Iceland, one unified network for communication is based on the Internet. Nearly all schools in the country are connected to this net. This allows an open communication between the student and teacher, as well as student and student. The Icelandic program allows student work to be dealt with by the teacher within 24 hours of receipt, if possible, and sent back to the student. The Internet has made it possible for students not to have to leave their places of work or their families for schooling and hence limiting the inconvenience of DLP on campus commitments. In turn, this aids in the completion of the course with the specified time limit (Agustsson 1997) A similar Swedish DLP states that the Internet has increased both their enrollment and student success because the Internet is a very feasible. It uses methods which are inherently simple, which are comparatively inexpensive, which work, and which can be easily mastered by the common individual (Agustsson, 1987).

On the whole, communication for these programs takes place via the network. Seminars are conducted on particular modules of the course over the network. The seminars can be held for single groups of students or for all students enrolled in the course, or for the entire network which gives students access to debate and current course information. The data network allows students access to various electronic databases both in Iceland and abroad. In addition to its wide base of uses, the Internet is also used for e-mail. Through e-mail, practitioners send instructional guidelines, send deadline reminders, point out errors and give guidance to DLP participants. E-mail also ensures teacher/student discussion is preserved in the computer, so the student always has access to what was previously written (Jeppesen, 1987).

3.1 Common Concerns of Worldwide Distance Learning Programs

The most acute concern of DLPs, from an international point of view, is the question of dropouts. High statistical dropout rates have been reported by many DL programs. Perhaps it is connected to the problem which concerns the motives adults have for enrolling for DLPs. In an attempt to clarify this problem Lampikoski (1975) asks which factors can be used to arouse interest in distance education? He believes the solution is found by demonstrating a greater need for adult education, a greater need of knowledge to cope with the rapid changes in society, a greater equality in education, and a greater amount of leisure time. If student interest is aroused and the realization of the importance of the program sets in and his enrollment in a DLP, then perhaps the dropout rate would lower. Another possible incentive to decrease dropout rates would be the use of the Internet as described above. For example, Lampikoski believes that the student has easy access to communication with the course instructor and faster feedback on questions and work, then student interest would be retained and student dropout rate would decrease.

Unlike other foreign DLPs, the UAE Distance Learning Program has incentives to ensure interest and completion of courses. Economic advancement in salary, rise in position in the workplace, prestige among colleagues and teaching districts, and a well-defined time limitation to complete all courses within the program. These reasons have curbed a high student dropout rate that is often lamented over in other DLPs.

3.2 Testing Forms In Distance Learning Globally

Although the literature does not address the types and formats of testing in DLPs, it does outline the administration of these exams. Exams, for the most part are given in one of the following ways. First, exams are sent by mail to schools that have undertaken to hold the examinations simultaneously for the DL students. The school guarantees that the examination is taken under the supervision of staff, and then sends the papers back to the university, where grades are awarded. Results are either mailed or e-mailed to students. Second, a written exam is administered at the University at the end of each semester. Another alternative, for students taking the course over computer is to be e-mailed the exam and take the exam at home. A last common alternative, is an oral examination taken at the university or over the telephone.

Each of these methods of examination of DLP students has its advantages and disadvantages. The method the UAE DLP utilizes is to administer the exams at each of the DLP Centers by staff during a two-week exam period. The challenges of which constitute the remainder of this paper.

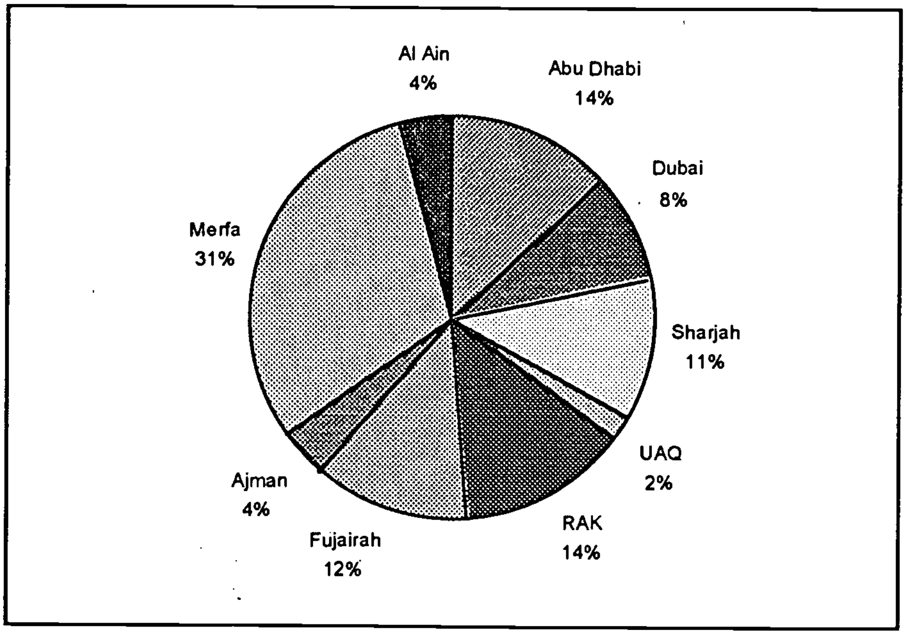
4.0 The Stratagem Behind the UAEU Distance Learning Program

The UAE DLP courses are sixteen hours in length. Lectures are divided over 16 weeks into eight two-hour seminars. Unfortunately, communication with the teacher is restricted between this two-week period. Only one full-time teacher on staff retains office hours one day a week, who may not teach the course the student is registered in. Student access and interaction with this teacher is limited by distance and time availability. UAE DLP students are not privy to Internet due to censorship restrictions placed upon the learner by the university.

Other internal guidelines for the UAE DLP include: students being required to attend six of eight lectures to pass a course. All course work must be turned in by assigned time and dates. Written assignments are turned in twice: once for teacher input, and a second time by the student with corrections and revisions completed. Two quizzes are given during the third and seventh lectures, a fifth week midterm, and a final exam upon completion of the eighth. Distance learning students complete examinations, at their DLP Center, within the same time frame as regular UAE University students. All UAE University student rules and guidelines apply to the DLP student population.

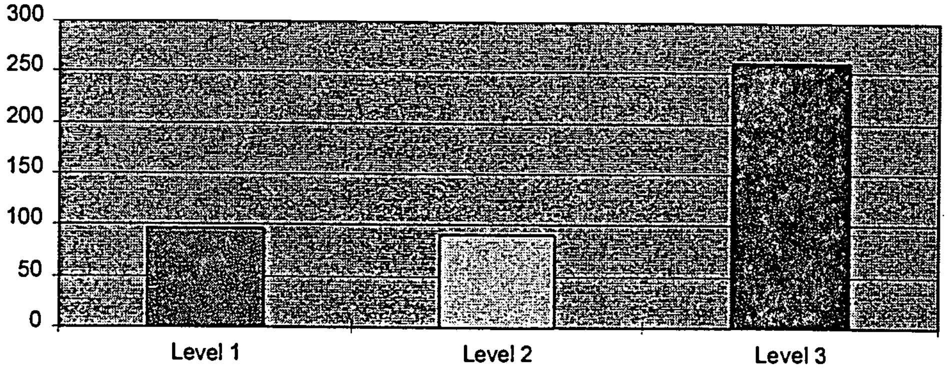
4.1 Centers

Each of the seven DLP centers is run by a full-time director who undertakes all program administrative and teaching responsibilities. The following indicates student group breakdown according to size and course. The majority of students are registered at the Merfa Center. As mentioned earlier, The Merfa Center enrolls regular 5- year university students, who are bused in from islands off the coast and villages up to the Qatari border.



4.2 Student Population

The semester under evaluation is Fall 1996. The total number of students was 312. The total for English Level 1 was 59 students, Level 2: 60 students, and Level 3: 193 students.



Who are our students?

Fall 1996

4.3 Curriculum

The curriculum follows UGRU English Levels 1 and 2, with the exception of the listening component. The objective of the UGRU English Program is to provide learning materials, activities and environments which promote mastery of related English language skills (listening, speaking, reading, writing, and structure) according to Giannotta (1998). English Level 3 departs from the UAE curriculum in that it focuses heavily on grammar, developing reading skills and paragraph writing. Each of these Level 3 skills are taught in a content-based unit. A listening component has been deleted from the curriculum. To compensate for not teaching or testing listening due to time constraints, cassette tapes and exercises are provided and marked for the students, but they are not tested on this skill.

5.0 The Four Challenges

Challenge 1

The first challenge faced by test writers during the academic 1996 year, was how to equalize and maintain *reliability* of test items and texts. According to Coombe and Hubley (1998), a test is considered reliable if it yielded similar results if it were given at another date and location. That is, will the test function in the same way at each Distance Learning Center? Also, will the score gained approach the "true score" of the examinee each time it is given, and in a consistent way with different examinees? For example, when Fatma in Fujairah sits down and takes the test, will her score be as *true* as Alia's in Abu Dhabi?

Sullivan and Higgins (1983) claim that the most common cause in the production of unclear assessment questions involves stating the items in a matter that permits more than one interpretation of what is being tested. One way we try to ensure reliability of grammar sections of exams is to write one model sentence and then vary it by one word 7 to 10 times, depending on the number of versions:

Example: I __________ to the store yesterday. (go)
She __________ to the store yesterday. (go)
They __________ to the store last week. (go)

A second way we attempt to ensure reliability of reading texts is a technical one. Each text is run through one of two computer programs Flesch-Kincaid or Right Writer. This way whether a reading text is on Mexico or Australia, students at each center will face a text that has been technically measured to be at the same reading level.

Challenge 2

The second challenge is how to modify, yet retain *validity*, when writing multiple versions of the same exam from a limited base of course objectives. By validity, as defined by Brown (1987), we mean, does the test measure what it is intended to measure? For example, there are a limited number of grammar points taught in English Level 2. With this minimal number of grammar objectives how do test writers produce questions that are valid, based on course content, but not predictable?

When a regular university 16-week course, which was written for more than 140 contact hours a semester, is reduced to one that meets once every other week for a total of 8 hours, a bare-bones curriculum left. Many target and course performance objectives had to be revised and cut to meet student needs, yet meet accreditation standards.

For example:
Tense Markers for simple present and simple past
He always __________ to class on Mondays. (go)
Today, it __________ is hot, yesterday it was warm. (be)

In what ways could these questions be answered? Can the *nuances* of English be taught and tested in this DL course with such a limited time frame? Would not testing for these nuances be considered as teaching/testing a bastardized English? In the end, to retain *validity*, only course curriculum taught to E1 and E2 was tested. Questions were not written to test the possible nuances of English.

Challenge 3

The third challenge is how to vary test items and text subject matter enough, yet safeguard *reliability* and *test security*, so that students who take the exams in the first few days, don't effectively disseminate format and possible variation of test items to students in other Distance Learning Centers. As Coombe and Hubley explained, security is part of both reliability and validity. Further, they suggest that cultural attitudes *toward collaborative test-taking* are a threat to test security and thus to reliability and validity. DLP students have been known to contact students at other centers, immediately after taking the test, and report verbatim what was on the exam. Thus, for students who have the exams on days 6, 7, and 8 a reliable pool of test questions have been memorized.

For example, in the Spring 96 semester DLP faculty produced 12 A and B versions for the EL2 testing dates. To alleviate this challenge of 8 separate test dates for EL3, faculty questioned whether it was necessary to write 8 A and B versions to ensure test security. It was found that statistically only 5 A and B versions were necessitated if administered by mixing versions. In this way, it was reasoned, each center would not be able to predict which version would be administered.

Finally, as mentioned previously, the UGRU faculty has no administrative control. This means we do not proctor our own exams. The DL faculty assigned to the center does. Problems arise from this method. The biggest security risk concerns itself with cheating due to inadequate space and number of proctors.

Challenge 4

The last challenge focuses on the production of stimulus material which is not culturally offensive to the examinee so as not to distract attention from the task. That is, in keeping with University General Requirement Unit concerns, how to produce culturally sensitive, but interesting and applicable, test material for the Gulf Arab students. As with UGRU students at UAE University, DLP courses avoid topics of religion, sex, politics and music. In addition to that, however, to be culturally sensitive we must be aware of the Gulf Culture itself. That is, test items and texts need to be reflect an awareness of tribal/family alliances, political and social tensions which are area specific to the Gulf, and an intercultural hierarchy.

For example:

A. The Shamsee tribe is larger than the Mansoori tribe (Merfa)

B. Kuwait has a stronger football team than the UAE. (World Cup Playoffs)

C. Careers Grid: The career of a police officer does not require university education, the work is dangerous, and the salary is low.

In example A, it is important to note that the Al Shamsee tribe and the Al Mansoori tribe have been warring for more than 100 years. Additionally, the majority of students at the Merfa center are Al Mansoori. At other DLP centers this is not a problem. Example B hits on a competitive point between Gulf nations. During the recent World Cup playoffs Kuwait beat UAE to go to the finals. Thus, because soccer is the number one sport here, reactions to such a statement could have caused an emotional response and taken the student off task. A last example of a culturally insensitive question was written in a career grid. The grid referred to the job of a policeman as not requiring a university education, dangerous, and has a low salary. Because the majority employers in the UAE are the police and military, the job of a policeman is considered to be of a high status. Something that requires special training and in fact pays a higher salary than many other jobs.

These test writing errors in cultural sensitivity could be considered obvious. Yet when multiple versions are written, attention to other factors such as validity and reliability seem to require more attention and cultural sensitivity is often overlooked.

6.0 Statistical Analysis of UAE University Distance Learning Test Scores

Finally, in an attempt to quantitatively reflect on the success or failure of the Fall 1996 final exam period, a comparative-statistical analysis was completed. A comparison method using descriptive statistics was used to identify the similarities and differences between the final results obtained between UAE DLP students studying English in the Level 1, Level 2 and Level 3 programs. A total number of 312 student results were examined and subjected to statistical analysis. A total of 59 English Level 1 student scores, 60 English Level 2 student scores, and 193 English level 3 student scores were subject to review over two semesters beginning in the Fall of 1996.

All students were placed into levels based upon their scores on a standardized placement exam written for Gulf Arab Learners. Students placed into English Level 1 had scores from 0 to 35. Likewise, students placed into English Level 2 received results ranging from 36-59 and English Level 3 from 60-100.

After a statistical review of the total student population in the DLP as a whole after course completion, 10% of the population passed with a final grade score of 90% or higher; 18% of the population passed with a score within the 80-89 percentile; 25% of the total population passed with a score within the 70 to 79 percentile; 26% of the total population passed with a score within the 60-69 percentile; and 21% of the total population failed and were subject to repeating the course.

Mean averages were devised and calculated in point value out of 40 as well as out of a percentage score of 100. The English Level 1 mean exam score was 27.15 out of a total of 40 points. This averages out to a percentage score of 68% based on the 100 point scale. Likewise, the English Level 2 mean score was 27.43 out of a total of 40 points. The average mean percentile working out to be roughly 69%. Low mean scores could be attributed to lack of exposure to English language, exposure to various dialects of English language, exposure to pidgin English which can be directly attributed to the fact that 80% of the country is run by expatriates from various countries and nations who use English as a medium of communication, and to lack of class interaction since the number of contact hours range from 14-16 hours in an entire semester. In contrast, English Level 3 scores showed that the mean average was in the 70th percentile with an average mean value of 28.17 out of a total of 40 points. It can be hypothesized that English Level 3 students scores were typically higher because before starting the program their exposure to English prior to the placement exam may have been more than those found in the two lower levels.

A total number of 12 E1 students, 16 E2 students, and 38 E3 students failed the final exams. The pass rate for the entire course at each level was significantly higher than what the mean scores from the final tests might seem to indicate. This is because a 35% teacher grade and a 25% midterm grade were awarded to students, in addition, to final exam scores. Because of these two influencing factors, E1 had an 80% pass rate and a 20% fail rate, E2 had a 73% pass rate and a 27% rate of failure and E3 had a 81% pass rate and a 19% rate of failure.

Although, students overall grades for the course may have been higher than their exam grade due to instigating factors mentioned previously, final exam scores in each level of letter grade assigned differed from the overall results. For example, a student with a failing final exam grade of 10/40, a failing midterm grade of 15/25 and a teacher grade of 35/35 could pass the course with a 60% without passing the valid and reliable exams. This is a challenge that the program is working to overcome.

However, in the scope of this study, the following final exam grades were awarded. The number of exams scoring higher than 90% were as follows: E1 5% (3 students); E2 8% (5 students); and E3 10% (20 students). The number of Bs: E1 11% (7 students); E2 20% (12 students); E3 27% (53 students). The number of Cs awarded to final exam scores resulted in: E1 18 % (12 students); E2 20 % (12 students); E3 19.6 % (38 students). The number of Ds awarded for examinations were as follows: E1 42% (25 students); E2 25% (15 students); and E3 22% (44 students). The failing examination population for the levels broke itself down in the following manner: E1 20% (12 students): E2 26% (16 students) and E3 19.6% (38 students).

The standard deviation presented in figure 1.1, shows the list of numbers is spread out around the average. The spread which is usually measured in quantity, reflects that the standard deviation scores were respectably reported at the following levels: E1 stdev (5.13), E2 stdev (6.98), E3 stdev (6.15). An alternative check was made to check the standard deviation scores by taking the average number of entries squared and subtracting the average of entries squared. Although, an *idealistic* standard deviation would or should be somewhere in the range of 1 or 2 degrees away from the mean, the standard deviations found were not as perfect as hoped. However, the standard deviations were extremely favorable as they bared close relation to the mean averages. STDs served as a shadowing indicator of how scores listed were closer to the statistical mean averages which the DLP has tried to achieve through writing valid and reliable tests.

Class	n	mean	trmean	median	stdev	min	max	Q1	Q3
E1f96	59	27.15	27.07	27	5.132	14	40	24	29
E2f96	60	27.43	27.75	28	6.988	0	39	23	33
E3f96	193	28.17	28.3	28	6.15	12	40	24	33

Figure 1.0

A summary of the data for the English Distance Learning students is represented in graph histogram (see figure 1.2, 1.3, 1.4). It is important for interpretation for the reader to understand that a histogram does not need a vertical scale to be interpreted; therefore, all class intervals will be presented in a horizontal format. It must also be noted that histograms have the ability to simply ignore total areas of information, however, the histogram representing the final scores for all levels of the DLP are accurate representations of the current data available. When plotting the percentages, longer blocks of class intervals were not represented as wholes; rather, individual percentages of class intervals were recorded.

The histograms used for this report represent the distributions as if the percents were spread evenly over each class interval. This was done by figuring out the original height of the block over each class interval and then dividing the percent by the length of each interval.

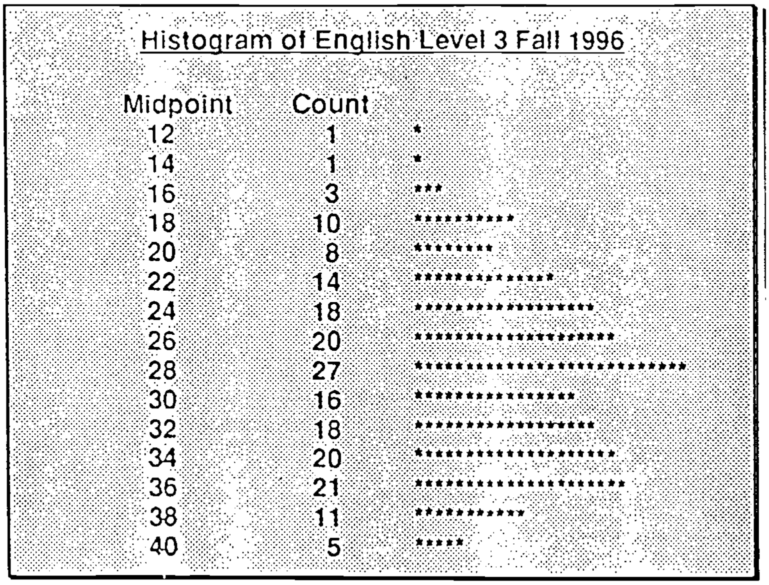
The histograms displayed in the figures, show the relationship between the average and the median. Figure 1.1 of English 1 shows a strong correlation between the average class scores and the median. However, the scores at either end of the scale show an uneven distribution which could have resulted due to a number of variables that have been discussed earlier in this paper. Figure 1.2 shows that scores at the midpoints between 20 and 35 range roughly around a center score of 27 with the majority of average class scores falling within the respective area. Again, the histogram for English 2 shows an uneven spread of averages that appear to be top-heavy showing that the majority of students passed with a score of 50% or higher. This again could be explained by the different final score compilations discussed earlier in the paper. Figure 1.3 demonstrates that the midpoint of 28 correlated the strongest with the average class interval score of 27. The histogram shows a more even bell-curve distribution of scores being displayed. The differences between the English 3 histograms versus those presented for the two other lower levels can be attributed to wither the number of students in the sample population or to exposure of repeated examination information either learned in previous courses or in the present course materials; of course, it could just as well as been a combination of both or other varying factors associated with DLPs, especially a program such as the one in the UAE where there are so many teachers of different nationalities teaching on many different sites.

```
Histogram of English Level 1 Fall 1996

Midpoint              Count
  14             1    *
  16             0
  18             1    *
  20             2    **
  22             7    *******
  24             5    *****
  26            10    **********
  28            17    *****************
  30             4    ****
  32             2    **
  34             3    ***
  36             5    *****
  ß3             0
  40             2    **
```

Figure 1.1

```
Histogram of English Level 2 Fall 1996

Midpoint      Count
  0             1     *
  5             0
  10            0
  15            2     **
  20           11     ***********
  25           15     ***************
  30           15     ***************
  35           13     *************
  40            3     ***
```

Figure 1.2 EL2



```
Histogram of English Level 3 Fall 1996

Midpoint    Count
  12          1    *
  14          1    *
  16          3    ***
  18         10    **********
  20          8    ********
  22         14    **************
  24         18    ******************
  26         20    ********************
  28         27    ***************************
  30         16    ****************
  32         18    ******************
  34         20    ********************
  36         21    *********************
  38         11    ***********
  40          5    *****
```

Figure 1.3 EL3

Finally, to ensure reliability, validity and test security of exams, tests given at different intervals were scrutinized by comparing and contrasting mean scores. Empirical proof from closer examination of test scores showed by comparing multiple repeated exam versions administered on different days that mean scores did not raise above one letter grade. The most significant proof was evident in the English 2 course, where men and women from the same emirate had the exam within a nine day lapse of time. Mean scores were virtually the same. Men scored a mean of 65% and women a mean of 60%. A third test was scrutinized from Sharjah emirate because their examination period fell equally between the two RAK testing periods for the same exam. The mean score was 80% which is a significant difference, as it poses a letter and half grade difference between the RAK men's score and the two letter grade difference between the RAK women's scores. Therefore, test reliability and security were maintained between the two exams given to the same distance learning area between the men and the women in RAK, but the high scores in Sharjah indicate that the women did significantly higher. Possible explanations for no difference between the men's and women's scores could be attributed to the culture of the society as education is sex-segregated even at higher levels of education. A possible reason for Sharjah students scoring higher on the English exams could have resulted from the high number of English speakers who work in the area where the students study. It is also possible that students from Sharjah were stronger language students or who may have had better teachers or received better teaching methods. The Sharjah teachers were female native speakers of English. Two male teachers, one of who was new to the program and the other whose first language is Arabic, taught the courses in RAK because of restrictions placed by the Ministry of Education banning female instructors from the University from teaching either male or female students.

7.0 Future Directions in Testing UAE Distance Learning Students

In the future, DLP testing should begin to follow UGRU testing procedures more closely. It is also believed that one unified exam date should be mandated. Additionally, class teachers should be able to co-procter exams for two reasons. First, to ensure valid student questions on exam procedures are clear. Second, to uphold UGRU testing policies. Because of the limited class time, perhaps alternative forms of testing could be implemented in future semesters. Another option is to pilot test formats.

8.0 Bibliography

Agustsson, H. 1997. "The Distance Education Program of Verkmenntaskolinn at Akureyri (VMA)" in *Educational Media Instruction Journal* 34: 2, 54-56.

Al Rawaf, Haya Saad. 1996. 'An Open University for Women in Saudi Arabia: Problems and Prospects.' PhD Dissertation, Loughborough University of Technology.

Brown, H.D. 1987. *Principles of Language Learning and Teaching*. Prentice-Hall: Englewood Cliffs, New Jersey.

Coombe, C. and N. Hubley. 1998. 'Creating Effective Classroom Tests.' Presented at TESOL International, Seattle Washington.

Dahllof, U., J. Lofgren and B. Willen. 'Evaluation, Recurrent Education and Higher Education Reform in Sweden.' Three papers to the 1978 Lancaster Conference on Higher Education.

Daniel, J.S. and B. Turok. 1975. 'Teaching by Telephone.' In *Ljosa* 133-140.

Dressel, D.L., M. M. Thompson. 1973. 'Independent Study: A New Interpretation of Concepts.' Quoted in Wedemeyer, C. 'Independent Study' in *The International Encyclopedia of Higher Education*, 1977 ed. San Francisco: Jossey-Bass.

Evans, T. and D. Nation, ed's. 1989. *Critical Reflections on Distance Education*. Falmer Press, Philadelphia, PA.

Giannotta, F. 1998. 'Curriculum Design' Draft for Comments. UGRU English Program, UAE University.

Gratton, P. and J. Robinson. 1975. 'Educational Broadcast: The Search for Priorities'. In *Ljosa*: 92-100.

Jeppesen, K. 1997. 'Distance Education in the University College Education, Iceland.' *Educational Media Instruction Journal* 34: 2, 57-59.

Lampikoski, K. 1975. 'Some Key Reasons an Individual Becomes Interested in Study by Distance Education. Paper to the European Home Study Council, Autumn workshop, Helsinki.

Perraton, H. 1981. 'A Theory for Distance Education.' Prospects. XXI (1), Pp.13-24.

Schley, N. and Canning, C. 1998 "Writing Effective M/C Tests", EMCEE, 4:2 P.7-8.

Schley, N. and Canning, C. 1998 "Good Testing Practices for Computer Based Math and Computer Courses Taught in a Foreign or Second Language", TESOL Arabia News, 5:4. Pp.16-17.

Sullivan, H. and N. Higgins. 1983. *Teaching for Competence*. New York, New York: Teachers College Press. Columbia University.

Wedemeyer, C. 1977. 'Independent Learning' in *The International Encyclopedia of Higher Education*. San Francisco: Jossey-Bass.

Willen, Brigitta. 1981. *Distance Education at Swedish Universities*. Uppsala Studies in Education.

Student Created Tests as Motivation to Learning

PHIL COZENS
Higher Colleges of Technology

This paper will look at classroom tests and, particularly, the involvement of learners in the creation of tests in order to increase motivation and lead to some instance of autonomous learning. The tests, themselves, are rather informal, but act as catalysts to 'something' else which I will try to define. It also looks at the concepts of 'face' with relation to the work of Scollon and Scollon in interethnic situations and Quirke who looks at the EFL classroom. It examines briefly student/teacher roles as they appear to be perceived by UAE students and how these perceptions can be used covertly to encourage student autonomy.

Heaton (1975) states, "Both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other." A statement with which most language teachers would agree. Tests according to Madsen (1983 P.4) "can help create a positive attitude to classes" and benefit students "by helping them master the language." Hughes (1989 P.1), unfortunately, does not appear to fully agree when he states in his opening paragraph, "Too often language tests have a harmful affect on teaching and learning; and too often they fail to measure accurately whatever it is they are intended to measure." Both, as do many others (Heaton, 1984; Rivers 1983 P.141) emphasize the importance of the backwash, or washback, effect on both teachers and learners. This, as Hughes (1989) points out, can be either beneficial or harmful, especially "if the test content and testing techniques are at variance with the objectives of the course" a situation which, unfortunately, often occurs. Carroll, quoted in Rivers, also warns of this when he states:

> The kinds of tests that are used to evaluate students can often have an adverse effect on students' learning....It is only natural for students to shape their learning efforts so as to be maximally successful on tests, and if the tests measure objectives that are in some ways different from those of the instruction, students will work towards those objectives and pay less attention to achieving other objectives.

Heaton (P.1) appears to concur when he states in his opening sentence "It is unfortunate that so many examinations have led to a separation of testing from teaching." He does, however, go on to say, "Tests may be constructed primarily as a device to reinforce learning and to motivate the student." Hughes also states that "testing should be supportive of good teaching and, where necessary, exert a corrective influence on bad teaching."

It is in this supportive role which classroom tests and, in particular, student created tests have an important role to play. Nunan (1988) has chosen to call this type of test 'micro-evaluation' which "is conducted at classroom level and involves teachers and learners," Heaton, when talking about classroom tests states "A test which sets out to measure a student's performance as fairly as possible...can be effectively used to motivate the student." This motivation can also occur when students are involved in the creation of tests for each other and appears be in line with Nunan (P.118) who states that both learners and teachers should be involved in the evaluation process as "Learners need to assess their own progress," and goes on to say:

> It is also argued that self-assessment by learners can be an important supplement to teacher assessment and that self assessment provides one of the most effective means of developing critical self-awareness of what it is to be a learner, and skills in learning how to learn.

He reinforces this when he adds:

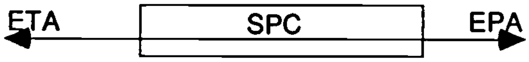
> ... the articulation of goals and assessments has particular value in a learner-centred curriculum, in which one set of goals will relate to the development of meta-cognitive skills on the part of the learners

This development of meta-cognitive skills, or learning how to learn, is one aspect of more independent learning which we are trying to address with the student created tests.

A barrier to this development, however, would seem to be students' perceptions of student/teacher roles in the classroom. There is, a definite preference for the teacher as "knower/informer" with students as "information seekers" situation posited in Ellis (1994), but the "seeking" does not seem to require any active participation. It is the product only which is important, be that a mark or proficiency band, not the process of obtaining it, although this may be the result of "a system which values scores and rank orders over actual success in learning."(Dickinson 1987 P.151). It may also, as Quirke suggests (1997 P.68), be a result of the importance of 'group face' to students in the EFL classroom "where all the personal needs are similar." and students "are often reluctant to put forward their own views in case they break the group's equilibrium." This equilibrium seems to result in some solidarity of purpose, and expectation, which encourages our students to believe that being in the classroom is the only effort required.

Scollon and Scollon (1990 Pp.167-169) have used a solidarity politeness - deference politeness cline to show how (Distance/Power can influence the 'face' of individuals and how asymmetric power relations can lead to misinterpretations of intent. In normal classroom situations students find themselves in a nominal - D(istance) - P(ower) situation which leads to a solidarity politeness position which, of course, poses no danger to either the individual or group face of the learners. The relationship with the teacher will often present itself as a +D +P situation which leads to deference politeness where, unless care is taken on the part of the teacher, both the individual and group 'face' can seem to be threatened. This can usually be averted, as long as pre-determined roles (knower - seeker) appear to be accepted by all parties. Face, after all, as Owen (1990 P.258) points out, "is a property of people living together in society, rather than a property of their own make-up."

Quirke (P.69) reinforces this when he states "Every type of social interaction is affected by, among other things, face" and that, "the classroom is one form of social interaction." Willis (1990 P.13) reiterates this when he states. "...a language lesson is a social event. There is more to it than simply learning a language." Face, in the classroom needs, therefore, to include everybody's interpersonal needs and seems, as posited earlier, to result from pre-defined expectations. Renaming the ends of the cline posited by Scollon and Scollon to model students' expectations with regard to student/teacher roles and their effect on face we find that all the extremities remain stable, the central area has some similarities which can lead to loss of the individual 'face', but not, surprisingly, the 'group' face. Students expect teachers to find errors. It is, after all, part of the responsibility of the 'knower', particularly if s/he is a native speaker, to find and highlight errors There is, therefore, no loss of face when this occurs. Peers do not, in most cases, evaluate each others' work and, therefore, the 'group face' and, in monolingual classes, 'cultural face' ensures the individual 'face' of the learners. We have, therefore, a cline of Expected Teacher Action (ETA) to Expected Peer Action (EPA) where learners feel 'safe' at both ends. There appears, however, to be a central area where students themselves have a new role; Student as Peer Critic (SPC).



In this area students are freed from the restrictions imposed on them by the 'group', particularly if they are functioning as part of a smaller 'group' within the whole., and as such they are now free to take on some of the responsibilities of the 'knower.' They must, of course, have been informed, in advance, of what they, and others, are 'to know', and not be the only ones responsible for knowing. It is the peer interaction of working in small groups within the whole group which allows for a change in role without loss of face and also allows involvement in a process rather than product prominent activity, although for the students, themselves, it is the product, i.e. the finished test, which is the important feature.

It was stated earlier, that UAE students take a product oriented attitude to learning, and, therefore, do not feel obliged to take a great deal of interest in the process, particularly as this not a simple procedure, highlighted in O'Malley and Chamot's statement (1990 P.216):

Two of the most important characteristics of procedural knowledge are that it is difficult to learn and it is difficult to transfer to new situations.

For this reason, the concept of student created tests allows the teacher to highlight a 'product' which the students are to produce, a goal with which they will happily comply, while, at the same time, encouraging the peer interaction which, hopefully, will have an effect on the process and meta-cognitive abilities of the students and motivate them at the same time.

The concept of peer interaction as a motivator for learning is not new, "Tutunis (1995) describes a writing project where students learned to improve their writing skills by writing for peers. thereby "writing with a reader (other than the teacher) in mind." and, in so doing, had "learnt more about writing through reading other papers critically." Bassano and Christison (1995) describe an activity where student-created visuals act as stimuli for peer interaction in the classroom context. Rea-Dickins and Germaine (1992 P.48) describe a task where learners in pairs or small groups examine each others' written work. One result of this as they point out is, "It encourages the development of a critical approach to written work..."This 'critical approach' results from the interaction encouraged by the tasks which all seem to focus on the process whilst having a product in mind. It is felt that this interaction and, more importantly, learning can also occur when students write tests for each other.

Harris (1969 P.13) points out that good tests have three qualities: validity, reliability and practicality, in other words they "must be appropriate in terms of our objectives." A student created test will probably not have a high reliability, but as Harris points out:

> ... for the usual classroom test, prepared by the teacher and used but once, the computing of reliability coefficients is scarcely ever practicable.

This, therefore, does not cause any major concern. The validity of the test will be partially ensured in two areas: content validity, in that the area of content has been set by the teacher and face validity, in that as the students, themselves, have set the tests they believe that it is a valid test. Those items not achieving the required level of validity being challenged by other members of the class. Criticisms of each others' tests is, in fact, an important part of the autonomous learning activities which often take place. Practicality is ensured in the current situation, in that most of the tests are administered either through simple CALL programs which are easy to use, or pencil and paper tests produced on a word-processor and administered in the classroom.

The concept of student created tests occurred as a by-product from a Pre-Intermediate class activity where students were each researching some aspect of life in Ras Al Khaimah. Besides other tasks, they were asked to produce two multiple-choice questions on their area of research which were then authored on to a multiple-choice CALL program to present as an example of the class's understanding of the project (See Appendix A). What surprised the teachers involved, was the interest generated by class members in their peers' questions. Not only did they want to find the answers, they were prepared to disagree (in English), if they felt the answers were incorrect. While the questions were in a General Knowledge Quiz format, they had been written in English and, as such, had required some independent work on the side of the students. Learning how to author the questions on to the software had also required learning a new skill and given practise in computer skills. [This particular series of tasks is now used as an introduction to the use of CALL software and is still able to generate discussion among new students]. This movement along the cline had been natural and unforced and, more importantly, unnoticed. They were the 'knowers', but were protected both by their 'group face' and the teacher's continued 'control' of the situation. The 'group face' of this particular group was reinforced by the fact that the students were all female UAE national students in an immersion-style classroom setting; were all known to each other and had been working together for at least two-thirds of a semester.

It was decided to try to extend this participation by having the same students (in groups) write a test in class and give it to their peers. This did not, however, generate the same willingness to comment on peers' work. The only matter of importance was the answer—the product, not the process. It was obvious, therefore, that the original interest had been created by other factors. Several similar attempts, with this and other similar groups, failed to achieve any interest until the writing of tests on specific areas of grammar was set as a class project. Each group was given a specific area of grammar to test, they were allowed to use any resources they wished, but not allowed to copy questions from other sources. They were also taught how to author the tests into two different CALL software packages. After one week, each group presented two disks with a copy of their test on each. Students, in pairs, were then asked to take each of the other tests. On this occasion students were not simply interested in answers, but were trying to find errors in each others' tests.

Later discussion showed that as students were aware of the areas other groups were going to cover, they had also tried to find information or, at least, revise these areas themselves. They wanted to find mistakes by their peers, but did not want them to find mistakes in their own work. Small group 'face' had become more important than whole group face and a competitive edge had been introduced, not by the teacher, but by the students themselves. This, in some ways, produced a dichotomy for students in that, on the one hand, they were the 'appointed knowers' of certain information and, therefore, should not be criticised, but, on the other, through their revision, or other learning, "knew" the other information and were in a position to criticise, they had moved along the cline to the SPC area. At the same time, they were still in a -D -P situation where both solidarity of politeness and solidarity of purpose prevailed, i.e. they were still friends and still had the same overall goals. Group face, therefore, was not in danger.

Two factors seem to influence the amount of 'learning' which occurs: time and the allocation of tasks, without these, normal classroom reticence to be exposed is present. Another factor, and perhaps the most important, is also evident, students do not appear to see this as 'work'. They have introduced the competitive element themselves, and, as such, see the tests more as a game than part of the learning process. It is, I feel, vital to retain this illusion as this willingness to 'learn' or 'revise' independently must be retained.

This procedure was later extended to examining the test questions and, where applicable, distractors themselves. Originally a teacher suggested idea, students at all levels have now started to question the validity of different questions and items. An example of a very good criticism which comes from a Pre-Intermediate class where groups had been asked to write a simple test on different types of preposition is the following;

Choose the correct answer:

1. Robert suddenly began to feel sick __________ the exam.

   a. while b. during c. in

Distractor 'a' was not felt to be a suitable item in that this test was on prepositions of time and 'while' was not a preposition. It is this type of analytical thought which can be encouraged by this type of student created test which results from an apparent feeling that as the test has not been set by 'the teacher' the score is not important, but finding errors in the test is. Other teachers have found similar results with several different levels and it has not only occurred within the 'English' classes. Some Business classes have also reported similar results, but only when time has been allowed to produce the tests outside the classroom and different groups have been allocated particular areas.

The apparent ability of these tests to encourage both some autonomy and analytical thought in students appears to come, as posited earlier, from the perception that they are both 'safe', and therefore do not pose a threat to the face of either individuals or the group as a whole. and are felt to be fun, and therefore are not important to the final grade. In an attempt to both encourage learning and have more student involvement in the classroom, a slightly different approach has been taken by one or two teachers to utilise this technique as a classroom exercise.

This has taken the form of a reading test which practises the final examination format used at the particular level targetted. Having students produce questions from a reading passage to match answers is a standard feature of both EFL textbooks and classroom practice, which is often enhanced in the classroom by interaction between different groups. In this procedure a reading text is given to students who are then split into small groups with each being allocated a specific area of the test: main theme(s) of paragraphs, vocabulary, comprehension, True/False questions, all features of their final examination. At the end of a specified time, groups then present their part of the 'test'. Whilst finding the correct answers did not always seem to be of the greatest importance, discussing the suitability of the test question was. Some multiple choice distractors were felt to be 'too easy' and in one case, students realised that the answers to two comprehension questions were to be found in later questions by the same group.

From this, students have, we hope, acquired some knowledge of what is required from their final examination and how to tackle the test By allocating areas of focus, duplication of questions, which often occurs when students are asked to produce questions themselves, if not totally avoided, has been reduced and by having group presentation of questions, the competitive, therefore fun, element is maintained and individual 'face' saved. Student reaction suggests that provided the technique is not overused, they are happy to enjoy this new way of 'doing a test' which, of course, does not count in the final grade.

What conclusions, if any, are to be drawn from this? Firstly, and perhaps most important, these are observations only. They have yet to be subjected to the rigorous scrutiny of a research instrument and, therefore, can only be treated as hearsay. However, they do seem to show, that female UAE students can be motivated to taking some responsibility, if only a very small amount, for their own learning if they are required to produce a test for their peers. Provided they are given time and an allocated area of expertise, they do seem prepared to make extra effort to produce a test which is their own and, very importantly, 'correct' This, of course, enhances the small group and, therefore, individual face of the students involved. They also appear to be prepared to investigate the other allocated areas of expertise in order to find errors. This competitive aspect seems to be the motivation for further learning by the students and the fact that the results are not part of the final grade. The production and taking of the tests are not work, but fun, therefore not to be taken seriously, therefore no face can be lost. As stated earlier, it is important that this illusion that no work has taken place if students are to benefit from this and start to further develop both critical abilities and learner autonomy. As Dickinson points out (1987 P.2) "Learners do not achieve autonomy by being told to do so." This, hopefully, is just another technique which may help students to find the beginning of the path that leads there and perhaps, overcome what is described in Ellis as the teacher's paradox.

> We seek in the classroom to teach people how to talk when they are not being taught.

Perhaps here we can teach them to think about thinking without thinking about it.

Bibliography

Bassano S. & Christison M. (1995). Drawing Out Communication: Student-Created Visuals as a Means for Promoting Language Development in Adult ESL Classrooms. In Reid J (ed), Learning Styles in the ESL/EFL Classroom, (Pp. 48-62). Boston: Heinle and Heinle.

Dickinson L. (1987). Self-instruction in Language Learning, Cambridge: Cambridge University Press.

Ellis R. (1994). The Study of Second Language Acquisition, Oxford: Oxford University Press.

Harris D.P. (1969). Testing English as a Second Language, New York: McGraw-Hill.

Heaton J.B. (1975). Writing English Language Tests, Harlow: Longman.

Hughes A. (1989). Testing for Language Teachers, Cambridge: Cambridge University Press.

Jones C. & Trackman I. (1987) Choicemaster, London: Wida Software Ltd.

O'Malley J.M. & Chamot A.U. (1990). Learning Strategies in Second Language Acquisition, Cambridge: Cambridge University Press.

Owen M. (1990). Language as a Spoken Medium. In Collinge N.E. (ed) An Encyclopedia of Language (Pp.244-280). London: Routledge.

Madsen H.S. (1983). Techniques in Testing, Oxford: Oxford University Press.

Nunan D. (1988). The Learner-Centred Curriculum, Cambridge: Cambridge University Press.

Quirke P. (1997) The Importance of Face in the Classroom and Teacher Training TESOL Arabia 95 TESOL Arabia 96 Selected Papers (Pp. 65-75). Al Ain.

Rivers W.M. (1983) Speaking in Many Tongues, Cambridge: Cambridge University Press.

Scollon R. Scollon S. (1983) Face in interethnic communication in Richards J.C. and Schmidt R.W. (eds) Language and Communication, (Pp.156-190). Harlow: Longman.

Tutunis B. (1995) Writing: a motivator in the 'standstill' period. In Modern English Teacher (Pp. 25-27) Vol 4 (3) July 1995.

Willis D. (1990). The Lexical Syllabus: A New Approach to Language Teaching, London: Collins ELT. Software.

"Student Errors—To Use or Not to Use? That Is The Question"

JACQUELINE EADIE, ALI ABDEL-FATTAH AND
HEDI GUEFRACHI
United Arab Emirates University

*Student errors should never be used In testing;
only grammatically acceptable English
should appear on test papers. Do you agree?*

This article gives a brief account of the debate conducted at the "Colloquium on Current Trends in English Language Testing" held in Al Ain on June 11, 1997. The debate was mediated by Jacqueline Eadie. Ali Abdel-Fattah and Hedi Guefrachi argued for the motion while Bob Shaw and Salah Troudi put forward their arguments against it.

Following a presentation as to the type of errors under scrutiny, members of the debating teams were asked to present their cases either for or against the motion stated above. Each side was allowed 10 minutes to speak without interruption, before the audience was invited to participate in the discussion for a further 15 minutes. The session concluded with a vote in which all those present were asked to show whether they supported or refuted the motion.

1. Test-Formats Incorporating Student Errors

Multiple choice distractors are the most common type of test, in which the four possible choices include items already known to be confused by students. Identification and correction tasks are also common, using a text containing a number of errors, specifically written with students' previous difficulties in mind to produce reasonably attractive and plausible alternatives to the correct answer.

Example In which student errors are not used:
taken from SLEP (Secondary Level English Proficiency Test)

They joined other families to ____________ a tribe.

a) transport
b) form
c) oppose
d) settle

Example in which student errors are used:
taken from MLTELP (Michigan Test of English Language Proficiency)

I ____________ to ride my bicycle in the park.

a) like
b) liking
c) am like
d) likes

Most Arab-speakers at elementary level in English are likely to select distractor c), while many will opt for d). They are highly attractive to these students, despite being grammatically unacceptable in English.

Example in which student errors are used:
taken from TOEFL (Test of English as a Foreign Language)

The development of the computer is expected for to greatly change
A B C
the way people live.
D

Distractor B is attractive to intermediate-level students from many L1 backgrounds, particularly speakers of Romance and Slavic languages.

Example of a test item using a combination of student error and direct production taken from classroom quiz used by the author. All these examples are wrong. Write the correct form on the line.

oreng
orang
oringe ____________
oronge

A quiz comprised of 20 such spelling problems was complied using previous errors made in class. (Interestingly, almost every student, even those who had previously given the correct spelling, failed to understand the task correctly and selected one of the wrong answers presented, thus demonstrating the power of the printed word.)

Example using a combination of student error and direct production taken from CEELT (Cambridge Examination in English for Language Teachers)

All the underlined parts of the letter extract contain errors. Write the corrections in the box below.

..... I have a second request. Going to England is wonderful opportunity for me. I want to use my time well. I would like to retain information about social and culture activities in London. Have the students any kind of reductions? Because my own country has a lot of possibilities in this aspect. ....

In essence, the motion under debate asks whether we should not in fact exclude the use of student errors in favour of direct testing through production, which is arguably a more reliable and accurate test of student knowledge and performance.

2. For the Motion

The speakers believe that student errors should NEVER be used in testing and that only grammatically acceptable English should appear on test papers.

2.1 The Significance of Interlanguage

Interlanguage is the language which ESL/EFL learners produce while learning the target language (Brown, 1987). Such learners utilise a range of different mental processes, including borrowing patterns from the mother tongue. In second language acquisition, this is termed 'transfer'. An amusing example originating from spoken Egyptian Arabic is 'I want to drink a cigarette.' A more confusing utterance also stemming from Arabic is 'From five years I went to the UAE.'

Overgeneralisation of previously learned rules from the target language can also result in a series of common developmental errors. For example, some learners overextend the pattern whereby most nouns are made plural by the addition of -s and add -s to irregular nouns (man-mans, foot-foots) as well. Similarly, beginners frequently over apply the past simple rule by adding the suffix -ed to irregular verbs (come-comed, go-goed). Similar developmental errors are also noted in children learning English as their mother tongue. Parents react by simply reinforcing the correct use; they do not expose their children to errors to help them acquire their native language (Pfaff, 1986). It is therefore reasonable to assume that second language developmental errors will similarly disappear over time as the learner experiences more exposure to the target language.

It can also be argued that it is inappropriate and unfair to test mastery over developmental errors while the corresponding interlanguage is still at the halfway point. The learner with only 50% mastery of the rule might receive detrimental exposure. That is to say, it might reinforce the error to see and recognise it on the test paper. The test itself might confirm a wrong hypothesis.

In summary, since we do not know for certain what is taking place in any student's interlanguage, we cannot find precise answers to the question of why he or she makes certain errors. Consequently, we do not know precisely what we are testing when we use student-generated errors as distractors. This implies they must be invalid.

2.2 Pragmatic Considerations

There are also powerful pragmatic arguments against using student errors in testing. Pragmatics can be defined as the study of language in communication. It investigates in particular the relationship between sentences, contexts and the situations in which they are used (Parker, 1986). In this field, a speech act is defined as an utterance used as a functional unit in communication. It may have both propositional meaning, expressing the speaker's physical state, and illocutionary force, revealing the message the speaker really wishes to convey.

Test-formats using student errors as distractors are generally oversimplistic, failing to provide adequate context and taking no account of pragmatic considerations. Testing grammar in the absence of meaning trivialises meaning and exaggerates the process of making mistakes. After all, most such errors are entirely harmless in terms of meaning and illocutionary force. "I am afraid from spiders" still carries the full intended meaning. Multiple-choice testing using a distractor set of prepositions including 'from' is arguably quite pointless and even harmful. For are we not overstressing accuracy and neglecting communication with this type of testing?

2.3 The Significance of Culture

Language learners in the United Arab Emirates (UAE) have a very limited exposure to English outside class. English is taught by non-native speakers of English in most schools and sometimes these teachers resort to the use of Arabic for certain classroom activities. This results in minimal exposure to English and early fossilisation in many cases. In such an environment, students cannot be expected to have developed the necessary level of discrimination between accurate and erroneous English structure and lexis. Using student generated errors with such a population can be regarded as trapping learners into making mistakes and is likely to be counter-productive.

Moreover, there is an additional cultural factor which argues strongly against the use of errors. Most UAE nationals are exposed to Asian speakers of English in the home and when shopping or conducting business. These Asian varieties are frequently different from the standard versions used in coursebooks and by teachers of English. Instead of seeing the product as necessarily wrong, should we not instead be sensitive to the role of English as a language of international communication? By ruling certain features of other varieties as wrong in our tests, we may be adding to the problem of confusion and linguistic imperialism.

"The design of distractors to trick the learners into confusing dilemmas is counter- productive." Oller, 1979: 256)

2.4 Language Learning Construct Validity

Language tests are based on theories of language learning. The theory of language transfer, whilst generally accepted, has not been proven conclusively, with the numerous anomalies seen amongst individuals from similar backgrounds. As an example, Arab students use 'married with' just as frequently as 'married from' in place of 'married to'. It is thus difficult to see any simple case of transfer here. Learners, even at comparable proficiency levels, form personal interlanguages which are distinct. This implies that their use of transfer also varies. If we are not clear about what kind of transfer is taking place during the learning process, then we are misguided in using transfer errors as distractors for testing purposes.

2.5 Test Construct Validity

Testing what has not been taught makes a test invalid. Tests should not include error correction unless contrastive analysis between the native, interlanguage and target language is included in the syllabus. There are special cases where error correction is given specific attention, such as the cramming courses followed by learners preparing for TOEFL. Learners who do not undertake such courses are gravely disadvantaged by the test. But the relative value of learning error correction for a test seems poor indeed when compared with that of dealing with errors in a meaningful context. It is not even a question of necessity to test in this way. Other respected proficiency tests like the IELTS and Cambridge series do not use errors. They use wholly authentic language and the only errors that appear on the test paper are the ones produced as the test-takers work through the test.

2.6 Face Validity

Tests that include errors usually have no face validity for many students, since most have been trained to deal with correct language and avoid mistakes. Moreover, learners in the Arab world believe strongly in the authority of the printed word. They also have a certain expectation of testing tradition acquired through their studies of Arabic. Exposing such students to errors in tests not only affects the validity of the test, but also lowers the credibility of the teacher or test developer. Experience has revealed that the use of student errors on the test paper can lead to great confusion.

3. Against the Motion

The speakers believe that student errors SHOULD be used in testing and that grammatically unacceptable English can appear on test papers.

3.1 Frequency of Student-Generated Errors

In any given teaching situation, teachers will see exactly the same errors produced by different students over and over again. We know that this phenomenon is the result of factors such as transfer from the mother tongue, overgeneralisation of rules and incorrect assumptions about the target language, although we can never be certain which cause underlies an error without consulting the student concerned.

Nevertheless, teachers are generally able to predict the errors that a student will make in production activities on the basis of such factors as mother tongue, prior teaching and the level of English. Let us consider a couple of examples. Teachers of Arab-speaking students at various levels find the present simple form so frequently rendered as 'I am study' 'He is live' that they often want and need to test whether a student has adequately mastered this problem area. Otherwise, it can lead to greater confusion with the continuous form and with passive constructions. Consequently, present simple is a popular item in class quizzes. An entirely different case is that of 'I can to drive' 'She must to finish' and so on. The English modal system contains a number of features that seem to be universally problematic for learners from almost all language backgrounds. These types of errors occur with such frequency that many teachers recommend their inclusion in tests on the grounds that they are problems inherent in the English system itself.

If certain errors are so frequent and predictable as to warrant special attention in the classroom, then it would appear to be almost negligent to avoid their incorporation in test items.

3.2 Face Validity

Using predicted errors in testing has face validity for teachers because they see the errors reproduced endlessly. It has face validity for students because they see that they and their peers are making the same mistakes again and again. Passing such test items, even in the face of temptation, means that the area of difficulty has successfully been mastered. If we fail to include predicted student errors in our tests and quizzes, then are we not in fact failing to test whether or not students have acquired this crucial knowledge.

3.3 The Significance of Interlanguage

Most teachers' basic understanding is that interlanguage is developing throughout the entire language learning process and they give explicit help in providing feedback as to when and why an error might have occurred. They compare, contrast, explain, use visuals, colours and mark things wrong with red pens! It would therefore seem likely that the appearance of common errors previously given attention in the classroom might help stimulate conscious knowledge, even if a student's performance is still impaired. There is an argument therefore that the use of common errors as distractors can actually help alert students and improve their test scores.

3.4 Convenience of Use

Student errors are both useful and convenient in a wide range of testing contexts. Multiple choice questions and short passages incorporating them can be written easily and specifically tailored on the basis of teacher experience.

They are entirely appropriate for diagnostic placement because they readily identify the student's needs and can be use to guide remedial work and curriculum development. In the classroom, they provide a wealth of items for discrete item quizzes that are easy and quick to sit, correct, score and analyse for structured feedback purposes. They are thus ideally suited to the medium of computer-generated testing. Students also enjoy generating error correction quizzes for their peers. Proficiency examinations such as TOEFL use errors characteristic of learners from numerous language groups, collected over an extended period. Proficiency exams, which are unrelated to any detailed syllabus, can exploit these very obvious markers to indicate the general level that the learner has reached.

It is, of course, inappropriate to include student errors in achievement tests for courses which do not contain error correction in the syllabus. This would break the basic rule of testing—only test what has been taught. However, there is a strong pedagogical argument for including systematic proof-reading so that students learn to correct their own work. It is better to teach learners to proof-read than to ignore the learning value of error correction. As learners improve their English, one of their most frequently expressed concerns is accuracy. They themselves are aware that they will be handicapped at higher educational, business and professional levels of development if they are unable to detect their own mistakes in written English.

4. Discussion

It was noted that, although we don't really know what we are testing with the use of student-generated errors, there exists such a huge corpus of common student errors that no experienced teacher can fail to be aware of their significance. Most people at the debate did NOT regard their use as the deliberate entrapment of learners into making mistakes, but rather as a way of addressing a very real phenomenon that needs attention in the classroom. It was also noted that student errors provide a very convenient basis for test-writing and marking.

"Why not use mistakes to highlight, to make students more aware? Then testing also becomes teaching!" (Mongi Al Baratly) "But tests shouldn't teach." (Bob Shaw)

It was mentioned that errors are indeed already heavily in use. They are most popular at the microevaluation level, as seen from the types of quizzes popular with teachers and students alike. While there is not enough time for assessing their validity and reliability, there is much face validity because the test is thus related to everyday problems. The teacher wants to find out what the learner really knows and the student wants to know how much he or she has progressed.

" Use the results for feedback sessions." (Tuhami Al Fayash)

"Mistakes are motivating, so don't be afraid to deal with them." (Ali Najeeb)

Although internationally important examinations like TWE and IELTS do not make of use of student-generated errors at all and instead assess grammar as a component within a written essay question, at the macroevaluation level we have MLTELP, SLEP, TOEFL, TOEIC and CEELT making extensive use of the formats under scrutiny here. These examinations relate to important steps for the learner, often proving decisive for entering university or getting a job. This sends a clear message to teachers and learners that they need to address the issue of errors.

"Errors are the opposition. If you ignore them, they will only get stronger." (Speaker unknown)

5. The Vote

For the motion	7
Against the motion	15
Abstentions	2

The clear majority voted against the motion. The debate concluded that it is indeed permissible to use student errors in testing.

References

Brown, H. D. (1987) Principles of Language Learning and Teaching. Prentice Hall Regents, New Jersey: Englewood Cliffs.

Oller, J. W. (1979) Language Tests at School. New York: Longman.

Parker, F. (1986) Linguistics for Non-Linguists. Boston: College Hill.

Pfaff, C. (1986) First and Second Language Acquisition Processes. Cambridge, USA: Newbury House.

Public Examinations Cited

USA

ELP: English Placement Test
Testing and Certification Division of the English Language Institute, University of Michigan.

MLTELP: Michigan Test of English Language Proficiency
Testing and Certification Division of the English Language Institute, University of Michigan.

SLEP: Secondary Level English Proficiency Test
Educational Testing Service, Princeton, New Jersey.

TOEIC: Test cf English for International Communication
Educational Testing Service, Princeton, New Jersey.

TOEFL: Test of English as a Foreign Language
Educational Testing Service, Princeton, New Jersey.

TWE: Test of Written English
Educational Testing Service, Princeton, New Jersey.

UK

CEELT: Cambridge Examination in English for Language Teachers
University of Cambridge.

IELTS: International Examination
The British Council, Manchester.

APPENDIX

The First Annual CTELT Program

Date: Wednesday, June 11, 1997

Keynote Speaker
(9:30-10:30 Conference Room)

Computer Based Language Testing: The Call of the Internet

GLENN FULCHER
University of Surrey, UK

This keynote address will look at the uses language testers have made of the computer over the last two decades and relate these uses to some of the theoretical concerns of language testing. These concerns include the nature of the measurement scale underlying particular tests, and the constructs test designers think they measure. However, it will be argued that studies in these areas have not done justice to the challenges or problems associated with computer based testing.

A prototype Internet listening test will be demonstrated as an example of the challenges of future innovation, whilst research into equity issues in delivering tests on the Internet will be presented to widen the debate on computer based testing.

Glenn Fulcher is the Director of the English Language Institute at the University of Surrey. He earned his Ph.D. with Charles Alderson at the University of Lancaster. As befitting someone with a wide background in all areas of testing, Glenn Fulcher developed the Resources in Language Testing Internet Homepage at http://www.surrey. ac.uk/ELI/eli.html. This valuable electronic resource for all language testers recently received commendation from the International Association of Language Testing for the dissemination of testing information. Fulcher's current areas of research include oral skills assessment, placement examinations, computer based testing, and item response theory.

You can reach Glenn Fulcher by email at g.fulcher@surrey.ac.uk.

First Session: 10:40 to 11:25

An Assessment of the Secondary School Certificate Examination

ABDULLAH LIBDEH
Faculty of Education, United Arab Emirates University

This paper aims at assessing the Secondary School Certificate Examination to determine the extent to which it fulfills the aims set by the Ministry of Education. The data will be gathered through content analysis and the sample will be a corpus of the tests administered in the last five years.

Abdullah Libdeh is an Associate Professor of TESOL at the Faculty of Education, at UAE University. He teaches ELT methodology at the Dept. of Curriculum and Instruction and is the Director of Field Experiences Office at the Faculty of Education.

The Design and Development of a Placement Test for English-Medium Tertiary Education in the U.A.E.

ELIZABETH HOWELL
Higher Colleges of Technology, Abu Dhabi, U.A.E.

This paper defines placement testing and compares its function with those of other types of language tests. It then briefly describes the theoretical and practical issues involved in placement test design. Subsequently, the paper reviews some of the possible options in the design of placement tests in different contexts, including both commercial and custom-made tests. Finally, the paper describes the design and development stages of the test currently in use in the Higher Colleges of Technology.

Elizabeth Howell has taught and tested EFL/ESL in Europe, the Middle East and the Far East for a quarter of a century. She is an RSA/UCLES DTEFLA Senior Practical Assessor and CELTA Course tutor. She believes firmly in the need for a close relationship between testing and teaching to ensure the improvement of both.

Exploring English Language Tests on the Internet

BOB CATTO
United Arab Emirates University

The Internet is a relatively recent resource in the Arabian Gulf. Many language teachers and testers have yet to explore the English language testing resources available on the Internet. This session provides a hands-on opportunity to examine and assess testing sites on the Internet.

Bob Catto is a Supervisor in the UGRU English Program at UAE University. He is co-chair of the TESOL Arabia Internet SIG. Bob has an extensive background in Educational Technology and is an experienced Internet user.

Second Session: 11:35 to 12:20

Examination and Curriculum Reform: The Oman Ministry of Education Perspective

JANET AL LAMKI
Ministry of Education, Muscat, Oman

The Ministry of Education of the Sultanate has undertaken a radical and wide-ranging programme of Education Reform. As part of this programme, the existing examination system has been subjected to review and a number of major changes have been suggested. The presentation will briefly set the context of the overall programme reform, indicate the importance of examination reform in this context, and deal with some of the major changes to the examination system which are presently under consideration. The rationale behind these suggested changes will be outlined.

Janet Al Lamki has been the Director of the English Language Department, Ministry of Education, Sultanate of Oman, for the past eleven years. She was educated in Cairo and at Georgetown University in Washington, DC.

The PET Test: Appropriateness for Placement Testing?

LORING TAYLOR
Dept. of English, Sultan Qaboos University, Oman

For the past five years the Language Centre at SQU has used the PET test to place all incoming freshmen. This instrument has also been used as an exit test in the College of Commerce. At present, the PET is being phased out for both these purposes. In order to make the PET ready for these functions, numerous changes were made to the exam, both at the time it was acquired and during the course of its use. This paper examines these changes and measures the reliability and validity of the PET against another instrument and against actual grades in English courses.

Loring Taylor is an Associate Professor and Deputy Chair of the Dept. of English at Sultan Qaboos University. He holds a Ph.D. from the University of California. He has taught in the U.S., Romania, Yemen, Oman, and Jordan.

Innovative Formats in Computer Based Testing

CHRIS HEAD AND NANCY HUBLEY
U.G.R.U., United Arab Emirates University

Computer based testing (CBT) provides opportunities to reconsider question formats for language testing. Although computerization of exams entails some inherent constraints, these are more than offset by new possibilities that CBT offers. This presentation explores English language testing formats using the CGE software developed in house at UAE University.

Chris Head, Assistant Head of the Math/Computer Unit, UAE University, led the development team for the Computer Generated Examination software. Nancy Hubley chairs computer based testing for the UGRU English Unit and is co-chair of the TESOL Arabia Testing and Internet SIGs.

Third Session: 1:40 to 2:25

Student Created Tests as Motivation to Learning

PHIL COZENS
Ras Al Khaimah Women's College, U.A.E.

This paper will look at classroom tests and particularly, the involvement of learners in the creation of tests in order to increase motivation and lead to autonomous learning.. It also looks at the concepts of 'face' in interethnic situations and the EFL classroom. This paper briefly examines Gulf Arab students' perceptions of student/teacher roles and how they can be used to encourage student autonomy.

Phil Cozens has taught English and Computers at Ras Al Khaimah Women's College since September 1994. Prior to his work at the Higher Colleges of Technology, he spent ten years teaching EFL in Hong Kong. Phil holds a recent MA in Linguistics (TESOL) from the University of Surrey.

Everything You Wanted to Know about UCLES

KATHY BIRD
U.G.R.U. English, United Arab Emirates University

The University of Cambridge Local Examinations Syndicate (UCLES) was established as a department of the University of Cambridge in 1858. Examinations in English as a Foreign Language have been administered in Cambridge since 1913 when the Certificate in Proficiency in English (CPE) was introduced. The examinations are widely recognized by universities, polytechnics, colleges, schools, and the business world. This session will provide you with a broad overview of most of the English as a Foreign Language examinations administered by the UCLES.

Kathy Bird is a lecturer in the UGRU English Program at UAE University. She coordinated the Ministry of Education Supervisor's Training Program. She has taught EFL in Oman and Greece.

Testing Perspectives: "Mr. Play it Safe" vs "Mr. Risk Taker."

ABDULLA SOLIMAN
Teacher Inspector, U.A.E. Ministry of Education.

In the age of information technology, the exam system is a catalyst for educational change. This presentation compares and contrasts two different test-related perspectives. Mr. Play-it-Safe and Mr. Risk-Taker. The former writes tests to please all parties concerned (students, teachers, and administrators), while the latter sets challenges that fit nicely with the spirit of the new era. Participants will engage in a workshop activity.

Abdullah Soliman is a teacher inspector for the Al Ain Educational Zone.

Innovations in TOEFL Testing

MARY CORRADO, Country Director AMIDEAST, U.A.E.
FRANK GIANNOTTA, Unit Head, UGRU English Program, U.A.E. University

The TOEFL Test offered worldwide by Educational Testing Services (ETS) is presently undergoing a number of major changes. This presentation will focus on two of the most important reforms: TOEFL 2000 and the new SPEAK test. Within a few years, the TOEFL will be administered in a computer adaptive testing format. This presentation will also explore recent changes in the oral component of the TOEFL, the SPEAK test.

Mary Corrado is country director of the AMIDEAST Program in the UAE. **Frank Giannotta** heads the UGRU English Program at UAE University. He formerly served in the Peace Corps in Turkey and directed the ESL Program at Duquesne University in Pittsburgh.

Test Writing for U.A.E. Distance Learning Students

LISA BARLOW, RITA MCDONAGH AND CHRISTINE CANNING
United Arab Emirates University

This paper addresses the challenges involved in producing multiple versions of exams for students in the UAE University Distance Learning Program. These involve 1) equalizing and maintaining reliability of test items and texts; 2) retaining validity while producing multiple exam versions from a limited base of course objectives; and 3) safeguarding reliability and test security when exams are given over a three week period. In addition, the paper discusses cultural appropriateness of test materials for Gulf Arab students.

Lisa Barlow directs the Distance Learning Program for the English Unit at UAE University, where she has served as on the testing and curriculum teams. Her colleagues, Rita McDonough and Christine Canning, both teach and develop tests for the Distance Learning Program. Rita's specialty is developing reading and writing materials for upper level research courses. Christine has produced an ESP English for Education course for distance learners. All three presenters have extensive experience in the Arabian Gulf region.

Fourth Session: 2:35 to 3:20

English-medium Content Area Testing by Non-Native Users of English

BOB HUNKIN
United Arab Emirates University

This presentation focuses on an analysis of subject-specialist tests in the Agricultural Sciences developed and written in English by faculty lectures whose first language is Arabic. The range and frequency of occurrence of various question types, the use of English in rubrics and questions, test styles and approaches will be discussed in relation to the stated testing objectives, linguistic competence, and educational backgrounds of the test developers.

In his eighteen years of EFL teaching and management, **Bob Hunkins** has been involved with a wide range of public examinations, and has developed placement, achievement, and proficiency tests. Bob is a lead teacher in English for Agriculture at UAE University. He holds an M.Ed. from Exeter University.

A Retrospective of RSA/DTEFLA Questions on Testing

CHRIS PEARSON
United Arab Emirates University

The UCLES/RSA DTEFLA examination is widely recognized as one of the best and most practical teaching qualifications in the world of ELT. Its annual written papers reflect the major current preoccupations and concerns in the field. This session will look at the DTEFLA exam questions that have been set on issues relating to Testing over the last ten years. We will use the DTEFLA exam to attempt to get an overview of the developments and trends in English Language Testing that have dominated the last decade.

Chris Pearson has been a teacher trainer for the last 27 and involved in the RSA Diploma for the last 9 years. A lecturer at the UAE

University, Chris is currently president of TESOL Arabia and was Conference Chair for the successful 1997 TESOL Arabia Conference.

Interaction in Oral Proficiency Discussions

JOHN POLLARD
Saudi Development & Training Company, Saudi Arabia

An a priori validation exercise has recently been completed focusing on an Oral Proficiency Interview/Discussion Test (OPD) developed specifically for Saudi Arabia. The test involves an interlocutor and a computer. This paper examines uses and abuses of multimedia technology with regard to authenticity in language testing, and offers possible new directions by tracing the main thread of debate that has accompanied the OPI over the last decade. Construct and consequential validity with regard to interaction will be considered apropos future applications and research development.

John Pollard has an M.A. in TESOL/TEFL. He has worked in the language testing field for 15 years in a variety of contexts. His past experience includes the British Council, UNESCO and the British ODA.

Student Errors: To Use or Not to Use? That is the Question

Chair: Jacqueline Eadie, U.A.E. University
Panel: Hedi Guefrachi, U.A.E. University
Robert Shaw, U.A.E. University
Ali Abdul Fattah, U.A.E. University
Salah Troudi, U.A.E. University
Josephine O'Brien, Higher Colleges of Technology

A panel of experienced English language teachers and testers will debate the issue of whether student-generated errors should be used as test distractors. Panel members include some Arabic-English bilingual speakers.

Jacqueline Eadie has an M.A. from the University of Edinburgh and an RSA Diploma. She is currently a lecturer in the UGRU English Unit at UAE University where she serves on the testing committee. **Hedi Guefrachi** is the Professional Development Supervisor at UAE University where his work involves liaison with the Ministry of Education. **Robert Shaw, Ali Abdul Fattah,** and **Salah Troudi** are all lecturers in the UGRU English Unit. **Robert** is editor of the TESOL Arabia Newsletter and **Salah** chairs the Teacher Education SIG of TESOL Arabia. **Josephine O'Brien** has many years experience teaching in the Arab world. She is currently a lecturer in the CD program at the Higher Colleges of Technology in Al Ain.

Fifth Session: 3:30 to 4:15

Oral Skills Assessment of Gulf Arab Learners: An Adventure in Futility?

MICHAEL BIRCHALL
United Arab Emirates University

This presentation describes the development and implementation of a testing component to complement the proposed oral skills component in the curriculum of the UGRU English Program at the UAE University. Two different types of oral assessment will be delineated: classroom-based evaluations and oral interview examinations. Issues relevant to the testing of oral skills at the UAE University will be discussed.

Michael Birchall is a lecturer in the UGRU English Program and is currently the head of level one.

Designing Multi-Purpose Written Communication Descriptors for the Higher Colleges of Technology

NICOLA MARSDEN
Higher Colleges of Technology, Abu Dhabi

This paper describes the development, implementation and validation of written communication descriptors for use in a system of tertiary, vocational training colleges in the UAE. The descriptors are used to produce inter-rater reliability amongst 300 assessors who evaluate a wide range of writing samples. The aim is to produce consistent scores which are comparable to externally validated writing measures for placement, progress and achievement tests.

Nicola Marsden has an M.A. from Manchester University. She is currently Coordinator of General Education at the Higher Colleges of Technology in Abu Dhabi. She is responsible for system-wide development and delivery of assessment in the CD Program. She has taught EFL in France, Saudi Arabia, Iraq, Bangladesh and the UAE.

Issues in F/SL Academic Listening Assessment

CHRISTINE COOMBE, JON KINNEY AND CHRISTINE CANNING
United Arab Emirates University

Unlike F/SL listening in an interactional setting, a major part of listening in a university context involves lecture comprehension, which differs in a number of ways. These differences have important implications for testing. Validity, reliability, practicality and authenticity issues in academic lecture comprehension will be discussed. An evaluation scheme for academic listening assessment instruments will be proposed in order to assist ELT professionals in the selection and production of effective tests for academic listening.

All three presenters are lecturers in the UGRU English Program at UAE University. **Christine Coombe** is Testing and Measurements Supervisor, TESOL Arabia Vice President and Co-Chair of the Testing SIG. **Jon Kinney** is chairperson of the Independent Learning and Tutorial Center. **Christine Canning** is chairperson of the Media Graphic and Visual Arts Committee.

Language Testing: A Means to Recognize L1 Transfer

MARVIN TAYLOR
United Arab Emirates University

L1 transfer effects Arab students test-taking strategies in a number of ways. This presentation uses specific examples of language transfer to highlight the role L1 plays in students' perceptions of test tasks and the answers they produce.

Marvin Taylor is a lecturer in the UGRU English Program at UAE University. He serves on the Testing Committee where his primary focus in L1 transfer and error analysis. Marvin's long-standing interest in Arabic stems from his graduate work at the University of Indiana and teaching in Saudi Arabia and the UAE.

Final Session: 4:45 to 5:30

Current Trends in English Language Testing
What are the issues and future directions?

FRANK GIANNOTTA: Panel Moderator

Panel participants:
Loring Taylor, Sultan Qaboos University
Bahia Diefenbach, U.A.E. University
John Pollard, Saudi Development and Training
Glenn Fulcher, University of Surrey
Abdullah Libdeh, U.A.E. University

This panel of experienced teachers and testers addressed a variety of questions and issues.

The Impact of High-Stakes Testing on Teaching and Learning

DIANNE WALL
Lancaster University

1. Introduction

The 'impact' of a test is the effect it may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole (Wall 1997:291). It is generally accepted that 'high-stakes' tests (defined by Madaus [1988:7] as 'those whose results are seen -rightly or wrongly - by students, teachers, administrators, parents, or the general public, as being used to make important decisions that immediately and directly affect them') will influence the way that students and their teachers behave as well as their perceptions of their own abilities and worth. High-stakes tests may have impact on the content and methodology of teaching programmes, attitudes towards the value of certain educational objectives and activities, the academic and employment options that are open to individuals, and in the long term, they may 'reduce the diversity of talent available to schools and society' (Ebel 1966, cited in Kirkland 1971:305).

The impact of high-stakes testing has been of interest to those in the field of general education for some time, but it was not until the early 1990s that language educators began to pay serious attention to the phenomenon. There is now a small but significant literature in our field which addresses the questions of whether tests really are as powerful as they are believed to be, where their supposed power comes from, what kinds of effects they have on teaching and learning, and what other factors in the educational context might influence what happened in the classroom. These questions are of importance not only to teachers and learners, who are the people who are most directly affected by them, but also to educational institutions and larger entities such as regional and national ministries of education.

In this paper I shall review some of the functions of high-stakes tests in society, and then present views from general education and language education about why test impact occurs and the forms it can take. I shall then discuss the need to explore other disciplines, particularly the field of educational innovation, in order to gain a better understanding of how to maximise the positive potential of high-stakes testing.

2. The Functions of High-Stakes Tests In Society

A number of specialists have written about the functions of high-stakes tests throughout history and the impact that they have on the education systems into which they have been introduced. One of the most accessible reviews is by Eckstein and Noah (1993). Eckstein and Noah claim that the first documented use of 'written, public, competitive examination systems' occurred under the Han Dynasty in China, about 200BC. The function of these examinations was to select candidates for entry into government service: they were used to 'break the monopoly over government jobs enjoyed by an aristocratic or feudal class' (p. 5). The impact of these examinations was substantial: Spolsky writes that it was 'to establish and control an education programme', in which prospective mandarins prepared themselves for a major professional hurdle (1995:55). This system lasted until the beginning of this century, and influenced the development of examinations for similar purposes under Frederick the Great in Prussia (the Abitur, 1748) and under Napoleon in France (the Baccalaureat, 1808).

The second function that Eckstein and Noah mention was to check patronage and corruption. Britain is given as an example of a country where people were able to gain entry into education or the professions on the strength of whom rather than what they knew. This situation began to change only in the middle of the 19th century, with the establishment of examinations for purchasing military commissions (1849), entry into the Indian Civil Service (1853), and entry into the Military Academies (1858). An important consequence of the introduction of these examinations was the establishment of numerous 'cramming establishments', which specialized in preparing students for the examinations.

The third function of high-stakes examinations was to encourage 'higher levels of competence and knowledge' amongst those who were entering government service or the professions. The intention was to design examination which would reflect the demands of the target situation: the key examinations in France were those controlling entry to the grandes ecoles; in Germany it was the Staatsexamen. Candidates who were preparing for these examinations would have to develop the skills which were relevant to the work they hoped they would eventually be doing.

The fourth function was that of allocating sparse places in higher education. An obvious example was Japan at the beginning of the 20th century, where examinations were used as a means of selecting only the most able candidates for the few places available at secondary and tertiary levels. Selection by examinatijon is still the rule in modern times, although the focus has changed from gaining entry into any institution to gaining entry into the most prestigious institutions at every level. The competition for places has led to what is commonly referred to as 'examination hell' and to the proliferations of juko and yobiko, which are basically 'cramming schools' for the most important entrance examination.

The fifth function of examination was to measure and improve the effectiveness of teachers and schools. Eckstein and Noah again use Britain as an example, describing how in the 1860s the government instituted a system of examinations to monitor the performance of primary schools which were receiving central funding. Inspectors were sent in to test pupils and to report their findings. The amount of funding that a school received depended on its students' performances. This system, which came to be known as the 'Payment by Results' system, had a drastic impact on teaching and learning, instilling drilling and cramming, and giving very little thought 'to the real training of the child, to the fostering of his mental (and other) growth'. (Holmes 1911: 107-108, cited in Stobart and Gipps 1997: 4).

The final function mentioned by Eckstein and Noah was limiting curriculum differentiation. In Britain in the 19th and early 20th centuries there was considerable resistance to the idea of centralised education, and all schools had the freedom to decide on their own curriculum and means of assessment. It was not until 1917, with the establishment of the School Certificate examinations, which were controlled by a number of examining boards affiliated with different universities, that schools had a common target they could aim for. These examinations were replaced by the General Certificate of Education (GCE) O-Level and A-Level examinations in 1951, and these were supplemented by the Certificate of Secondary Education (CSE) examinations in 1965. These examinations exercised an 'indirect control of the curriculum' which continued until 1988, when the National Curriculum for England and Wales took on the role of setting objectives and standards for primary and secondary education.

Though many of the examples given here are from Britain and Western Europe, it is likely that examples exist in many countries, at national level and at the level of individual institutions, such as universities or prestigious organisations. Readers are invited to think of examples in their own education systems!

3. Views from General Education

The impact of high-stakes testing has been a subject of interest in general education for many years, and educationalists have offered strong views both in favour and against using tests for the purposes described in Section 2 above.

One of the best known advocates was Popham, who coined the term 'measurement-driven instruction' (MDI) to refer to situations in which high-stakes tests could lead to educational improvement (1987). Popham argues that if such tests are 'properly conceived and implemented' then focusing teaching on what is assessed by the tests is a positive activity. In order for a test to be 'properly conceived and implemented' it must meet five conditions:

1. It must be criterion-referenced, because 'the descriptive clarity of well-constructed criterion-referenced tests gives teachers a comprehensible description of what is being tested'.

2. It must contain defensible content - that is, important, not trivial, knowledge and skills.

3. There must be a manageable number of targets. Popham claims that during the 'heyday of behavioral objectives' teachers were faced with 'endless litanies of minuscule instructional targets'. It is important to test more general targets, which subsume smaller ones.

4. The test must provide for instructional illumination - that is, it should encourage teachers to design 'effective instructional sequences'.

5. Instructional support must be provided to teachers so that they cope with the test's demands.

Other specialists who have presented arguments in favour of using tests are Morris (1961), Frederiksen and Collins (1989) and Heyneman and Ransom (1990).

Morris writes that tests can be used as 'a means of maintaining standards, as an incentive to effort... and as tool of social engineering' (1961: 1). Although the term 'social engineering' has negative connotations, Morris believes that tests also have positive potential. It is necessary to examine 'the social philosophy of those who order the manufacture of the tool and direct its operation' (P.25).

Frederiksen and Collins (1989) recommend that one of the factors that ought to be considered in the evaluation of tests is their 'systemic validity'. A systemically valid test is one:

> ... that induces in the education system curricular and instructional changes that foster the developmetn of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time (P.27).

Heyneman and Ransom (1990) draw on their experience and that of colleagues conducting research for the World Bank and similar organisations in developing countries. They too argue that tests 'can be a powerful, low cost means of influencing the quality of what teachers teach and what students learn in school' (P.105), given that the right conditions are in place. These include high-quality test design (cf. Popham's conditions above), and good communications between the testing agency, students, teachers, school administrators and others concerning the performance of students on the test and implications for future teaching and resourcing.

Perhaps the strongest critic of Measurement-Driven Instruction is Madaus (1988), who condemns it as 'nothing more than psychometric imperialism' (P.84). He predicts only negative effects if testing is used as the 'primary motivating power of the educational process'. He states that

> Measurement-driven instruction invariably leads to cramming; narrows the curriculum; concentrates attention on those skills most amenable to testing . . .; constrains the creativity and spontaneity of teachers and students' and finally demeans the professional judgement of teachers (P.85).

Madaus reviews a number of studies on the impact of testing on teaching and presents a set of seven 'principles' which summarise his own position:

1. The power of tests and exams to affect individuals, institutions, curriculum or instruction is a perceptual phenomenon. If students, teachers or administrators believe that the results of an examination are important, it matters very little whether this is really true or false. The effect is produced by what individuals perceive to be the case.

2. The more any quantitative social indicator is used for social decision making, the more likely it will be to distort and corrupt the social processes it is intended to monitor.

3. If important decisions are presumed to be related to test results, then teachers will teach to the test.

4. In every setting where a high-stakes exam operates, a tradition of past exams develops, which eventually de facto defines the curriculum.

5. Teachers pay particular attention to the form of the questions on a high-stakes test (for example, short answer essay, multiple choice) and adjust their instruction accordingly.

6. When test results are the sole, or even partial, arbiter of future educational or life choices, society tends to treat test results as the major goal of schooling, rather than as a useful but fallible indicator of achievement.

7. A high-stakes test transfers control over the curriculum to the agency which sets or controls the exam. (Madaus 1988:88-97)

Other specialists share Madaus' concern about the negative effects of testing, amongst them Haladyna, Nolen and Haas(1991), who use the term 'test score pollution' to refer to practices which 'increase or decrease test performance without connection to the construct represented by the test' (p 4),and Fullilove (1984), who writes of educational systems where 'the examination tail wags the curriculum dog'.

4. Views from language education

There was little discussion in the language education literature of the impact of tests on teaching before the early 1990s. What discussion there was consisted mainly of definitions and brief explanations given in language testing handbooks, claims about the importance of tests that did not include reference to research, and expressions of hope or worry concerning the impact of specific examinations on the teaching contexts they were being introduced into.

The discussions in language testing handbooks tended to focus on the effects of testing on the classroom, rather than on the effects on individual students or on society. Some writers use the term 'washback' and others 'backwash' to describe these effects, stating that tests can influence 'what and how the students choose to study and ... teaching procedures' (Finocchiaro and Sako 1983:311), 'teaching and learning' (Hughes 1989:1), or simply 'instruction' (Bachman 1990:283).

Some references to test impact are very brief. Madsen, for example, claims that 'an occasional focus on grammar or vocabulary or mechanics can have a good 'backwash' effect on the teaching of writing (1983:p 120), without offering any support for his argument. Heaton (1990) devotes more attention to the topic, claiming that teachers will always base their teaching on exams which are used for selection purposes and which are designed by bodies external to the schools (p 16). However, he generalises about the influence these examinations will have:

> If it is a good examination, it will have a useful effect on teaching; if bad, then it will have a damaging effect on teaching. (ibid)

Davies also believes that 'all language tests and examinations exert an influence on the curriculum and on the teaching', and states that if this influence is indeed inevitable then test designers should 'at least try to ensure that the exams are good ones' (1977:42). He suggests two ways of ensuring that the washback from exams is beneficial: designing a 'closely detailed and wide-ranging syllabus' (a public statement of exam requirements - see Alderson, Clapham and Wall 1995:9) and making sure that the exam itself illustrates clearly the types of proficiency that are required.

Finocchiaro and Sako (1983) refer to the 'four persistent problems in testing', the last of which is the 'degree to which testing either enhances instruction or, alternatively, distorts it through various feedback effects from the tests' (p 11). They claim that testing should not distort learning if the testing and the teaching 'both derive from sharply defined objectives based on sound inter-disciplinary theory'; however, they do not elaborate on what this theory should contain or how the objectives and theory will ensure that tests 'will be a positive motivating force for student and teacher alike' (p 41).

Weir (1990) believes that an evaluation of communicative tests should include the systematic gathering of data on construct, content, face and 'washback validity' ('a measure of how far the intended washback effect [is] actually being met in practice' - see Morrow 1986 for further discussion), while Hughes (1989) devotes several pages to his own guidelines for achieving beneficial backwash, which include testing the abilities that need to be encouraged, sampling widely and unpredictably, using direct testing, using criterion-referenced testing, basing achievement on objectives, ensuring that the test is known and understood by teachers and students, and providing guidance for teachers who do not understand how to teach towards the test's demands (p 44-46).

There is another group of publications which express hope in the power of tests to cause changes in the future, but which do not offer evidence that the hoped-for effect has appeared. Swain (1985) describes a test which she and colleagues developed in Canada for use in French immersion situations, and states that 'Work for washback' was one of the principles that guided the team's thinking. They believed that they could promote positive washback by involving teachers in all stages of the testing process and that they should prepare detailed support materials to help them to administer and mark the tests and 'to suggest alternative teaching-learning strategies' (pp 43-44). Swain does not comment, however, on whether the washback of the test was as positive as the team desired.

Pearson (1988) refers to tests as 'levers for change', and discusses attempts in Sri Lanka to reinforce innovations in other parts of the curriculum by introducing tests which match these innovations. He states that it is 'vital' for the tests to have a beneficial effect on teaching, but he cannot report actual outcomes as the tests had not yet been introduced when the article was written. Similar hopes are expressed by Wesche (1987), concerning the

Ontario Test of English as a Second Language, but again there is no evidence that the test achieved what it was supposed to.

Other language educators fear the harmful effects that influential tests may have: Raimes (1990) laments the 'proliferation of coaching and test-specific instruction materials' for the Test of Written English, and Norton Peirce (1993) is concerned that the TOEFL Reading Test may encourage an approach to reading texts which may not match what test-takers need outside the testing situation.

There were only a few empirical studies available before the early 1990s, amongst them Wesdorp (1982) into the use of multiple choice tests in The Netherlands; Li (1984) into the effects of the Matriculation English Test (MET) in China; Hughes (1988) into the effects of a high-stakes EAP test in Turkey; Khaniyah (1990) into the possible effects of the School Leaving Certificate in Nepal; and Shohamy (1993) into the effects of three different tests introduced into the Israeli school system. Li and Hughes report on the positive influences of the tests they were involved with, while Khaniyah and Shohamy report more negative effects. Wesdorp's study was interesting because it compared teachers' perceptions of the washback of new test with evidence from questionnaires, materials analysis and student performance. He concludes that although teachers expressed grave concerns about the effects of new testing techniques, the so-called backwash effects are a myth.

> If they do exist, they must be so weak or small that our research methods cannot detect them. (P.102)

The turning point for studies into the impact of language tests came in the early 1990s, with the publication of an article called 'Does Washback Exist?' (Alderson and Wall 1993). This article declared that the concept of washback, at least as perceived by language educators, was too vaguely defined to be useful, and that much of what was written about the power of tests over teaching was based on assertion rather than empirical findings. The authors argued that the concept should be explored more thoroughly, and they presented their own ideas about the possible effects that tests could have, whether washback was inevitable, whether its form could be predicted by the form of the test and what other factors there were that might also contribute to its nature.

Alderson and Wall presented a number of 'Washback Hypotheses', ranging from general to fairly specific, and stated that educators should specify the type of washback they want to promote and precisely what they are looking for when they evaluate whether washback has occurred.

The hypotheses are as follows:

1. A test will influence teaching.
2. A test will influence learning.
3. A test will influence what teachers teach.
4. A test will influence how teachers teach.
5. A test will influence what learners learn.
6. A test will influence how learners learn.

7. A test will influence the rate and sequence of teaching.
8. A test will influence the rate and sequence of learning.
9. A test will influence the degree and depth of teaching.
10. A test will influence the degree and depth of learning.
11. A test will influence attitudes to the content, method, etc. of teaching and learning.
12. Tests that have important consequences will have washback.
13. Tests that do not have important consequences will have no washback.
14. Tests will have washback on all learners and teachers.
15. Tests will have washback effects for some learners, but not for others. (Pp.120-121)

They argued that if washback is to be taken seriously, 'then we need to examine it critically, and see what evidence there might be that could help us in this examination'. (P.121)

Alderson and Wall also discussed the methodology that should be used when investigating washback, stating that the few studies which had presented evidence relied on surveys of teachers' self-report data or on test results rather than on analyses of classroom behaviour. They cited Smith 1991 as a good example of the range of methods which could be useful:

> We employed direct observation of classrooms, meetings and school life generally; interviews with teachers, pupils, administrators, and others; and analysis of documents. (1991:8)

They also discussed the importance of accounting for what occurs in the classroom, rather than just describing it, taking into consideration variables which emerge from the literature on motivation and educational innovation.

This article was accompanied by Wall and Alderson (1993), which described research into the washback of a new O-level examination in English in Sri Lanka. The authors made explicit statements about the type of washback they expected to find and described a complex research programme which included a baseline study (a description of teaching before the examination was introduced) and classroom observation on a large scale. The research revealed a significant amount of washback on the content of teaching, but little to none on the methodology employed by teachers. In-depth interviews with teachers indicated that there were many factors which affected how teachers reacted to the examination, including an inadequate understanding of many of the principles underlying the textbooks on which the new examination was based. Other factors included inadequate training opportunities, school management problems, difficulties in resourcing etc.

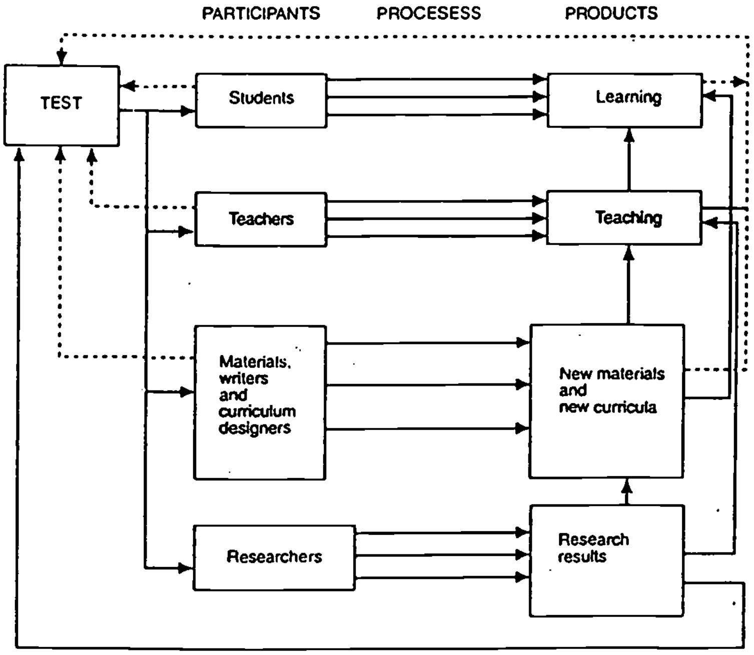
These papers led to further investigations of test impact. In 1994 the Educational Testing Service commissioned a study of washback, with the intention of incorporating new insights into its work on TOEFL 2000. The study took as its starting point the papers by Alderson and Wall (1993) and Wall and Alderson (1993), and further contributions were requested from Hughes (1993) and Bailey (1993, revised 1996). One of the products of this interaction was Hughes' observation that tests could have in impact on the participants, processes and products of education:

> The trichotomy into participants, process and product allows us to construct a basic model of backwash. The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practising the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work. (Hughes 1993:2)

Hughes advises that at least five conditions have to be met before all of the possible washback effects can occur (presumably, all of the positive effects that the test designers would wish for):

> Success on the test must be important to the learners, teachers must want their learners to succeed, participants must be familiar with the test 'and understand the implications of its nature and content', participants must have the expertise which is demanded by the test (including teaching methods, syllabus design and materials writing expertise), and the necessary resources for successful test preparation must be there. (Pp.2-3)

The second product was a comprehensive review of washback by Bailey, which attempts to explain what washback is, how it works, how positive washback can be promoted, and how washback should be investigated. She combines the hypotheses from Alderson and Wall (1993) and the Hughes' (1993) distinction between participants, processes and products to produce her own 'basic model of washback', illustrated in Figure 1 below:



Bailey identifies a number of different participants, including researchers, and the types of products that might be affected by an examination, and illustrates how these products might affect other products as well (e.g. research results could contribute to materials and curricula, or to teaching). She suggests a distinction between 'washback to the learners', which is the result of supplying 'test-derived information' to the test-takers, and 'washback to the programme', which is the result of supplying information to all the other participants in the education system. She suggests that five of the Alderson and Wall hypotheses (2,5,6,8 and 10) fit into the 'washback to the learners' area, and she provides a number of examples of the processes that learners might engage in when preparing for important tests. These range from practising items which resemble test items to enrolling in special test-preparation courses. She states that six of the hypotheses (1, 3, 4, 7, 9, and 11) fit under the 'washback to the programme' heading; however, she does not specify what kinds of processes the participants (in this case, the teachers) might participate in. She states only that there is room here for future research.

Bailey concludes her review with a set of questions which she felt should be asked of any 'external-to-programme' which was intended to promote positive washback:

- Do the participants understand the purpose(s) of the test and the intended use(s) of the results?
- Are the results provided in a clear, informative and timely fashion?
- Are the results perceived as believable and fair by the participants?
- Does the test measure what the programme intends to teach?
- Is the test based on clearly articulated goals and objectives?
- Is the test based on sound theoretical principles which have current credibility in the field?
- Does the test utilise authentic texts and authentic tasks?
- Are the participants invested in the assessment process?

5. Recent Research Into Test Impact In Language Education

Lam (1993) developed a series of 10 hypotheses to explore the impact of the New Use of English Examination in Hong Kong. He investigated whether the new exam had affected the following: the amount of time that schools dedicated to English language teaching, whether schools set aside special time to study for one particular section of the exam, the attitude of the teachers and their perceptions of the attitudes and abilities of the students, the quality of English language textbooks, the content of the teaching, and the students' language performance. He found a combination of positive and negative impact in most of the areas, and gave interesting explanations about how different factors in the context might be interacting with one another to produce a more complicated picture than the examination designers might have predicted. Of particular interest was his reference to a 'teacher culture', which meant that different teachers reacted in different ways to the examination, depending on their length of experience, their own language competence, their understanding of the aims of the test, their own motivation and commitment to the professions, and their fears of an increasing workload.

Cheng (1997) attempted to trace the impact of the revised Hong Kong Certificate of Education Examination in English (HKCEE) from the time of the first official announcement that the exam was to be revised. She discovered that the new examination had a quick and forceful effect on the types of materials that teachers were using and on the activities that they were presenting in their lessons. She suggests, however, that these changes were changes of 'form' rather than of 'substance', and that teachers were more influenced by commercial publishers' understanding of the new HKCEE than by their own. Further investigations revealed that although the exam (or the publishers' understanding of the exam) influenced lesson content and some aspects of teacher behaviour, 'it has not changed (teachers') fundamental beliefs and attitudes about teaching and learning, the roles of teacher and students, and how teaching and learning should be carried out'.

Shohamy, Donitsa-Schmidt and Ferman (1996) report on the long-term effects of two of the three tests that Shohamy had investigated in 1993, one in Arabic as a Second Language and one in English as a Foreign Language. They found that whereas both tests had had some impact on teaching when they were first introduced they had very different effects several years later. The impact of the Arabic test had almost disappeared: there was little preparation for the test, no new materials had been published for several years, there was little awareness of the test or its content, and those who were aware of the test felt that it was of poor quality. In contrast, the EFL test had continued to have impact on the content and methodology of teaching. New teaching material had appeared, there was a high degree of awareness of the exam, the test created anxiety amongst teachers and students, and the subject enjoyed high status. Shohamy and her colleagues concluded that washback can change over time and that the form that washback will assume depends on several factors: the importance of the test, the status of the language, the purpose of the test, the format of the test, and the number of skills and particular skills which are tested.

Alderson and Hamp-Lyons (1996) present an investigation into the effects of the TOEFL examination in a language institute in the United States, analysing TOEFL preparation classes and 'normal' classes being taught by the same teachers. They found that there were differences between the TOEFL and non-TOEFL classes for each teacher, but that the differences between the teachers was at least as great as the differences between the types of classes. They conclude that it is not the test alone which determines what will happen in the classroom, but rather a complex set of factors, including the status of the test, the extent to which the test is 'counter to current practice', the extent to which teachers and materials designers think about appropriate ways of preparing students for the test, and the extent to which teachers and materials designers are willing and have the ability to innovate. (1996:296)

Watanabe (1996) discusses whether there is any connection between university entrance examinations in Japan and the prevalence of grammar-translation teaching in that country. He analyses the teaching which takes place at a yobiko (examination preparation centre) in central Tokyo, comparing the lessons given by two different teachers to prepare their students for two different university entrance exams - one which emphases grammar-translation and one which does not. Watanabe concludes that it is too simple to expect that examinations will affect all teachers in the ame way: like Alderson and Hamp-Lyons (1996) he considers that the personal characteristic of the teachers (in this case, educational background and beliefs about teaching) and, possibly, the proximity of the exam in terms of time have an important role to play in how teachers conduct their lessons.

6. The Importance of Innovation Theory

It is important to note that the more recent research into washback has tended not to restrict itself to descriptions of test impact but has also attempted to explain why impact has or has not appeared and why it has taken on certain forms if it has. A significant step forward for the field of language testing would be to construct a model of washback which took account of the many factors which may play a part in determining why teachers react to tests in the way they do. A number of language educators have attempted to provide guidelines for creating positive impact in the past: Davies (1968) laid out minimum requirements in a report to the West African Examination Council, Hughes (1989) proposed seven guidelines, Shohamy (1992) set out five principles of test design, and Bailey (1996) offered eight questions that should be asked of any external-to-programme exam. Bailey also designed the model of washback in Figure 1, which took into account the Participant-Processes-Product distinction from Hughes (1994). These offerings reflect current thinking about what is desirable in language testing (validity, direct testing, criterion referencing etc) and so will appeal to educators who wish to introduce 'communicative' ideas into traditional teaching and testing settings. They also pay some attention to the people on the receiving end of this type of innovation, and remind us that teachers and students must understand the tests they are preparing for (Hughes 1989), teachers must receive help if they do not understand (Hughes 1989), schools must receive feedback from testers (Shohamy 1992), and teachers and principals must be involved at different phases of the assessment process since they are the ones who will have to implement change (Shohamy 1992).

What are not included in these lists, however, and this seems not to be discussed in language testing, are references to the settings in which tests are to be introduced, the resources that are available to support their introduction, and the way that innovation should be managed. This may not be a problem for those who are concerned with testing on a relatively small scale (within a single institution or a cluster of closely linked bodies); however, those who have to manage developments at a regional or national level would surely find it useful to consider factors beyond the test itself and whether teachers understand what is expected of them. If language testing does not pay enough attention to these factors, then it is important to examine the work being produced in other disciplines--particularly the work which is based on research findings rather than speculation.

In 1993 Alderson and Wall wrote that in order to understand how washback works 'it will be important to take account of findings in the research literature in at least two areas: that of motivation and performance, and that of innovation and change in educational settings' (p 127). In 1996 Wall made use of insights from innovation theory to explain the type of washback that appeared in Sri Lanka after the introduction of the new O-Level English examination. She reviewed a number of key concepts in educational innovation, explaining how these concepts were manifested in the setting where her data was gathered and outlining how they were being applied in more recent test development projects.

Wall's work is the most recent in a series of studies which have been published in language education, which deal with how best to introduce innovation and change. Others who have contributed to this area include Markee (1993), Kennedy (1988), and Henrichsen (1989).

Markee (1993) argues that language educators should adopt a 'diffusion-of-innovations' perspective in order to understand the many factors that affect the success or failure of new developments in language teaching:

> A 'diffusion-of-innovations perspective ... provides curriculum specialists, materials developers, and teachers with a coherent set of guiding principles for the development and implementation of language teaching innovations. Furthermore, it supplies evaluators with criteria for retrospective evaluations of the extent to which these innovations have actually been implemented. (p 229)

Markee states that language educators should make a careful analysis of the context into which they wish to introduce their innovations, and investigate each component of Cooper's (1989) composite question: 'Who adopts what, where, when, why and how?'. The 'Who' component, for example, refers to all of the participants in the innovation process: Who was the originator of the idea? Who will benefit from it? Who will bear the burden of the implementation? The 'Why' question refers to what are called the 'attributes of the innovation': Is it appropriate for the situation it is being introduced into? Is it well enough described? Will the effects be visible enough to convince the users to use it enthusiastically? Markee discusses each of the components and explains how studies carried out in fields outside language education can contribute to our understanding of how to go about dealing with educational change.

Kennedy (1988) discusses many of the same questions, and offers particularly interesting insights into the nature of the context into which innovations will be introduced. He argues that would-be innovators need to think not only about whether their proposed changes (new materials, new methodology, new tests) will fit the language classroom, but also about whether they will be in harmony or in conflict with other aspects of the setting. He proposes 'a hierarchy of inter-relating subsystems in which an innovation has to operate', which is reproduced in Figure 2:

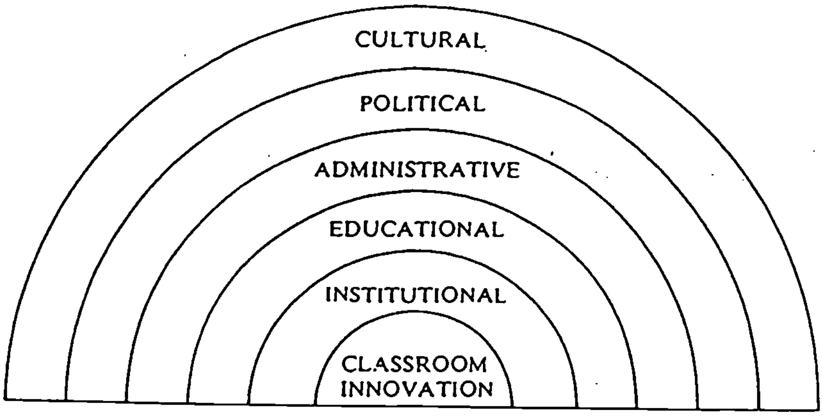


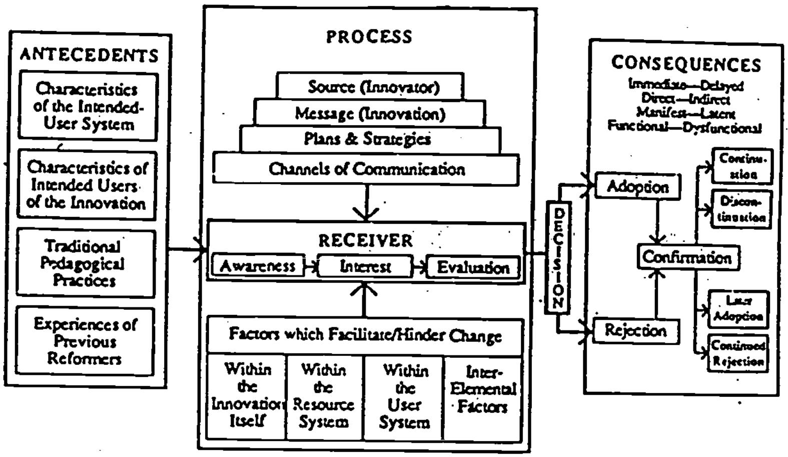
Figure 1: The hierarchy of interrelating subsystems in which an innovation has to operate

Kennedy stresses the hierarchical nature of the system, and that the outer rings in the diagram represent the most powerful influences:

> ...the cultural system is assumed to be the most powerful as it will influence both political and administrative structure and behaviour. These in turn will produce a particular educational system reflecting the values and beliefs of the society in question, a system which must be taken into account when innovating within an institution and ultimately in the classroom. (1988:332)

He warns that problems can arise if those who wish to innovate are 'outsiders' to the system, who do not understand the subsystems represented in the outer circles, or choose to ignore them, or try to change them. (op cit: 333)

Perhaps the most comprehensive model of the factors which influence change in education is that of Henrichsen (1989), who gives a detailed account of an attempt to introduce the Audiolingual Method in Japan in the 1950s. Henrichsen seeks an explanation for the failure of this innovation, and argues that it is necessary to analyse not only the 'antecedents' of the situation into which innovators plan to introduce change, but also factors within the situation which might facilitate or prevent change from occurring. The antecedents include the characteristics of the intended user system (this corresponds roughly to Kennedy's 'hierarchy of inter-relating subsystems), the characteristics of the intended users (this corresponds to the 'Who' questions in Cooper's composite question (see discussion of Markee [1993] above), traditional pedagogical practices, and the experience of previous reformers. The factors which facilitate or hinder change include the attributes of the innovation itself (this corresponds to the 'Why' in Cooper's question), the characteristics of the resource system (this refers to the innovator and the innovation teams), and so on. These factors can be seen in Figure 3 :



The analysis that Henrichsen proposes is complex and is bound to be time-consuming, but he maintains that unless a thorough investigation is carried out into all these factors, time and effort put into attempts to innovate are likely to end with failure.

7. The Challenge

It is accepted that high-stakes testing will have an impact of some kind on the participants, processes and the products of language education, but there is no fool-proof way of predicting how great the impact will be or exactly what form it will take. However, with the increase of research not only in the area of washback but also in innovation theory, we now have more understanding of the kinds of things that might go wrong in attempts to introduce curriculum change through testing. It is clear, however, that more research is needed in both these fields, so that future innovators will be able to use their resources more efficiently and more effectively.

References

Alderson, J C, C Clapham and D Wall. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J C, and D Wall. 1993. Does Washback Exist? *Applied Linguistics* 14 (2):115-129.

Alderson, J C, and L Hamp-Lyons. 1996. TOEFL Preparation Courses: A Study of Washback. *Language Testing* 13 (3):280-297.

Bachman, L. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bailey, K. 1996. Working for Washback: A Review of the Washback Concept in Language Testing. *Language Testing* 13 (3):257-279.

Cheng, Liying. 1997. How Does Washback Influence Teaching? Implications for Hong Kong. *Language and Education* 235 11 (1).

Cooper, R L. 1989. *Language Planning and Social Change*. Cambridge: Cambridge University Press.

Davies, A. 1977. The Construction of Language Tests. In Testing and Experimental Methods, Volume 4 of *Edinburgh Course in Applied Linguistics.* Oxford: Oxford University Press.

Eckstein, M A, and H J Noah. 1993. *Secondary School Examinations: International Perspectives on Policies and Practice.* New Haven: Yale University Press.

Finocchiaro, M, and S Sako. 1983. *Foreign Language Testing: A Practical Approach.* New York: Regents Publishing Co.

Frederiksen, J R, and A Collins. 1989. A Systems Approach to Educational Testing. *Educational Researcher* 18 (9):27-32.

Fullilove, J. 1992. The Tail That Wags. Institute of Language in Education 7:131-147

Haladyna, T M, S B Nolan, and N S Haas. 1991. Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution. *Educational Researcher* 20 (5):2-7.

Heaton, J B. 1990. *Classroom Testing.* Harlow: Longman.

Henrichsen, L E. 1989. *Diffusion of Innovations in English Language Teaching: The ELEC effort in Japan,* 1956-1968. New York: Greenwood Press.

Heyneman, S P, and A W Ransom. 1990. Using Examinations and Testing to Improve Educational Quality. *Educational Policy* 4 (3):177-192.

Hughes, A. 1988. Introducing a Needs-based Test of English Language Proficiency into an English-medium University in Turkey. In *Testing English for University Study,* edited by A. Hughes. London: Modern English Publications.

Hughes, A. 1989. *Testing for Language Teachers*. Edited by M. Swan, Cambridge handbooks for Language Teachers. Cambridge: Cambridge University Press.

Hughes, A. 1993. Backwash and TOEFL 2000. Unpublished manuscript. University of Reading.

Kennedy, C. 1988. Evaluation of the Management of Change in ELT Projects. *Applied Linguistics* 9 (4):329-342.

Khaniyah, T R. 1990. Examinations as Instruments for Educational Change: Investigating the Washback Effect of the Nepalese English Exams. Unpublished PhD dissertation, Department of Applied Linguistics, University of Edinburgh, Edinburgh.

Lam, H P. 1993. Washback - Can It Be Quantified? Unpublished MA thesis, University of Leeds, Leeds.

Kirkland, M C. 1971. The Effect of Tests on Students and Schools. *Review of Educational Research* 41(4):303-350.

Li, X. 1990. How Powerful Can a Language Test Be? *Journal of Multilingual and Multicultural Development* 11 (5):393-404.

Madaus, G F. 1988. The Influence of Testing on the Curriculum. In *Critical Issues in Curriculum: Eighty-seventh Yearbook of the National Society for the Study of Education*, edited by L. N. Tanner. Chicago: University of Chicago Press.

Madsen, H S. 1983. *Techniques in Testing*. Oxford: Oxford University Press.

Markee, N. 1993. The diffusion of innovation in language teaching. *Annual Review of Applied Linguistics* 13:229-243.

Morrow, K. 1986. The Evaluation of Tests of Communicative Performance. In *Innovations in Language Testing*, edited by M. Portal. Windsor: NFER/Nelson.

Norton Peirce, B. 1992. Demystifying the TOEFL Reading Test. *TESOL Quarterly* 26 (4):665-689.

Pearson, Ian. 1988. Tests as Levers for Change. In *ESP in the Classroom: Practice and Evaluation*, edited by D. Chamberlain and R. Baumgardner. London: Modern English Publications.

Popham, J. 1987. The Merits of Measurement-driven Instruction. *Phi Delta Kappa* May: 679-682.

Raimes, A. 1990. The TOEFL Test of Written English: Causes for Concern. *TESOL Quarterly* 24 (3):427-442.

Shohamy, E. 1993. *The Power of Tests: The Impact of Language Tests on Teaching and Learning.* Washington, D.C.: The National Foreign Language Center.

Shohamy, E. 1992. Beyond proficiency testing: a diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal* 76 (4):513-521.

Shohamy, E, S Donitsa-Schmidt, and I Ferman. 1996. Test Impact Revisited: Washback Effect over Time. *Language Testing* 13 (3):298-317.

Smith, M L. 1991. Put to the Test: The Effects of External Testing on Teachers. *Educational Researcher* 20 (5):8-11.

Spolsky, B. 1995. The examination-classroom backwash cycle: Some historical cases. In *Bringing about Change in Language Education*, edited by D. Nunan, R. Berry and V. Berry. Hong Kong: University of Hong Kong: Dept of Curriculum Studies.

Stobart, G and C Gipps. 1997. *Assessment: A Teacher's Guide to the Issues. 3rd edition.* London: Hodder & Stoughton.

Swain, M. 1985. Large-scale Communicative Testing: A Case Study. In *New Directions in Language Testing*, edited by Y. P. Lee, C. Y. Y. Fox, R. Lord and G. Low. Hong Kong: Pergamon Press.

Wall, D. 1997. Test Impact and Washback. In *Language Testing and Assessment*, edited by C. Clapham and D. Corson. Dordrecht: Kluwer Academic Publishers.

Wall, D. 1996. Introducing new tests into traditional systems: insights from general education and from innovation theory. *Language Testing* 13 (3): 334-354.

Wall, D and J C Alderson. 1993. Examining washback: the Sri Lankan Impact Study. *Language Testing* 10 (1):41-69.

Watanabe, Y. 1996. Does Grammar-Translation Come from the Entrance Examination? Preliminary Findings from Classroom-based Research. *Language Testing* 13 (3):319-333.

Weir, C J. 1990. *Communicative Language Testing*. Hemel Hempstead: Prentice Hall.

Wesche, M. 1987. Second Language Performance Testing: The Ontario Test of ESL as an Example. *Language Testing* 4 (1):28-47.

Wesdorp, Hugo. 1982. Backwash Effects of Language Testing in Primary and Secondary Education. *Journal of Applied Language Studies* 1 (1): 40-55.

Myths of Testing and the Icons of PET, TOEFL and IELTS

GRAEME TENNENT
United Arab Emirates University

In giving this talk I should perhaps summarize the main suppositions behind it. The first is that there is a tendency towards using established British and American tests such as the three in the title, PET, IELTS and TOEFL. This seems to be the case: Higher Colleges of Technology in UAE use the PET and IELTS and Sultan Qaboos in Oman also used the PET and IELTS in a modified form. In the university here they have used the Michigan Placement Test and the issue is never far away that we should enter our students for TOEFL, IELTS or some such exam.

My next supposition is that these examinations have taken on an iconic aura—they have become icons of respectability. Cassell's dictionary defines icon as:

> *in the Eastern church a sacred image, picture, mosaic or monumental figure of a holy personage usually endowed with miraculous attributes.*

There is a sense in which the signifier has become detached from the signified. An example might be the case of Princess Diana where the iconic signifier of a Mother Theresa-Christ-Virgin Mary figure seems remote from the signified, a rather intellectually limited product of a dysfunctional aristocratic family with a penchant for cavalry officers, Egyptian playboys and easy emotionalism. I think the examinations referred to have equally become detached from their initial signification, certainly in the Gulf context.

Before I look at the iconisation and the exegesis of that process and my own iconoclastic intentions, it would be appropriate to consider the said PET, TOEFL and IELTS examinations. Forgive me if I am boring the initiated.

The PET

The Preliminary English Test is described in The Longman's Guide to English Language Examinations as:

> *An elementary examination representing about 350 hours from beginner level and testing appropriate oral and written skills. Successful performance in the test should equip students with language abilities which would enable them to enjoy a normal social life in an English-speaking country in line with the recommendations of the council of Europe's Threshold Level.*

The Threshold Level suggests the following areas of interest:

Social interaction with native and non-native speakers of English
Dealing with official and semi-official bodies
Shopping and using services
Visiting places of interest and entertainment
Travelling and arranging for travel

Using media for information and entertainment.
Medical attention and health
Studying for academic/occupational/social purposes

I am glad they tacked on the last bit although I am not quite sure what studying for social purposes is.

So we have this examination designed for European holiday makers being used as an entrance placement test in one gulf university and as an exit test in another, although, to be fair, HCT use PET as a kind of quality check on their own exit criteria for their certificate course. That is something I will come to later. Lets leave this anglo-centred world of railway stations and advertisements for semi-detached cottages and lonely hearts for the more rarified academic world of IELTS and TOEFL.

IELTS or The International English Language Testing System is the creation of the British Council, the University of Cambridge Local examinations Syndicate and The International Development program of Australian Universities and Colleges. Its main purpose is:

> *" to find out if your ability in English will meet the demands of a course of study or training in Britain, Australia or anywhere else where the teaching is done in English."*

So this is the exam to determine proficiency before entry into study in Britain or Australia and it tends to reflect the social aspects of student life particularly in the listening section. Interestingly it has virtually no predictive value of success in study according to Alan Davies 1990. However despite its faults and they are considerable it does appear to offer a fair statement of proficiency. Whatever its claims to universal application in English medium education, it is primarily a test for those going to UK or Australia.

It uses a banding system quite similar to the ESU Bands where university requirements would be about 6 plus depending on the course. It is of limited value in determining proficiency at the lower end of the scale.

TOEFL or Test of English as a Foreign Language is probably the best known as it is used widely outside the USA as a means of determining language ability. Its aims are stated as:

> *"to provide valid scores indicating the English proficiency of non-native speakers seeking admission to colleges and universities in the United States and Canada. Also used in other countries by institutions where English is the medium of instruction ."*

Davies and West (1989) draw an equivalence between TOEFL scores and the Cambridge exams.

TOEFL	CAMBRIDGE
677	Diploma of English Studies
550-500	Certificate of Proficiency
475-450	First Certificate in English

Like IELTS, TOEFL is of little use in determining the value of lower scores. Quite simply that is not what it is meant for.

I shall not go into the English-Englishness or the American Englishness of these examinations. We are all aware of the confusions of condos, duplexes, B&Bs, cottages and semi-detacheds. The computer on which I write this shows equal confusion!

Suffice it to say that these are examinations with specific functions despite their attempts to broaden their catchment areas. The rhetoric of each of them shows this urge to widen their markets. Pennycook 1994 in his book THE CULTURAL POLITICS OF ENGLISH AS AN INTERNATIONAL LANGUAGE (pages 156 and 157) outlines the hard business approach of TOEFL and the British Council in the hugely profitable market of examinations. Pennycook gives figures for 1987 where TOEFL generated 14 million dollars and the British examination boards around 9 million dollars. These are not philanthropists! I quote from Pennycook quoting a British Council Corporate plan of 1 990:

> *Its goals are not only to promote a wider knowledge of the English language abroad but to increase the demand for British examinations in English as a foreign language.*

I do not condemn them for not being philanthropic. The British Council and USIS are here to promote their own products. They are, to use the jargon, players in the new world of English as an International Language. For we are beyond a British centred English and even an American centred English although the latter's influence is certainly the greater. Just as we have entered the time of post-colonial literature we should be entering the world of the post British and American Imperialist language.

Why then are we still suffering from a dependency on these examinations? Exams which are fine in themselves in their own situations but which are not directly appropriate to our experience here in the Gulf.

I shall argue that there is still a post-colonial dependency which leads to the iconisation of these exams as measures of reliability and validity, a process where, as I suggested before, the signifier has become detached from the signified creating the semi-mystical icon.

If we look at the Chart of examinations produced by the ESU (English Speaking Union) in 1989, we can get a glimpse of why testers are drawn to these examinations.

The reasoning goes like this: we have a low level course. Lets use PET which features around band 3 or 4 or we have intermediate to advanced students lets use TOEFL or IELTS. If our students prosper in these exams then we can state our own courses are at those band levels and thus have some credibility. It doesn't matter that these exams are not quite appropriate to our situation. They are at the right band level and they carry the icon of respectability, the British and the American trademark. Let me cite a few examples.

In Oman they used the PET as an entrance placement test or at least part of it. The very fact that only part was used immediately detracts from it being the PET and of course PET is not a placement test. This was replaced by a specially designed IELTS for lower levels. IELTS is not designed for lower bands so what they were using was not IELTS but something new which claimed the icon status of the original. So why bother with IELTS? Why not use your own test incorporating the ideas of the IELTS test if that's what you think constitutes a good placement test? Answer, because it was not the test but the icon of standardised respectability which counted. Informal sources tell me that teachers did not like the PET but administrators loved it. Hearsay but possibly revealing.

The same examinations are used somewhat differently at the Higher Colleges of Technology here in UAE. They use the authentic full-blown PET as an exit test for their low level certificate programme. In fact it is an adjunct to their own examination intended therefore as a check on their own assessment. The logic is reasonable: if our tests are at band 3/4 then they will correlate to the PET. The idea is that PET is used as an outside check on standards, a one-off picture of proficiency, but the effect is somewhat different. Teachers and students are aware that it is necessary to pass both the internal and external examinations in order to pass the course so naturally as good teachers they try to prepare students for the PET. This is the classic washback effect where the course begins to take on aspects of the PET course. It may end up that it simply becomes a PET course which of course is something different from what was intended. PET ceases to be a snapshot of proficiency but becomes the course itself, hideously inappropriate as it may be. In a similar way the IELTS is used on the higher level diploma programme with, I am told, similar washback effect. Here we can see the power of the icon and the desire for external validity.

To give another example: When I arrived here I took charge of testing for the foundation course where two thousand students had to be placed in four levels. It was decreed that the Michigan Placement test would provide the aura of objectivity and external respectability. The results were predictable: 1,950 students should have been in level one, 30 in level two, fifteen in level three and five in level four. I exaggerate but of course the students performed in a hideous clump which made discrimination impossible. Fortunately that experiment was abandoned but I'm sure the search goes on for the external test which will do the business. Even in my present position in the department of English Literature and Language, from time to time there will be a call for us to use TOEFL. It hasn't happened but if it does I have no doubt that all literature will be forgotten and we will become a TOEFL preparatory department.

So why is it that testers and administrators are so drawn to these icons of testing?

One reason often cited is that external agencies such as employers need to understand the levels of English proficiency reached by students. There is some sort of extraordinary myth that employers know all about PET, IELTS, are fully appreciative of TOEFL scores and can tell a Band 3 from a Band 4. This is patent nonsense. Most teachers are not fully conversant with these issues! They may assume that a graduate of a British or American University will have a high level of proficiency but that is a different matter. So what is it?

I have suggested that there is a post colonial dependency syndrome. We can not know if we are doing all right unless we carry the stamp of British or American approval. Our teachers cannot be trusted and our testers and administrators cannot be relied upon to assess honestly, and anyway it has no meaning unless it is tied to, affiliated to, accredited to, or standardised by some British or American institution.

Or to summarise: we are not responsible enough to assess our own students.

If teachers, testers and administrators are not equipped to give accurate descriptions of proficiency and to ensure that standards are maintained then they should be dismissed and replaced by those who can do the job! Why not hire the TOEFL, PET and IELTS designers to come and write and administer tests for our courses? No, of course not. The people here have been hired because they are professionals, skilled and trained for the job. They are quite capable of testing, teaching, writing courses, administrating as anybody else.

What is lacking is the trust between those involved. There is also a lack of confidence in ourselves. There is an unwillingness to take full responsibility. There is an unwillingness to accept the passing of the British American hegemony over the English language and to assert our independence in the new order of English as an International language. It is not just native speakers who carry their "British is Best" and "America is right" slogans. Nonnative speakers educated in Britain and America have a vested interest in the status of their qualifications and perpetuate the iconic status of Anglo-American qualifications. Hence the dependency on icons of reliability, the touching of the magic talisman in the hope its powers will rub off.

To sum up: I do not believe we need these tests. We are quite capable of designing our own. We are quite capable of banding to provide comparisons with other qualifications. However until teachers, testers and administrators have confidence in each other and are willing to take full responsibility for the education of their students, we might as well continue to bow before the icons of PET, TOEFL and IELTS.

Copyright Infringement:
What are the Legal Rights of Educators as Test Writers?

CHRISTINE M. CANNING
United Arab Emirates University

Before making important copyright decisions, consult a knowledgeable copyright lawyer, the Copyright Office, or a trusted publisher or agent who has an up-to-date understanding of the law. As a writer, a teacher, and/or a test developer, you are trading in your words, your talents, and what you are really selling to your institution, publisher or editor is your copyright. Copyright protects nearly every original piece of work that you create.

Copyright does not protect your idea, but it does protect how you "express" your idea. Publication is no longer the key to obtaining the copyright of a work. It is true that certain foreign works originally published without notice had their copyrights restored under the Uruguay Round Agreements Act (URAA), but now it is important to create a copyright notice on materials created by an individual or institution. It is important for teachers and test writers to note that copyright is secured automatically when the work is created and a work is considered created when it is in its fixed copy or best edition form.

There are many misconceptions about copyright law. Firstly, there is no such thing as an "international copyright" that will automatically protect an author's writing throughout the entire world. Protection against unauthorized use in a particular country depends, basically, on the national laws of that country. However, most countries do offer protection to foreign works under certain conditions, and these conditions have been greatly simplified by international copyright treaties and conventions. World intellectual property organizations follow and adhere to United States laws as the basis for decisions. Only twice has US law been amended by world intellectual property organizations in order to protect foreign works under the GATT and URAA agreements. If you choose to copyright in a foreign country you will be advised to send the work to the United States for an interim stamp. Therefore, it is simplest and best to register your work with the United States of America Copyrights Office. The American government has provided in its general laws that copyright protection in the USA is "available for all unpublished works, regardless of the nationality or domicile".

Copyright is a form of protection to the authors of original or intellectual works or properties for both published and unpublished works. A copyright is not just one right, but a bundle of rights. Copyright protection subsists from the time that the work is created in its fixed form. According to section 106 of the 1976 Copyright Act, the owner of a copyright is entitled to:

- Reproduce the copyrighted work in copies or phonorecords.
- Prepare derivative works based upon the copyrighted work.
- Distribute copies of the phonorecords of the copyrighted work for public sale or other transfer of ownership, or rental, lease or lending.

- Produce the copyright publicly and to display the work. (In the case of literary, musical, dramatic and choreographic works, pantomimes, and motion pictures and other audio visual works.)

However, there are some b asic facts that test writers should understand, before they lend, borrow, or author any materials, if they want to prevent copyright infringement. It is also important to note, that professional test writers are not given the same lenient "use" of copyright materials as teachers.

How does one claim copyright as a teacher or test writer? As was previously mentioned, copyright protection subsists from the time the work is created in its fixed form. The copyright of the work of authorship immediately becomes the property of the author who created the work. Only the author or those deriving rights from the author can rightfully claim copyright. An original work may be submitted formally to the copyright office by first telephoning 001-202-707-9100 and ordering the TX form for nondramatic works. Then, the author must follow the statues of Title 17 which require the "best edition of a work". The law states that to submit printed textual matter the following guidelines must be met:

A. Paper, binding and packaging should be:

1. Archival-quality rather than less-permanent paper

2. Hard cover rather than soft cover

3. Library binding instead of commercial binding

4. Trade edition rather than book edition

5. Sewn rather than glued bindings

6. Stapled rather than plastic-bound

7. Bound rather than loose leaf, except when future loose-leaf insertions are to be issued. This can include binders and indexes as they are part of the unit for publishing, sales or distribution.

8. Slip case covers instead of nonstop cased

9. With protective folders rather than without for broadside submissions

10. Rolled rather than folded for broadside works

11. With protective coatings rather than without (except broadsides which should not be coated

If there is computer software that also needs to be copyrighted officially, then the computer information must be submitted in the following format:

1. With documentation and other accompanying materials that make it the best edition.

2. With the best edition of the work accompanying the program.

3. Not copy-protected instead of copy-protected.

4. Formats:
   a. PC DOS or MS DOS on a 5 1/4" disk or a 3 1/2" disk
   b. Apple Macintosh works should be on a 3 1/2" disk and optical media should be in its best edition such as a CD ROM.

All works and forms must be accompanied by a money order for 20 USD.

If a teacher or test writer does not want to formally submit his or her work to the copyrights office, they can claim copyright by simply repeating the following steps and they are protected:

1. Write the symbol for copyright or the word "copyright"

2. The year of first publication

3. The legal name of the owner of the copyright. No abbreviations or nick-names are allowed.

   Examples: Copyright 1998 Christine M. Canning
   Copyright 1998 TESOL Arabia

Test writers and teachers can avoid copyright infringement by using works that are not protected. It is important for teachers and test writers to note that an idea can not be copyright; instead, a person can only copyright how they "express" the idea. The 1976 Copyright law states that the following works are not protected:

1. Works in public domain.

2. Titles, names, short phrases and slogans; familiar symbols and designs, ingredients, colors.

3. Ideas, procedures, methods, systems, processes, concepts, principles, discoveries, devices as distinguished from a description, explanation, and illustration.

4. Works consisting entirely of information that is common property or contains no original authorship (ie., standard calendar).

The first unprotected work is probably the most important to educators and test writers. To avoid any form of copyright infringement, it is important to select work in the public domain. For a work to be in the public domain, it must meet one of the following criteria:

1. It was published more than 75 years ago and the copyright was not renewed.

2. The work was first published or copyrighted between 1930-1963 and does not show copyright renewal.

3. A work published without prior notice between January 1, 1976 and February 28, 1989.

4. A work created by a government employee while in office as part of his or her duties.

For example, if you were to produce Shakespeare's *A Winter's Tale*, you may use a man in a bear suit instead of an actual grizzly bear because his works are in the public domain. However, if you were to cast a person of color in a Tennessee William's play you would be in violation of copyright law because his works are in the private domain. The author never intended race to be introduced as a factor into his work. Therefore, black actors who have sued and challenged the copyright to be the main character in *Cat on a Hot Tin Roof* have been denied. Copyright protects the spirit of the work.

If as a teacher or test writer, you are unsure of the copyright status of a work, you can research a record by submitting the following:

1. Title of the work and any possible variants.

2. The names of the author, including any possible pseudonyms.

3. The name of the probable copyright owner, which may be the publisher or producer.

4. The approximate year when it was published or registered.

5. The type of work involved.

6. Something to identify the work, book or periodical.

7. A possible registration number.

8. 20 USD.

9. Mail all of the information to:

    Reference and Bibliography Section
    LM-451
    Copyright Office
    Library of Congress
    101 Independence Avenue—SE
    Washington, DC 20599—6000

It is important to note that not all searches are conclusive. For example, obtaining foreign copyrights may not hold up in other countries without a second interim copyright from the US Government. A teacher or test writer can check the copyright status by looking at how long the copyright on the work lasts. The general rules are as follows:

1. Before 1977, about 50 years.

2. 1978 - present and works "made for hire" - 100 years from creation or 75 years from publication.

3. In many foreign countries, the extensions may only last 20 years.

Teachers, test writers and administrators would be well advised to look at the renewal of the copyright. Anything before 1977 can be optionally renewed; however, anything after must go through renewal of registration for the copyright to be owned. Application for a copyright must be applied for in the 28th year after the original certificate by December 31st of that year which is inclusive of derivative works.

If the copyright claimant dies, the claimant can terminate the copyright or renew it in his/her name. To check on the claimant the law is clear on the order to who it belongs to for legal purposes. When the owner of the copyright dies, it goes automatically to the widow or widower of the author, then the blood children of the author. If these people do not exist, then the will of the author is considered. If no one is listed in the will, then it goes to the executor. If the executor declines, it goes to the next of kin.

Teachers and test writers should be aware that just because a work is registered, it does not mean that it is copyrighted. Registration is not a condition of copyright protection. The only thing that registration does is that it establishes a public record of the copyright claim. This means that before an infringement suit may be filed in a court of law, registration is necessary for works of US origin and for foreign works not originating in a "Berne Union" country. Registration allows the claimant to ask for statutory damages and attorney's fees provided you register it within 3 months. Most importantly, it allows Customs Officials to prohibit copies of your work from illegally entering other countries.

This does not mean that educators cannot use a copyrighted work. As teachers and test writers you are allowed some leniency under the fair use laws.

Fair use allows you to use a portion of a copyrighted work for such purposes as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research. However, to claim fair use as an educator, the following must be proven under sections 106 and 106a:

1. The purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes.
2. The nature of the copyrighted work.

3. The amount and substantiality of the portion used in relation to the copyrighted work as a whole.

4. The effect of the use upon the potential market for or value of the copyrighted work.

The courts have ruled that a "specific exemption freeing certain reproductions of copyrighted works for educational and scholarly purposes from copyright control is not justified..." However, the international courts agreed that "There is a need for greater certainty and protection for teachers". As a result the following has been added to the concept of the fair use laws:

1. The fact that the work is unpublished shall not itself bar a finding of fair use, if the above is considered.

2. International courts have ruled time and time again on this doctrine because again and again, no real definition of the concept has ever emerged.

3. Photocopying is for "non-profit educational purposes":

The following rules under the 1976 Copyright Act and the Revised March 1, 1989 Copyright Act can be applied to educators without infringing on the copyrights of others:

1. No more than 9 copies from copyrighted works can be given to a student in a single class.

2. Students can't be charged more than the photocopying fee.

3. Teachers/test writers are prohibited from:

    a. Copying in order to create their own textbooks as a substitute for a compilation, anthology, or a collective work.

    b. Copying from works intended to be "consumable" in the course of study or of teaching. These works include: workbooks, exercises, standardized tests, and test booklets and answer sheets—and like consumable materials.

    c. Copying as a substitute for the purchase of books, publisher's reprints, or periodicals.

    d. Copying without a notice of inclusion of the copyright status

The courts have ruled that an educator may:

1. Copy an article for scholarly research or to teach his/her class:

    a. by copying no more that one chapter of a book.

b. by copying no more than an article from a periodical or newspaper.

c. by copying a short story, short essay, or short poem, whether or not from a collective work.

d. by copying a chart, a graph, a diagram, a drawing, a cartoon, or a picture from a book, periodical, or newspaper.

e. make multiple copies of a work provided it meets the requirements of the brevity, spontaneity, and cumulative effect test.

Teachers should prove that when photocopying or using a copyrighted work that they have met the requirements of the brevity, spontaneity and cumulative tests, as they are stated below under the general laws of Section 107 of the Copyright Revision Bill:

The Brevity Test:

1. Less than 10% of a work or @ 2500 words

2. No more than 1 picture or 1 illustration

3. Specialty works should not be photocopied. If they must be copied, no more than 10% or @ 2500 words may be borrowed from the given text.

The Spontaneity Test:

1. The copying is at the instant and inspiration of the individual teacher.

2. The inspiration and decision to use the work and the moment of its use for maximum teaching effectiveness are so close in time that it would be unreasonable to expect a timely reply to request permission.

The Cumulative Test:

1. The copying of the material is for only one course in the school in which the copies are made.

2. No more than one short poem, article, story, essay or two excerpts may be copied from the same author, nor more than three from the same collective work or periodical volume during one class term.

3. Section 1 & 2 may be voided if it refers to a current news event.

If you choose, as a teacher or test writer, to create your own original works, who owns it? Does it belong to you or your employer? What is a "work made for hire" and who owns the rights to these works? Section 101 of the copyright statute states:

1. A work prepared by an employee within the scope of his or her written contracted employment.

2. A work specially ordered or commissioned for use as a contribution to a collective work (including instructional texts, tests, and answer materials for tests) provided that the parties sign an official agreement that the work shall be considered a work for hire.

3. Authors of joint works are joint co-owners on copyright.

4. Copyright in each separate contribution to a periodical or other collective work is distinct from copyright in the collective work as a whole and vests initially with the author of the contribution.

Mere ownership of a book, manuscript, painting, or any other copy or phonorecord does not give the possessor the copyright. The law provides that transfer of ownership of any material object that embodies a protected work does not itself convey any rights in the copyright. The creator of an original work generally owns the copyrights. There is an exception, however, for "works made for hire". In other words is it the party who commissions and pays for the work, rather than the actual creator, who owns the copyright? And when is a work actually "for hire"?

First, unless expressly excluded by contract, all works created by employees within the scope of their employment are "for hire". This will normally not include works created on your own time that are unrelated to your employment. So, if you are employed by a newspaper, or hired by a software publisher to write documentation, your employer owns the copyrights in the works you have been paid to create. If you use the copies of these works at your next job, you are infringing on your former employer's copyrights.

Second, certain specified categories of works including translations, compilations, and parts of audiovisual works are considered "for hire" if they have been specially commissioned and a signed writing identifies them "for hire". Therefore, if you are not an employee and you have not agreed in writing that your work is "for hire" or otherwise been assigned your rights, you will generally continue to own the copyrights of your works even if others paid you to create it. Although it is important to note, that they will have the right to use your work for the express purposes for which they paid you. Teachers and test writers must educate themselves about their legal rights in order to protect their works from being unfairly used or exploited by their employers.

Author's Note: This article is not meant to take the place of standard legal advice. A Copyright Lawyer should be contacted before pursuing any legal matter. The author and editors waive their rights of responsibility for parties who ignore this warning of obtaining legal consul. Copyright:1998 Christine M. Canning.

References

Federal Register, 62:191, P.51065.

Gopher: marvel.loc.gov/telenet: locis.loc.gov

1989 Copyright Act.

Sections 101, 106, 106a 107, 108-110 of the 1989 Copyright Act.

Library of Congress Circulars 1-22 of the Copyright Act.

1992 Amendment to the Copyright Act.

C-Testing: A Theory Yet to be Proved

NEIL MCBEATH
Royal Air Force of Oman

Introduction

The physicist Dr. Stephen Hawking once remarked "I rely on intuition a great deal, but I have to go on to prove it" (Desert Island Discs, BBC Radio 4, 25/12/92). This is precisely the approach that I have adopted with regard to C-Tests, which were first introduced to Oman in 1986. This followed the dissemination of a policy paper (Cleary; 1986) which was less a discussion document than a statement of intended policy. C-Testing had arrived, it worked, and the Examinations Cell of the Directorate of Education intended to incorporate a C-Test in the Sultan's Armed Forces English Level 3 final examinations, as well as employing them as an assessment device for new recruits.

This decision caused some controversy, as the C-Test is a modification of cloze procedure. It is based on the mutilation of text by the "rule of two" (Raatz & Klein-Braley 1982; 123). Four to six short authentic texts are chosen, and in each case the first and last sentences appear in their entirety. Between these sentences, every second half of every second word is deleted, unless the word has an odd number of letters, in which case the larger portion is deleted. Single letter words like "I" and "a" are ignored in the word count, but the entire C-Test should contain 100 deletions, and only entirely correct restorations are counted in assessment. The C-Test is best described by Klein-Braley (1985) and it resulted from her research at the University of Duisburg. Specifically, she set out to examine Taylor's (1953) findings, which led to the development of cloze tests. She concluded that in the case of cloze techniques, "the anchor point in the measurement scale is missing" (P.8)

It is Klein-Braley's contention that any test constructed for foreign language learners ought to be one in which native speakers obtain a perfect score, and yet Taylor's research had proved that native speakers were failing to do this with cloze.

The Prima-facie Case against Face Validity

By deleting more words, and every second word up to a maximum of 100, while leaving the first letter visible, it was intended that the C-Test would aid native speakers to the extent that they would be able to achieve a perfect score, but this same technique was found to impede non-native speakers when it was first used in Oman. Immediately after the introduction of the C-Test, there were reports of Omani servicemen becoming so frustrated that they abandoned C-Tests when they were only half completed (Andrewartha; 1987).

Bachman (1990; 288) points out that "the 'bottom-line' in any language testing situation, in a very practical sense, is whether test takers will take the test seriously enough to try their best" and the Omani servicemen's rejection of the test appeared to indicate that, for some at least, the C-Test failed to match this criterion. More importantly, the adverse reaction appeared to support Anderson's (1971) finding that students who scored 44% or less on an exact-word cloze passage had reached such a "frustration level" that they were unable to make sense of the text even when given help by a teacher.

Zeichner (1988) has indicated that most student examinees regard testing as a meaningful experience, and so a total rejection of the exercise is a powerful statement of the candidate's affective disposition. The feedback, though negative, is an indication that, for the examinee, the C-Test lacks face validity, as the FL student is immediately placed in a disadvantageous position.

The schema theory model (Bartlett 1932; Rumelhart and Ortony 1977; Rumelhart 1980; Carrell and Eisterhold 1983) claims that the comprehension of a text is an interactive process between the reader's (or listener's) background knowledge and the text. Rumelhart (1977;267) discusses the text "The policeman held up his hand and stopped the car." explaining how this, in North American society, provides a schema of a traffic cop signalling to a driver to stop. This, in turn, suggests other schemae not specifically mentioned in the text - the policeman will be wearing a uniform, the car will stop because the brakes have been operated etc.

Schema theory is part of what van Brederode (1996;30) characterises as "knowledge structures" and he cites Fillmore's (1982) work on frames, Lakoff's (1987) concept of Idealised Cognitive Models and Shank's (1971; 1982) work on scripts. He does not mention, however, that "knowledge structures" may be culture specific (Meyer 1991; Robin 1985) with the attendant possibility of cross-cultural misunderstanding.

An L1 student, faced with a C-Test, can bring socio-cultural knowledge to bear on the initial sentence, and this will assist in bottom-up processing to reach a composite meaning for that sentence. This composite meaning, in turn, will be combined with socio-cultural knowledge to construct conceptual dependencies to help predict the meaning of the next sentence, and so on. FL students frequently lack the socio-cultural knowledge to construct conceptual dependencies in this way, and hence are less likely to frame the textual context or frame the text's continuation in this way.

This critique of the C-Test, however, has been ignored by the main contributors to the literature (Klein-Braley 1984;1997; Klein-Braley and Raatz 1984; Strawn 1985; Cleary 1988; North, Hirst, Petty and Scott 1988; North 1991; Dornyei and Katona 1992; 1993/94; Connelly 1997) most of whom are enthusiasts who have based their ideas on Klein-Braley's original research, without attempting to investigate her ideas any further.

Strawn, Cleary and North, however, have made alterations which take their work beyond the limits of Klein-Braley's research, and have applied C-Tests to EFL situations. The bulk of the construct validation carried out by Klein-Braley was conducted with German speakers, and had little connection with EFL, but Strawn has employed English language C-Testing with

University level Koreans, Cleary with Omani servicemen, and North with international students at private language schools in Germany and Britain.

Four Hypotheses Examined and Discussed

Klein-Braley (1985) sought to establish the validity of four separate hypotheses:

1. If the same C-Test is administered to subjects at different stages of language development, then the C-Test scores will become successively higher as the subject becomes more proficient in the language. (P.84)
2. Subjects learning a language 'naturally' will exhibit similar behaviour on C-Tests in that language. (P.86)
3. If texts have an inherent 'C-Test processing difficulty' which is independent of the subject groups involved then it will be possible to discover characteristics of the texts which can be used to predict the rank order of difficulty of texts, possibly even the actual empirical difficulty levels, for specific subject groups. (P.89)
4. Learners with more efficient language processing strategies will make higher scores on C-Tests. (P.97).

In the case of Hypothesis 1, data was gathered from an unspecified number of L1 speakers of English aged nine, eleven and thirteen, who were given an English C-Test, and from 3 German speakers who were administered a German C-Test. These German L1 speakers were attending secondary schools at Gymnasium, Realschule and Hochschule level, but Klein-Braley gives no indication of where these schools were located. If they were located near her home university in Duisburg, in North-Rhine Westphalia, this might be a significant factor, as this is an area of Germany where the language of adolescent students is unlikely to be marked by non-standard dialectal interference. Clyne (1984; 68) indicates that dialects "are used far more in the south than in the north of the Federal Republic where they are more stigmatized" and that they are more likely to be used by those engaged in agricultural occupations than by workers in towns.

A northern, urban setting such as Duisburg is likely to produce speakers whose everyday use of German approximates to the "High" standard taught in schools, where in Switzerland a diglossic situation pertains, "High" German being used for formal spoken and all written communication, and "Low" German being used informally. It is possible, therefore, that Klein-Braley's research would have produced different results had it been carried out in a rural area, or in Southern Germany, Austria or Switzerland.

As it stands, the evidence produced by Klein-Braley validates Hypothesis 1, but only so far as it applies to L1 learners studying their own language. The data is confined to children and adolescents from a West European background, and may not be applicable to adult learners, L2 learners or FL learners.

Hypothesis 2 was investigated using data gathered from 197 German L1 learners, 203 Greek speaking and 186 Turkish speaking learners of German as L2, all at primary school level. The evidence validated the hypothesis, but only with regard to children, and children learning German within a host environment.

Dulay, Burt and Krashen (1982) analyse the results of a number of studies on child and adult language acquisition order, and declare that "children of different language backgrounds learning English in a variety of host country environments acquire eleven grammatical morphemes in a similar order" (Pp.207-09). They also state that "the contours for the acquisition sequences of the children and adults studies are very similar (P.209) but the conclusions of research conducted on English acquisition order need not be valid with respect to German. No comparable research appears to have been conducted for German and so whether Klein-Braley's findings can be extended from their undoubtedly valid German conclusions to encompass English L2 learners, adult learners and EFL learners is another matter, and one on which no research appears to have been attempted.

Hypothesis 3 was tested with data drawn from three different groups. The first consisted of 276 German speaking L1 students at two, different, primary school levels, who were given 16 tests, eight of which were common to both groups. The second cohort consisted of 67 German university students of English, who were given English texts, and the third group consisted of "Finnish FL learners" - Finns who were, in fact, learning English, as they were used to check the texts given to the German university students. Once again, the hypothesis was validated with regard to German L1 students at primary school level, leaving no doubt that C-tests are valid assessments at L1 level.

With regard to L2 or FL learning, however, the position is more complicated. Klein-Braley's use of a sample of German and Finnish speakers means that the hypothesis was tested with subjects whose first languages are marked by high degrees of internal cohesion. Both Finnish and German inflect in number, gender and case, depend heavily on agreement, and use case-governing prepositions (Hajdu 1975; Herbst, Heath and Dedering 1980), while a peculiarity of German is its use of capital letters to indicate nouns. It might be assumed therefore that speakers of both Finnish and German are likely to approach a reduced redundancy format in the same way, and hence this choice of sample is more likely to produce a high correlation than, say, the choice of Greek and Turkish speaking children that was used to test Hypothesis 2.

At this point, it is worth noticing that Dornyei and Katona's endorsements of the validity of C-testing are based on data gathered exclusively from Hungarian EFL learners. Their contributions to the literature are important because they are based on empirical research and a larger sample than that used by Cleary (1988). Even so, the use of a sample of 102 university students again raises two questions.

Firstly, to what extent are these subjects aided by the fact that they are speakers of a language which exhibits a high degree of internal cohesion? (Honti 1979). Secondly, are their C-Test scores affected by their overall level of education? The most recent positive research suggests that this may be an influential factor. Klein-Braley's (1997) comparison of C-Testing with other reduced redundancy formats is based on an experiment conducted with university level students at Duisburg, while Connelly (1997), is concerned with postgraduate students at the Asian Institute of Technology.

On Klein-Braley's own admission, she is unable to validate the fourth hypothesis, though Raatz (1984) discovered a correlation between C-test scores and a non-verbal intelligence test conducted with, again, German speaking schoolchildren. A subsequent test with a slightly more advanced group of children produced a rather higher correlation, which is further evidence in support of Hypothesis 1, but Hypothesis 4 cannot be regarded as proven on the strength of so small a sample.

Of Klein-Braley's four hypotheses, therefore, the first is validated by L1 speakers in childhood and adolescence only; the second validated by L1 speakers in childhood and L2 speakers in childhood in a German speaking host environment; the third is validated by L1 speakers in childhood and well educated FL learners from highly inflectional, predominantly Finno-Ugric linguistic backgrounds; and the fourth remains unsupported.

The only feature common to the four hypotheses is that three have been validated with L1 speakers in childhood, suggesting that C-tests are almost certainly an excellent method of assessing the competence of German speaking children studying German in an L1 environment, and that their validity may well be extended to other children studying their L1 in natural environments; to adult and adolescent learners of L1 in similar situations, and possibly also to L2 learners of German living in a host environment.

The Hypotheses Extended

This is a long way from suggesting, as do Strawn, Cleary and North, that the same techniques can be applied to FL situations in Korea, Oman and Britain, and these claims are robustly refuted by Jafapur (1995). By conducting the largest experiment to date (202 English native speakers at university in America, and 325 English majors at university in Iran), Jafapur first proved that native speakers "did not reach the 'perfect performance criterion' envisaged by Klein-Braley and Raatz" (P.199). Native speaker scores were high, but only on one test out of twenty was a perfect score achieved. Jafapur asserts that a change to the deletion ratio, or to the deletion start, obviously produces C-Tests of differing difficulty, and hence the claim that the sampling is representative (Klein-Braley and Raatz 1984;136) is disproved.

His research also proved what I had long suspected in the light of the Omani servicemen's reactions to Cleary (1986). 64% of his non-native speakers reported negatively on the C-Test, many claiming that it appeared to be more a puzzle than a test of linguistic ability. This is an interesting comment and one which partly explains behaviour that I have noticed with my own students. It is not uncommon for students to omit one or more lines of a C-Test, and then complete individual items almost at random. The students have realised that they will receive credit for those items which are correctly completed but any real attempt to construct a coherent prose passage has been abandoned in favour of an approach which owes more to guesswork and problem-solving.

The "puzzle" criticism is also significant in the light of Strawn's adaptation of the C-Test for use in Korea. He took a sample passage from Modern Freshman English II, published by Yonsei University English Department, to demonstrate how the C-Test avoided the disadvantages inherent in the random N-th word deletion found in conventional cloze, and while in this instance he left blanks of equal length in the mutilated words, he also suggests that one way of simplifying the test would be to leave blanks indicating the number of letters that have been deleted from each word.

There are three problems here. Firstly, the use of a specifically written in-house test violates Klein-Braley's insistence that C-Tests use authentic materials as the basis of text production. Secondly, the passage appears to be taken from an EAP text, while Klein-Braley declares that "Examinees with special knowledge should not be favoured by specific texts" (1997;64). Thirdly, the use of a single passage also violates the concept of using several, different short texts. It is therefore open to question whether Strawn is actually using a C-Test at all.

Another difficulty here is that examination of the data on Hypothesis 4 leads Klein-Braley to the conclusion that the subjects who perform best on C-Tests are those who can appreciate the syntactic cohesion of the text and simultaneously understand the semantic relationships at work within the syntax. This is a far more sophisticated thought process than that envisaged by Strawn, for where Klein-Braley's "good" subjects can "chunk" the text to complete the C-Test, Strawn reduces the exercise to a guessing game; focusing attention on individual-words, and justifying the criticisms of Jafapur's subjects.

North (1991) working with EFL students at Eurocentre Bournemouth, also made use of non-authentic texts and indicated each missing letter with a dash. He went further than Strawn, however, by "taking care not to split graphemes like "ch", "th" etc" (P.174) and he also admits that the texts used were often frequently edited "to influence where deletions would occur, and to ensure that the text came to a reasonable end after the last deletion" (P. 174). He was concerned that a continuation would result in redundancy in the text, and this, in turn, would give the students clues and increase their use of reading strategies. He does not, however, seem to notice that his adaptation of the rubric by retaining graphemes effectively restores every instance of the use of the definite article - "th" can only be "the" or "thy".

Redundancy in the text, refusal to mutilate graphemes and the indication of missing letters are not concerns mentioned by Cleary (1988), but he also cites a passage which was purpose written "so as to sample the course of study that the subjects had followed at SOAF language schools" (P.28) and which "resulted in a passage which was undoubtedly artificial and not altogether coherent above paragraph level" (P.28). This again violates Klein-Braley's ruling on the use of authentic material, and substitutes a series of connected paragraphs for different texts. What is more important, however, is Cleary's claim that these tests "could not be 'taught to'" (P.27). Jafapur having proved that my intuitions on face validity had been correct, I decided to construct an experiment to determine whether increased practice would increase subject's scores.

A Preliminary Experiment

Following Strawn and Cleary, a C-Test was purposely written. The text was based on a conventional cloze passage which had been extensively trialled with SAF Level 2 students (Appendix 1). It was introduced with a head-and-shoulders photograph to focus the subjects' attention and was 71 words long, yielding 32 items. The average number of words per sentence was 7.1.

A major departure, however, was my decision to avoid the rival claims of "conventional" C-Tests (rule-of-two deletion with the first half of the mutilated word visible) as against the approach which indicates the number of missing letters. I chose instead to delete all but the first letter of each test item, thus increasing the complexity of the test overall and reducing the difference between individual items.

The subjects consisted of 46 young soldiers from the Force Ordnance Service. They were chosen because of their similarity to Cleary's initial sample "SOAF students are all in their late teens or early twenties. They share the same professional (i.e. military), educational, cultural and language background and they have all followed, or are following, the same course of English. Therefore, errors tend to be common to the group rather than idiosyncratic" (P.27). The FOS students, however, were more cohesive than the SOAF sample, as many had undergone basic training together, and they had lived and worked together in the same military establishment.

The control group of 25 students consisted of one Level 1 class of FOS personnel, who had been selected according to their service numbers. On arrival in the unit, these men had been given an English Literacy assessment, and had scored between 65% and 90%. Anyone who failed to achieve the 65% mark was allocated to a Literacy Level class, while those who scored above 90% were given further assessments by the official RAFO examination team, and were generally found to be suitable for immediate enrollment at a higher level than Level 1.

The control group was administered the C-Test, on the day after their final examination at SAF Level 1 (Day 64 - 2/4/96). They had had no prior exposure to C-Tests, but no time limit was imposed. Once the C-Test had been completed, the scripts were collected and filed. They were not, however, marked immediately. The results of the control group are listed in Table 1.

Table 1

Subject	Score	Subject	Score
1	30	14	23
2	29	15	22
3	29	16	21
4	29	17	21
5	28	18	19
6	27	19	19
7	27	20	19
8	27	21	18
9	26	22	18
10	24	23	17
11	24	24	17
12	23	25	14
13	23		

Mean 22.96 Variance 38.39 Standard Deviation 6.196

The experimental group of 21 students consisted of a second class of FOS personnel selected, again, by service serial number. This class was also administered the C-Test, under examination conditions, but without a time limit, on the day after their final examination at SAF Level 1 (Day 64 - 13/8/96).

Prior to that date they had been exposed to five other C-Tests. These C-Tests had been presented on five successive days in the week immediately prior to the examinations (days 58-62) of the course, and had been completed as class exercises. The students had worked on the C-Tests independently, and then each C-Test had been revised and completed with the aid of an OHP transparency. No mention had been made of the coming experiment, however, and the subjects had been informed that the C-Test was a new type of exercise that they would encounter at SAF Level 3.

To ensure standardisation of marking, the experimental group's scripts were collected and marked at the same time as those of the control group. I marked both sets of papers in the same afternoon, and the results of the experimental group are listed in Table 2.

Table 2

Subject	Score	Subject	Score
1	31	12	22
2	30	13	21
3	27	14	21
4	27	15	20
5	27	16	20
6	27	17	20
7	27	18	19
8	25	19	18
9	25	20	15
10	25	21	15
11	24		

Mean 23.14 Variance 20.51 Standard Deviation 4.52

Nunan (1992) indicates that different populations can only be established once the population mean has been established, and these figures are shown in Table 3.

Table 3

Group	Mean	Standard Deviation	Standard Error	Population Mean
Control	22.96	6.196	1.239	20.486-25.618
Experimental	23.14	4.25	0.986	21.148-25.092

The figures are close, but the difference has been enough to encourage me to design a further experiment to determine the extent of exposure to C-Tests at which the differences might become statistically significant.

A Design for Further Research

Since the conclusion of the experiment with the Force Ordnance Service personnel, I have been posted to the Armour School at MSO Shaafa. This posting has meant that I will be unable to conduct any further research with FOS personnel, but it has given me access to a cadre which is more homogeneous.

Owing to the remote location of the Armoured Brigade, MSO personnel are given basic training at Shaafa, and are then allocated to units with the Armoured Brigade according to manning requirements. I intend, therefore, to select a control group from the personnel presently under training, cross referencing age and civilian education experience to ensure that they are as similar in background as possible. This will give me a control group which is less subject to personal variations than was the case at FOS, where classes were arranged simply on the basis of the men's service numbers.

The control group will be enrolled at the Armour School for a SAF Level I English course, and in the same way as at FOS School, they will be administered the C-Test, under examination conditions, but without a time limit, on the day after their final SAF Level I examinations. Following that, I will conduct a t-test to allow comparison between the FOS control group scores and those of the MSO personnel. I do not anticipate that the t-test will reveal a significant difference, but it is possible that the more careful selection of the MSO cadre will be reflected in a smaller deviation from the mean.

Following this experiment, I will select the MSO experimental group according to the same criteria, attempting to match age and civilian education differentials as closely as possible to those of the personnel from the control group. These men will again follow a SAF Level I English course, and will receive the C-Test on the day after their final examinations, but prior to that they will (a) receive 5 practice C-Tests, and b) they will be informed that a similar test is to be administered at the end of the course.

This change in criteria is intended to test Cleary's assertion that the C-Test "can not be taught to". I believe that while the FOS experiment has failed to prove that casual exposure to C-Tests is, alone, likely to improve test scores, Hughes (1992) and Yeo (1994) both suggest that "positive backwash" from assessment may effect an improvement.

In the MSO experiment, therefore, the students will be allowed to work through each C-Test independently, after which the texts will be revised as a class activity using an OHP. The students' attention will be directed towards cohesive devices, to collocations and to reading strategies that will enable them to decode the text (as is already done with SAF Level 3 students). Little and Singleton (1990;14) mention that "in filling C-Test slots our subjects tended to give priority to a ready lexical solution over morpho-syntactic and more general semantic issues", but with the MSO personnel every effort will be made to encourage them to "chunk" the text. Following the administration and marking of the final C-Test, the results will be analysed and subjected to a t-test to determine whether there is a significant difference between the control and the experimental groups. I will conduct a second t-test to determine whether there is a difference between the FOS and MSO experimental groups, in the expectation that the MSO experimental group will reveal test scores that are significantly higher than both the MSO control group and the FOS experimental group. This difference can then be checked and confirmed using an analysis of variance (ANOVA).

Implications of the Research

If the results of the experiment confirm what I anticipate, then Cleary's contention that C-Tests can not be taught to will be revealed as an opinion that was never tested empirically. The results will also prove that the deliberate exposure to C-Testing which occurs during preparation for the SAF Level 3 examinations is almost certain to lead to an improvement in the candidates' scores, and that C-Tests are not neutral instruments.

At the present time, in a period of 13 weeks, SAF Level 3 students are likely to encounter over 20 C-Tests in their course core materials alone. There are further examples in the standardised progress tests, and these may be supplemented with further C-Tests devised by individual teachers. The very existence of all this material suggests that, in Oman, teachers who are working with C-Tests have intuitively disbelieved Cleary's assertion, and that they have provided extensive practice in a bid to strengthen their students' chances in the final examination.

If my research produces the results that I anticipate, then the judgement of these classroom teachers will be vindicated, and the effort that they and their students have expended in preparing and completing C-Tests will be seen to have paid short-term, instrumental dividends.

At the same time, however, the claims made for the reliability of the C-Test will be proven to be false, and it will be a moot question whether the C-Test, in any form, should remain a part of the SAF assessment procedures.

If my research is inconclusive, on the other hand, it may well be that additional practice in C-Tests has little effect on student test scores. In that case, Cleary's claim that C-Tests are a reliable measure of ability will be justified, and teachers might do well to spend less time on C-Tests as exam practice, and concentrate instead on more meaningful task-based skills (Willis 1996) that will give students a greater overall competence.

Bearing in mind McNiff's (1995) reminder that any self-reflective research includes the necessity to accept that one could be wrong, I still believe that too much was initially claimed for the C-Test. Cohen, Segal and Weiss (1984) and Carroll (1986; 1987) all urged caution and more research before C-Testing could be widely adopted, and subsequent research has proved that their warnings were wise. Jafapur's research have definitely exposed weaknesses in two areas, and I believe that my research is likely to uncover another. Far more research is needed, with larger samples of subjects, but Cleary's early endorsement of the C-Test as a tool in Oman looks increasingly like an intuition which has yet to be proved.

References

Anderson, J., 1971. "Selecting a Suitable 'Reader"—Procedures for Teachers to Assess Language Difficulty". *RELC Journal* 2/3.

Andrewartha, J., 1987. Personal communication from English Language School (South), SOAF Salalah.

Bachman Lyle, F., 1990. *Fundamental Considerations in Language Testing* Oxford. Oxford University Press.

Bartlett, F.C., 1932. *Remembering; A Study in Experimental and Social Psychology*, Cambridge, Cambridge University Press.

Carrell, Patricia L. & Joan C. Eisterhold, 1983. "Schema theory and ESL reading pedagogy" *TESOL Quarterly* 17/4, Pp.557-73. Reprinted in Patricia Carrell.

Carroll, John B., 1986. "LT+25, and beyond? Comments" *Language Testing* 3/2 Pp.123-29.

Carroll, John B., 1987. Review of Klein-Braley C. and Raatz U. (Eds) C-Tests in der Praxis; Fremdsprachen und Hochschule *Language Testing* 4. Pp.99-106.

Cleary, Christopher, 1986. The C-Test at Lower Intermediate Level—An Appraisal SOAF.

Cleary, Christopher, 1988. "The C-Test in English; Left Hand Deletions". *RELC Journal* 19/2 Pp.26-37.

Clyne, Michael, 1984. *Language and Society in the German Speaking Countries*, Cambridge, Cambridge University Press.

Cohen, A.D.; Segal M. & Weiss Bar-Siman-Tov R., 1984. "The C-Test in Hebrew" *Language Testing* 1 Pp.221-25.

Connelly, Michael, 1997. "Using C-Tests in English with Post Graduate Students." *English for Specific Purposes* 16/2 Pp.139-150.

Devine, Joanne & David Eskey (Eds) 1998. *Interactive Approaches to Second Language Reading* Cambridge: Cambridge University Press Pp.73-92.

Dornyei, Zotlan & Katona Lucy, 1992. "Validation of the C-Test among Hungarian EFL learners" *Language Testing* 9/2 Pp.187-206.

Dornyei, Zoltan & Katona Lucy, 1993. "The C-Test; A Teacher-Friendly Way to Test Language Proficiency" *English Teaching Forum* 31/1 Pp. 34-36.

Dulay, Heidi; Burt Marina & Krashen Stephen, 1982. *Language Two*, Oxford, Oxford University Press.

Fillmore, Charles J., 1982. "Towards a descriptive framework for spatial deixis". In R.J. Jarvella and W. Klein (Eds) *Speech, Place and Action*, London, Charles Wisley Pp.31-59.

Hajdu, Peter, 1985. *The Finno-Ugrian Languages and Peoples*, London, Andre Deutsch Herbst Thomas; Heath David and Dedering.

Honti, L.,1979. "Characteristic Features of Ugric Languages (Observations on the Question of Ugric Unity" *Acta Linguistica Academia Scientorum Hungaricae XXIX/1-2*, Budapest, Akademiai Kiado, Pp.1-26

Hughes, A., 1992 . *Testing for Language Teachers*, Cambridge, Cambridge University Press

Jafapur, Abdoljavad, 1995. "Is C-Testing Superior to Cloze?" *Language Testing* 12/2 Pp.194-216

Kebir, C., 1994. "An action research look at the communication strategies of adult learners" *TESOL Journal* 4/1 Pp. 29-31.

Klein-Braley, Christine, 1984. "Advance prediction of Difficulty with C-Tests". In T. Culhane; C. Klein-Braley and D.K. Stevenson (Eds).

Klein-Braley, Christine, 1985. "A Close-up on the C-Test; A Study in the Construct Validation of Authentic Tests" *Language Testing* 2/1 Pp.76-104.

Klein-Braley, Christine, 1997. "C-Test in the context of reduced redundancy testing; an appraisal." *Language Testing* 14/1. Pp. 47-84.

Klein-Braley, Christine & Raatz Ulrich, 1984. "A Survey of Research on the C-Test" *Language Testing* 1 Pp.134-46.

Kral, Thomas (Ed) 1994. *Teacher Development; Making the Right Moves* Washington. D.C. USIS Pp. 270-73.

Lakoff, G. , 1987. *Women, Fire and Dangerous Things*, Chicago, University of Chicago Press.

Little, D. & Singleton D., 1990. "The C-Test as an elicitation instrument in second language research". Paper presented at AILA '90, Thessaloniki, Greece. 16-22 April.

Martin, Hans,1980. *Grimm's Grandchildren; Current Topics in German Linguistics*, Harlow, Longman.

McNiff, Jean, 1995. "How can I develop a theory of critical self-reflection?" *Studies in Continuing Education* 17 1/2 Pp.86-96.

Meyer, Meinert, 1991. "Developing Transcultural Competence; Case Studies of Advanced Foreign Language Learners". In Dieter Buttjes and Michael Byram (Eds) 1991. *Mediating Languages and Cultures*, Clevedon, Avon Multilingual Matters.

North, Brian, 1991. "Standardisation of Continuous Assessment Grades." In Charles Alderson and Brian North (Eds) 1991. *Language Testing in the 1990's*. London Modern English Publications in Association with the British Council Pp.67-77.

North, Brian; Hirst Laura; Petty Chris & Scott Roger, 1988. "Testing Time for Students" *EFL Gazette* 101 Pp.6-7.

Nunan, David, 1992. *Research Methods in Language Learning*, Cambridge, Cambridge University Press.

Raatz, Ulrich & Klein-Braley Christine, 1982. "The C-Test: A modification of the cloze procedure". In T. Culhane, C. Klein-Braley and D.K. Stevenson (Eds).

Raatz, Ulrich, 1984. "The Factorial Validity of C-Tests" In T. Culhane, C. Klein-Braley and D.C. Stevenson (Eds) *Practice and Problems in Language Testing 7*. Colchester, University of Essex.

Robin, G.,1985. *Crosscultural Understanding*, Oxford, Pergamon Press.

Rumelhart, D.E. , 1977. "Understanding and summarising brief stories". In D. LaBerge and S.J. Samuels (Eds) 1977. *Basic Processes in Reading; Perception and Comprehension*, Hillsdale N.J. Lawrence Erlbaum Pp.265-303.

Rumelhart, D.E., 1980. "Schemata; the building blocks of cognition". In R.J. Spiro; B.C. Bruce & W.E. Brewer (Eds) 1980. *Theoretical Issues in Reading Comprehension* Hillsdale N.J. Lawrence Erlbaum Pp.33-58.

Rumelhart, D.E. & Ortony A., 1977. "The representation of knowledge in memory". In R.G. Anderson; R.J. Spiro and W.E. Montague (Eds). *Schooling and the Acquisition of Knowledge* Hillsdale N.J. Lawrence Erlbaum, Pp.99-135.

Shank, R.C., 1982. *Dynamic Memory*, Cambridge: Cambridge University Press.

Shank, R.C. & Abelson R.P., 1977. *Scripts, Plans, Goals and Understanding* Hillsdale N.J. Lawrence Erlbaum.

Strawn, Dwight J., 1985. "The C-Test; Another Choice". AETK News 4/4. Reprinted in *English Teaching Forum XXVII/1* 1989, P.54.

Taylor, W.L., 1953. "Cloze procedure—a new way for measuring reliability" *Journalism Quarterly 30* Pp. 414-38.

Van Brederode, Tom, 1996. Collocation Restrictions, Frames and Metaphor Unpublished Doctoral Dissertation University of Amsterdam.

Willis, Jane, 1996. *A Framework for Task-Based Learning* Harlow Addison Wesley Longman.

Yeo, Serena, 1994. "Working for positive backwash—a laudable aim?". In Wendy Scott and Susanne Mulhaus (Eds) 1994 *Languages for Specific Purposes*. London CILT in Association with Kingston University School of Languages.

Zeichner, Moshe, 1988. "Classroom Testing; The Examinee's Perspective" *Studies in Educational Evaluation 14/2* Pp.215-233.

Why Teacher, That's Your Job

PHIL COZENS
Sharjah Mens College

The paper discusses the reaction of students, and their teachers, after they were asked to prepare the reading component of their second progress test. The rationale behind student-designed tests and their possible motivational effect on students is also touched on.

Introduction

This paper is on a project that covers several student created tests which were written specifically for a progress test taken by four Level 2 (Intermediate Level) classes at the Ras Al Khaimah Women's College towards the end of their second semester. Each class was involved in the creation of two tests, although the classes were informed before starting that only one would be chosen for the Progress Test, provided it fulfilled certain criteria. This was, in some ways, an ambitious extension of previous experiments with student created tests, but also partially the result of two presentations attended at the 1998 TESOL-Arabia conference. In the first (Hubley, Coombe & Stuart) Hubley stated that her best testing experience was whilst she was studying Mediaeval History where the professor on the course had told students that they should prepare questions, both individually and in groups and that he would select the best questions presented and actually use them for the examination. This had motivated students to work together to master materials in order to write good questions which, they hoped, would be selected. The second was a presentation on Learner-Centred Testing for ELT by Coombe and Kinney in which they asked how much student involvement teachers permitted within their classrooms.

In some ways the project was more about empowering students, as Hubley suggested, than the actual assessment process itself. It was a way of promoting awareness of not only what is required of the student, but also of what teachers and examiners are seeking. As the project was cross-curricular, the aims were twofold, i.e. from the English teachers' perspective students were being asked to develop greater understanding of the way to approach texts in a 'test' situation, whilst the Computer teachers wanted to make them more aware of the importance of information contained in the format, in that students were being asked to become more aware of such non-textual clues as individual fonts and layout.

After the TESOL conference the Director of Ras Al Khaimah Women's College was approached and asked whether the Level 2 students could write one component of their second progress test. Whilst this was an important part of students' semester grade, it was not, as described by Coombe, a 'high stakes' examination; the progress tests counted as only 5% of the final grade with each skill being evenly weighted. With his approval, the English supervisor was then consulted. He also agreed to the suggestion, provided the other English teachers did not object. After a short meeting with the other teachers, it was decided that each of the four classes would be visited to gauge their response. After further discussion, it was decided, in order to have some form of test security, that each class would produce questions for two passages and that only one would be used. Students were also informed that if another class became aware of the contents of their tests, neither would be used. In an attempt to make the exercise cross-curricular, it was also decided to approach the Critical Thinking Business teachers for each class and ask them to provide one passage which they wanted their class to read. These were then checked to ensure that they fell within the specifications of the reading skills examinations.

The Procedure

Shortly after the midway point of the second semester each class was visited and the objectives explained. With the exception of one slightly less enthusiastic class, the students were quite excited by the idea and each class provided one or two subjects on which they would like to have a text. These included, the environment, language and business. Suitable texts were found and distributed to the teachers of the different classes. Prior to the students actually writing the questions, several reading tests and comprehension passages were examined and different types of question identified. These were then examined in greater depth and some guidelines drawn up by the students themselves. With one particular class, the questions of what an achievement test was trying to do were discussed, and why, if the whole class scored 100%, it could not be counted as a good test. This concept, was for some students, very difficult to accept.

Before the actual passages were handed out, teachers and students in all classes decided, as a group, on an outline for the work. First of all, the classes, in groups, were to prepare questions of a particular type. This was to be followed by class piloting of the questions and finally, once the questions were decided on, the rubrics were to be written. Several possible test-writing problems were also identified and highlighted: obvious answers, duplicated questions and those which provided the answers to earlier questions.

With one particular class, after being given the first text, in this case about the environment, the different groups were each asked to produce two different types of question. Allowing the students to work in small groups prevented any unnecessary stress or loss of 'individual face' whilst, at the same time, permitted them to accept the position as 'knower' (Ellis 1994). At the end of this stage, members of these groups then presented their questions to the other class members. This was quite a difficult time, on occasions, as none of the groups were happy when their own questions were criticised, particularly if they were deemed to be too easy or factually incorrect. This did, however, lead to some heated discussion on which questions were acceptable and which were not. Questions which included the predicted problems occurred, but these were happily relegated to the bin and others produced. The concept of student as peer critic, as posited by Cozens (1997), would seem to be in place here, generating great involvement in the product without either loss of face or focusing on the process itself.

The questions were then re-examined and the concept of paraphrase initiated with some, but not all classes. All, however, now started to show an awareness of the importance of looking beyond the written word and dictionaries and thesauri were brought into play. Lexical items were examined, particularly where used in an unusual form, and synonyms discovered. This then led to a second look at the questions available and after re-examining the wording, a final choice was made. Up to this stage, approximately 4 x 45 minute periods had been spent on the passage, a time frame which appeared to be true of most classes, and the rubrics still had to be decided for the single passage so far completed. In this particular class, this took place in the computer period allocated to the English class. The different groups in the class were asked to type up their questions in a form which they felt would give other students extra information they would need. This usually took the form of applying bold, underlining or italic formatting to particular words and letters. Once this finished, each group printed out their part of the test and circulated it for further discussion. In at least one case, this led to revisions, for example the inclusion of shading to emphasise that certain areas on a table did not need answers, thus preventing other students from unnecessary stress. It was hoped that by having students spend time on the writing of rubrics for their questions and discussing how they could help students answer correctly, that this would encourage the students themselves to actually read the rubrics of subsequent tests.

After completion, students were then asked to reflect on what they had done whilst making the test, what, if anything, they had learnt; why they thought they had been asked to complete the task and if they thought the exercise was a waste of time and why. These were collected and some are attached as an appendix.

Problem Areas

Obviously, not all the classes followed exactly the same format, but each class produced at least one test. There were, of course, several inadequacies with the process. Firstly, none of the participants involved, particularly the initiator, was aware of how long the task would take. Secondly, despite the problem of test security, asking students to produce two tests did not take the students' attitudes into consideration. Although, in many cases, they actually produced better questions on the second test, they were no longer interested in the process and, therefore, it would probably have been better to to have required only one. The question of the validity of the test also arises; as a learning experience for students it was valid exercise. As a way of empowering students in test-taking strategies, it was also valid, but whether it was a useful testing instrument is debatable. The concept of content validity was partially satisfied, in that each passage was of a similar length and reading difficulty. The test questions themselves, while being student-generated, were still checked for reliability by both the students involved as well as their teachers, although they did not, themselves, overtly interfere in the process. Face validity is partially there, in that the students were made aware of the effort which was required to make the test, some were, however, a little bemused by the fact that they were making a test for themselves. Creating tests for each other was completely acceptable, but for themselves seemed inappropriate to some.

Another basic testing problem is the practicality of the test. In order to ensure that each class takes both a test that they have written and one from all the others, the time allocated for the test had to be extended by approximately 20 minutes. While this was not a great problem, it did, in some ways, cause some unnecessary strain for students in that they were under test conditions longer than was normal and it also necessitated the use of time allocated to other subject areas.

Feedback

Other teachers involved in the process were also asked to comment on what, if anything, had been achieved. The majority felt that as an exercise for assisting students to exploit a text more fully, it was successful. It also appeared to have succeeded in its attempt to make students more aware of ways to approach questions and what to look for in the instructions, as can be seen from the comments of several teachers below.

> Feedback from XXX2 was very positive - at least for the first passage. As ever, they enjoyed the sub-group approach to the activity, and were spurred on by the notion that their questions were to be used on other classes. This inspired them to devise more challenging questions.

As a tool to encourage ,deeper understanding of a text, it certainly worked well. Students pored over dictionaries to puzzle out synonyms and other word forms to rephrase the original text for True/False tasks. In another section, paragraph headings/matching was an excellent tool for summarising. A third group devised an interesting cause/effect matching task, which encouraged more thought on the relationships and structures involved. The short answer section pushed the fourth group to focus on the specific information needed, rather than accepting guesswork and mere lifting from the text.

Later, when sub-groups exchanged questions and tried to answer them, this prompted valuable class discussion of the suitability of questions and correctness of answers.

All in all, this proved a stimulating exercise which generated useful thinking about approaches to understanding the requirements of reading questions.

Now, as regards the logistics of using these as a progress test, this might prove more contentious

Students found a sense of maturity with having to think like the teacher, and a sense of cynicism too though, I think, reading between the lines...

They were aware too of the pitfalls of question design and happy to exploit them, can't blame them.

Students felt rewarded by being offered the insight the exercise gave them.

Inevitably, some students appreciated the subtitles more than others.

Students took it seriously.

These comments clearly indicate that whilst there were some reservations on the tests being used an a testing instrument, there was clear support on both the motivational and learning aspects involved in the process.

Addendum

As posited earlier, the actual use of the tests did not prove to be as useful as originally intended, but, despite the time expended, the exercise appeared to have fulfilled some of its aims, it was felt worth repeating in a less ambitious form.

After the tests were administered and the results analysed, it was found, not surprisingly, that most students had fared better on their own tests, although no-one was able to achieve 100 per cent on any test. One test, in particular, appeared to be easier than the others, and students who attempted this achieved better results overall. For this reason, and in order to standardise the grading, students were only marked on those questions attempted and an overall percentage grade calculated. A second grade was calculated by taking each student's average over their two best scores.

Conclusion

The concept of students creating tests for each other has long formed a part of the classroom teacher's arsenal, particularly because of its apparent ability to motivate students. Expecting students at this level to produce their own examination was, perhaps, slightly ambitious, but as a way of empowering students in test taking strategies it did appear to have valid outcomes. More importantly, however, is the fact that UAE students were prepared to spend an extended period working on a particular text, whilst, at the same time, accepting criticism, although not always willingly, of their work. These students also showed that they could work together, were able take the initiative to consult other reference sources in order to produce what they felt were both valid and reliable examinations and proved that given the responsiblity were able to maintain test security. For these reasons, as stated previously, it is felt that such experiments are worth repeating.

Appendix 1

Student Comments

Student answers to the following reflective questions.

1. What were you thinking when you tried to make the questions?

2. Why do you think that the teachers asked you to do this?

3. What, if anything, have you learned from doing this?

4. If you think it was a waste of time explain why and if not, say why not.

1. Because we took this article before, so we tried to find some specific and difficult questions.

2. To guess the type of questions that they include in final exams and to see our abilities in forming exams questions.

3. Read carefully in order to form good questions.

4. We can use this time in something more important and useful like exams to collect marks or do something to improve our skills so that it will help us in our final and what we are doing now has nothing to do with our final.

I look at the short clear answer which is easy to find.

Because the teachers want to show that they are very kind to students.

It just help me to read the essay very clearly and that will not help me in the test because we do not have enough time to read in the test.

It was not wasting time because really we want more reading. It also help us notice which quetions that may ask us in the test. So that will not reduce our time in the test.

Note our real problem is the time of the test, not the test itself.

To make clear different questions to be not so easy or so difficult to answer.

To be easy for students to expect the type of questions on exams and give them experience in making questions.

I learnt the type of different questions and to read the questions carefully before answering them.

Good way to collect marks and be ready for the final exam and to have some idea about making questions.

Looking for clear information, or statistical / short answers.

To learn which types of information we should look for and to look where we can look to find the answers.

I learn it is not important to understand all the question, but it is important to find the key words.

It is not a waste of time, if we trained on this before and if we really try to find the answers for all the questions which students make we learn a lot.

I'm looking for clear answers, for example vocab, names and percentages.

In my opinion it will help me to know and learn how to put questions and improve my reading and we want to show the teachers the kind of questions that we could answer in a short time.

How to do some questions.

It's not wasting time because it help us learn how to improve our reading and also I learn new vocab. I think it is helping me in my reading speed too.

I'm looking for clear and short answers . Also for basic things like names, why it's important etc.

I think the reasons are that they want us to improve our reading and how we could make good questions.

I learn to read everything in the task, know eveything in the exam.

It is not a waste of time because we improve our reading and learn how to do good questions which is improving help our English in general like how we learn new vocab.

I looked for the vocab.

To share each others' ideas in group work and improve our reading skills

New vocab.

It's not a waste of time because we work and practise reading.

I don't know

I don't know

New vocabulary

Because it might not come in the test.

I looked for new words.

maybe it's one of a million ways to pass the English test.

Nothing

It's a waste of time, but we could writting the meanings of difficult words and practising them.

I tried to learn the new meanings.

because, as you said, it helps us to copewith tests.

I learnt to trick students and make them miserable.

It's a waste of time because you forced us to do this and we could take more reading articles instead.

I don't really think that it is a waste of time, because when I know how instructors make questions, it will be fairly easy for me to answer. The only thing that I could think of thast could improve is by writing other questions first.

I genuinely believe that I learnt a lot of things that I can't mention it now.

Example Text and Questions

This Land is Your Land

After years of industrial pollution, the world's big businesses are now banding together to clean up their act, says Ursula North

A. The hippies of the 1960s coined the name Mother Earth for our planet and praised her powers as a nurturer of all life. Three decades later, Mother Earth has grown tired and weak, struggling to support her human charges.

B. At the heart of this struggle is our poor management of agricultural land, which is destroying the very basis of human productivity - the soil. Soil takes thousands or millions of years to form - and a year or two to destroy. More than 25,000 million tonnes of material are removed from farmland every year, not counting the soil that is inevitably lost to natural erosion.

C. In the United States, 44 per cent of cropland is affected by soil erosion; in El Salvador, 77 per cent of all land is eroded; and 38 per cent of Nepal's fields have had to be abandoned as a result of land degradation. A recent global assessment of land degradation estimates that 15 per cent of the world's land has now been degraded by human activities such as overgrazing, deforestation and over-exploitation. The worst affected areas are the drylands, which cover 47 per cent of the world's land area. Here land degradation caused by human action is called desertification, a term used not to describe the spreading of deserts but the creation of them. In the drylands - where desertification affects 47 per cent of rain-fed cropland - the lives of hundreds of millions of small-scale farmers have been ruined.

D. Recurrent droughts exacerbate the problem.The African drought of 1984 to 1985 affected more than 30 million people in 21 countries. Of these, 10 million were forced to leave their homes to become known as environmental refugees. Death, disease, malnutrition and disability are haunting the lives of these people as they continue to suffer intolerable living conditions.

E. We are also guilty of mismanaging the planet's forests and woodlands. Forests provide fuel, building materials, foods, fodder, medicines, fibres and employment for millions of the world's poorest people. But many are losing their access to forests as a result of deforestation, currently estimated at 16.8 million hectares a year. Deforestation, too, produces its own environmental refugees. Millions of people have been forced to leave their homes in Central America, the Caribbean, Africa and Asia.

F. Lack of fuel wood, on which some 2,000 million people depend, is one of the crucial issues. Currently 11,300 million people are consuming firewood faster than it grows locally. Fuel wood is scarce in most developing countries and it becomes scarcer as every year passes. The situation ought not to be as severe as it is. The world's live stockpile of wood amounts to 315 billion cubic meters, which generates a growth of six billion cubic meters a year, twice as much as the world needs each year. But much of this forest growth occurs in thinly populated areas of Alaska, Canada and Siberia, leaving other areas, notably South-East Asia and Latin America, perilously short of timber.

G. Logging is not the only cause of deforestation. Clearing land for agriculture, forest fires, air pollution, and slash and burn agriculture - which allows no time for tree regeneration - all take their toll. The effects of deforestation are wide ranging. Forests are home to many peoples and many species, so when forests disappear, so do their inhabitants. Forests prevent soil erosion and provide one of nature's principal means of water management. When trees are removed, torrents of water are allowed to run unchecked down steep hillsides, causing avalanches and flooding. For instance, when the Himalayas were covered in trees, Bangladesh suffered a major flood only about twice a century. Now one flood every four years is the average. Trees also play an important role in stabilising climate.Deforestation is responsible for one-quarter to one-third of the atmospheric carbon dioxide which now threatens to produce global warming.

H. The worldwide response to land degradation has been slow and painful. Restoring degraded land is expensive, and prevention is the best form of cure. Even so, there have been isolated examples of progress. In Pakistan, 32 projects have succeeded in reducing salinisation of irrigated land to 28 per cent from 40 per cent, and in Hungary and Bulgaria millions of hectares of land are being either protected from erosion or restored. As for combating deforestation, less than five per cent of the world's forests are formally protected, but some countries have made major progress. Brazil has launched an ambitious system of forest parks and conservation areas covering 15 million hectares.Costa Rica has protected 80 per cent of its remaining wild lands. The Ivory Coast has banned log exports. Bolivia has declared a five-year moratorium on logging concessions, and several countries are developing non-destructive uses for forests.

I. Improved wood stoves have helped reduce consumption of fuel wood, notably in Nepal and Central America. On a global scale, the Tropical Forestry Action Plan was launched by four international agencies in 1985 to encourage sustainable forest development. So far, 81 countries have joined the scheme, but many nations have yet to add their signatures to it. Unless they do, Mother Earth will one day have nothing left to give.

Source: *Khaleej Times* 886 words

This Land is Your Land

Section A

True, False, No information

1. In the most developed countries the consumption of fuel wood is rare.

2. Soil takes a million years to form.

3. International agencies aimed to develop the tropical sustainable forest.

4. 30 years ago the land sustained the degradation of human activities.

5. In Africa, people suffered from death, disease and malnutrition because of two years' drought.

Section B

Find the correct cause for each of the problems and then write the corresponding letters from A - G for questions 1 - 5.

1	15% of the world's land has now been degraded by human activities such as ..	A	forced people to leave their home town.
2	Desertification is leading to ...	B	flooding and avalanches.
3	The environment problems ...	C	reduction in the consumption of fuel wood.
4	Bad farming methods cause ...	D	overgrazing, over-exploitation and deforestation.
5	Improved wood stoves caused ..	E	degradation of plants.
		F	clear land, no life and air pollution.
		G	death, disease and disability.

Section C: Answer the questions In brief

1. Give three examples of how people have caused degradation of Mother Earth. (3 words)

2. Why have people migrated from their own countries? (1 word)

3. What is the importance of woodland? (4 words)

4. Where has the export of timber been forbidden?

5. Why has our Mother Earth become destroyed? (3 words)

6. What is the proposal that appeared to help Mother Earth from destruction? (3 words)

Section D: Choose the most suitable heading for each of paragraphs A - I

1. Taking away the basic needs by the human's avidness.
2. Preventing trees from playing their important role in nature.
3. Increasing in soil erosion by human activities.
4. Projects have been launched in different countries.
5. The exacerbation of the problem in Africa.
6. More plans to save Mother Earth.
7. From the power to the weakness.
8. Human overspend wood than his needs.
9. The dramatic degradation of the soil.
10. How to help the earth by technology.
11. International attention to the problem.

References

Coombe, C. and Kinney J. Learner-Centred Testing for ELT. TESOL Arabia '98 25-27 Mar: Al Ain: 1998.

Cozens, P. Student Created Tests as Motivation for Learning. CTELT 97 11 Jun, Al Ain. 1997.

Ellis, R. *The Study of Second Language Acquisition.* Oxford: Oxford University Press, 1994.

Hubley, N, Coombe C. and Stuart R. Empowering Students in Test Taking Strategies: *TESOL Arabia* 25-27 Mar, Al Ain: 1998.

North, U. This Land is Your Land *Khaleej Times.*

Computer Generated Visuals and Reading Texts

CHRISTINE M. CANNING,
LISA BARLOW AND CECILIA KAWAR
United Arab Emirates University

1.0 Abstract

This paper will examine the role of computer generated visuals that compliment reading texts used in testing situations. Issues such as vision and visual processing; visuals and their relationship to language learning; visuals and testing; imagery and instruction; the effects of visuals on vision and reading; research in F/SL reading; background information on reading texts used in current CGE readings; and the scope and limitation of computer generated readings will be discussed. Finally, information on how to create computer generated visuals for use on language tests will be disseminated.

2.0 Vision and Visual Processing

David Sless (1981) coined the term the "thinking eye" because he believed that "vision is the instigator of thought, not its handmaiden." He is correct in both metaphorical and biological terms. The eye is a complex part of the body. Neutral tissues are developed in order to make use of incoming visual information. The evolutionary catalyst for the development of the brain was the need to process visual information. As Sless states, "Vision is the seat for intellect". In medical terms, the eyes are the first to appear on an embryo and the brain being a subsequent outgrowth. In structural terms the "eyes have not grown out of the brain, but the brain has receded from the eyes" (Polyak, 1968). The eye is not biologically separate from the brain. It is actually part of the organ. In other words, the brain is a part of the eye. Kant (1981) reported that " vision and thinking are one process; they can't be separated, either logically or physiologically".

2.1 Visuals and Language Learning

The use of visual prompts can serve as an aid to language learning, but does a visual aid or distract in a testing environment? It can be argued that visuals can cue responses by limiting or expanding a point of reference. It has been documented that mental images are the primary symbols of thinking (Berkley 1710). Recent publications have suggested that graphic ideation helps develop ideas worth communicating because images can serve as an interactive stimuli and that words may serve as substitutes for images (Canning, 1998). Presently, there are mixed reactions in the field to the effect of visuals in second language assessment practices.

Researchers have found that visuals tend to isolate a word from a particular context. A visual can remove the grammatical and semantic information, therefore, allowing the F/SL learner to associate other words in various contexts with the visual prompt. The learner is able in turn to process the input, associated words and visual images and rebuild connections between structures and meanings in the target language. On the other hand, recent research suggests that "subjects divert their attention away from the visual task to the imagery" (Lemly and Reeves, 1992). Therefore, it is important to examine whether or not perception deteriorates as less attention is paid to the visual task.

2.2 Visuals and Testing

Picture prompts are commonly used in standardized tests such as ACER, SPEAK Test, and The Borrgotta Test of Visual Response. Visuals present learners with the ability to predict, deduce, and infer information presented on language examinations. There is scant research in the area of visuals and testing in second language studies. The primary focus of research with regard to visuals has been completed in the field area of cognitive psychology. In language journals qualitative studies are more frequently found than quantitative studies in regards to the effect of visuals on second language learning.

The role of visuals in English Language Testing (ELT) is vital to ensure "genuine validity" (Caulfield and Smith, 1982). Caulfield and Smith coined "genuine validity" to mean a test's ability to communicate with its audience. The use of pictures in examinations can be tested not only in traditional programs and classroom settings, not through individual-competency based tests, and through large and small scale listening and aural testing. Picture items can be developed to test whether the learner understands the syntax and structure of the target language. Students can "see" an immediate meaning in terms of vocabulary recognition provided the item exists in the first language.

In ELT a picture can make a situation seem reasonably authentic. The picture can be used to test structured vocabulary, functions, situations, and skills. Visuals used as testing prompts can be employed to measure semantic and associative clusters. In addition, visuals allow students to focus on a whole or a piece and they can allow for decontextualization or contextualization of the second or foreign language.

Pictures, in testing situations, give learners options, alternatives, chances for interpretive response as well as patterns to help in answering exam questions. Moreover, visual testing prompts can aid in measuring syntactic, phonological, lexical and cultural proficiency.

In a testing environment, careful picture selection should be emphasized. Overall, pictures on tests should make a statement, be comparative, be interpretive and to the point. The picture should permit strategies for organizing knowledge. Most importantly, visuals need to serve as a bridge to enhance learning, sensory acuteness, and the testing situation as a whole.

2.3 Imagery

Do we see before we think or vice versa? If asked to count the windows in your home, do you take a mental picture of the room and start counting? Interestingly enough, research has shown that most people do including the blind. In 1975, a group of researchers (Jonas et al.) found that imagery instructions facilitated learning and recall. What made this experiment so unusual is that the participants were all legally blind since birth. It is difficult to imagine how people blind since birth were able to form pictorial images when instructed to do so. The group of researchers found that with visual images the participants scored higher and had noticablely better scores in areas of recalling information. These findings suggest that imagery instructions can affect recall and learning capabilities. Anderson and Bower (1973) observe that imagery instruction force a subject to think about a process or a word more carefully. They suggest that better recall is presumed to result from the formation of a more elaborate code.

2.4 The Effect of Visuals on Vision and Reading

Hershensonn (1980) reports that most readers hold the page 30-50 centimeters from their eyes, depending on the length of the arm, the height of the desk or table, the size of the print, the weight of the book and the amount and position of the lighting. He further describes that most print in books designed for adults measure 3 millimeters high, which makes the typical visual angle of each letter about .25 in degree. His research has shown that a four letter word occupies about 1 degree across the retina. Therefore, according to his research, if the average distance moved by a particular reader is 8 character spaces, then the average movement is 2 degrees. With this information in mind, how does the printed letter as a visual image effect the reading comprehension of the learner?

A computer generated reading text with the letters serving as a visual representation, are of a different size, color and texture of that found on a paper exam. The screen and background could also serve as potential barriers or distracters for the examinee. Coupled with the effect of a pictorial image, learners could be advantaged or disadvantaged depending on the effect of the projected or nonprojected visual stimuli.

Gibson and Levin (1975) argue that reading is usually defined as an extraction of meaning from a text. They further assert that reading is a process by which the written or printed symbols are translated into a representation in which meaning is already accessible—a translation to a form of language from which the reader makes clear that the reading process is intimately tied to other language processing, especially the ability to extract meaning from speech. However, isn't reading much more than printed speech? As Clark and Clark (1977) point out, the visual components are so different from the auditory ones and "that reading and listening, while sharing the same language, make very different demands on the information process in skills of the perceiver". Therefore, it is important to look at how subjects discriminate using visual materials in the form of texts and pictures while reading identical texts on computers as well as on paper. It is essential that any data collected be examined for adaptational effects that may or may not have interfered with visual processing as it applies to reading in testing situations.

3.1 Background Information on Reading Texts Used in Current CGE Readings

Current research on the effect of computer generated reading texts tends to suggest that reading computer generated texts is advantageous to some learners. First, these texts accommodate a diversity of learning styles: linear, top-down, and bottom-up thinking (Pruisner, 1995). For a student who reads in a linear manner, a computer text allows him to "abandon page-bound reading" and encourages him to see that reading is a continuous thought. In other words, reading while scrolling down a computer screen until the end of a text, instead of page turning, reinforces the idea that a text is one continual thought not several ideas that begin and end with each individual printed page. Additionally, these texts help students who follow a top-down mode of learning. Chains, flow charts and concept maps help students anchor abstract concepts and aid in problem solving. In addition they can test ideas against facts or solve specific contents, types and levels of users. Lastly, reading texts with computer graphics can help learners with bottom-up thinking learning styles. Pie charts, grids, and graphs help students scan, sort and organize information as they read and answer tasks.

3.2 Guidelines of Computer Generated Reading and Graphics

Although most current research on computer generated graphics utilized in a testing situation deal with listening and writing tasks, some rough ideas on the production of effective CG graphics can be postulated. First, the graphic should support the reading content: i.e., it should resemble its referent and not impose incorrect structures on the information its conveys (Williams, 1993). Second, the graphic should support the nature of the testing task. That is, perceptual interpretation of the graphic should not oppose the student's higher level cognitive process in understanding the reading text. Third, a CG reading visual should increase motivation and significantly improve the student's attitudinal level toward the exam task. Fourth, the graphic should trigger the student's imagination. It should help him develop schema, make connections and instigate an internal debate of the test questions being posed from the reading text. Fifth, the CG graphic should relate to attainable goals; that is, it should not assume a level of difficulty that is above and beyond the exam goals and objectives and the ability of the student himself. Lastly, the CG reading visual should demonstrate real-world application of knowledge (Sultan and Jones, 1995). Maps of fictitious places and charts with obviously inflated or deflated statistics should not be produced and presented in an exam situation. Attempts at such poor testing practices will only affect the face validity of the exam itself, and confuse the student in his endeavor to address the exam task correctly, coherently and successfully.

3.1 Background Information on Reading Texts Used In Current CGE Readings

Computer generated exams (CGE) were first implemented for midterm examinations in Fall 1995 at United Arab Emirates University in the University General requirements Unit for English. English Level One (EL1) was the first level to incorporate the CGE format into its curriculum. The CGE Programs for assessment were later utilized in the higher levels of English Level Two (EL2) and English Level Three (EL3). The exams were constructed to measure levels of comprehension for the English for Academic Purposes (EAP) curriculum.

Despite being an academic university program, a typical EL3 passage is written for a tenth to twelfth grade level of reading because of limited exposure to English previous to entering the university. The average EL3 passage contains 300-500 words at a 10-12th grade level. EL2 consists of an average word count of 200-500 words at a 7 -9th grade level. While an EL1 text is constructed with 100-300 words at an elementary level of between 5-7th grade. Word counts and readability measured with a commercialized computer readability software called the Fleisch-Kincaid.

Initially graphics were only utilized in the EL1 CGEs at a primitive level of chart reading or for aesthetic value instead of supporting the reading content and the nature of the testing task. Research has proven that spatial distances, color content, and size affect a learner's ability to comprehend a projected or non-projected visual. The incorrect usage of a visual has the potential to distract the learner's ability to successfully comprehend, process and respond to the testing task.

EL2 and EL3 implemented computer generated graphics in 1995. Although these graphics did not always take on the form of a physical picture, the use of colored words; time clocks; classification charts; maps; time lines and family trees were used to enhance the face validity of the computer generated exam. Whether these graphics enhanced student ability to comprehend and develop schema to successfully complete the testing task is unsubstantiated.

3.2 Scope and Limitations of Technology

Visuals have been used in different task-type questions on CGEs in the format of cloze exercises; fill-in the blank; true, false and not enough information; and multiple choice questions. Color usage is limited to 16 basic colors because of the technology available to the unit. Perhaps one of the biggest limitations in CG reading graphics is that a common language of graphics must be presented for graphics to be universally understood in CGEs (Pruisner, 1997).

4.0 Introduction to Creating Computer Generated Visuals

Computer generated visuals (CGV) can be produced by classroom practitioners, material writers and test developers to help learners master a second or foreign language (F/SL). CGVs can be constructed by practitioners who wish to create their own CGEs or Hypercard programs. A multitude of methods is available for designing appropriate CGVs for reading texts.

The role of background associations and background knowledge in language comprehension has been formalized as *schema theory* which has as one of its fundamental tenets is that a text, any text, either spoken or written, does not by itself carry meaning. Rather, according to schema theory, a text only provides directions for listeners or readers as to how they should *retrieve* or construct meaning from their own, previously acquired knowledge (Carrel and Eisterhold, 1987). Over the years research has shown that the use of colored pictures improves assessment scores of children learning their L1 because the visual serves as stimuli and activates schemata.

Living and working in a Gulf Arab environment, EFL practitioners must attempt to use graphics that stimulate background knowledge in culturally appropriate formats that will help learners to identify and to relate the CGV picture with the reading task in order to aid with the language learning process. CGVs need to be reviewed, revised, and updated on a regular basis in order to check their valid function and purpose in a regulated testing environment. Importing graphics, photos, and clip art, capturing images and saving them with the desired file extension such as a PCX file, can be helpful to a test developer.

Programs such as Windows Paintbrush, Superpaint, Arts and Letters, Arts and Letters Express, Micrografxs, Coral Draw 5, and Coral Draw 7 can serve as helpful tools to illustrate reading texts as well as target vocabulary and themes. Most of the programs allow users to choose an extension.

Classroom practitioners and test writers can search for a graphic in an encyclopedia such as Encarta or find something on the Internet and "capture" the graphic. Next, save the captured picture as a PCX file and reduce it to 16 colors to make sure it will run on a CGE program. If the captured graphic needs some "touching up" such as obliterating an unwanted part or name, the cloning tool can be used to make a similar fill that will appear as ordinary background.

Clip art can be imported from a program that allows you to change features. For example, the UGRU English Unit at UAE University uses "culturally appropriate" materials. If a figure is imported wearing short western clothing, the artist imports the clip art into PaintBrush and saves it in a PCX file. Then the image is altered and the graphic is added using long skirts and long sleeves. Another viable option would be scanning the desired graphic to the accompanying program. Again the scanned materials would need to be saved as a PCX file if they were to be used in a CGE program. Graphic artists may prefer to work with the New Windows Computer Generated Exam Authoring Program, which allows for 250 color graphic capabilities. Test writers and practitioners must consider that a strong disadvantage to this program is the amount of disc space used and BMP file time consumption. This program enables tone substances to make pictures with smoothness and without grainy effect from the old PCX 16 color files.

4.1 Process of Creating a Computer Generated Visual

Below is a list of steps for the production of CGVs for the novice:

- Choose the theme and program for the desired visual for the reading text.
- Decide if you need a schematic drawing or a life–like example. Note: if you draw a basic design or sketch with the mouse, use a basic program like Windows Paintbrush or Paintbrush Pro.

- Fill in the design with colors and save as a BMP file.

- Go to another program that converts your BMP file to PCX and click on SAVE AS. Your choices will be varied. Click on PCX.

- Find a color reduction choice and save your colors to 16 if it is necessitated. Details can be lost when CGVs are reduced to 16 colors. To 'touch up' details that are important to the questions on the worksheet or exams, you just click on the TEXT TOOL to write your labels or names.

- Go to you exam file and copy the 16 color, PCX file into your exam file. As you program type the name of the graphic, such as CATS.PCX at the end of every answer...example.

```
Q          =  What color are these cats?
C          =  black
C          =  grey
C          =  white
C          =  tan and black
A          =  tan and black
Graphic    =  CATS.PCX
```

You must repeat the command Graphic=CATS.PCX after each A into to allow the learners to refer to the graphic after each question.

The GroupHeader will give specific directions about what commands the students will see; therefore, as a graphic programmer, you must indicate how to toggle into the graphic and escape again to the answer section.

Toggling can be done in the following manner:

```
GroupHeader = {Lmagenta:Press F10 to see the picture.}
```

This command will instruct the student on how to access the accompanying graphic to use to obtain the correct answer. The L Magenta is a color that is added to the instructions to make the text more appealing to the students.

Any kind of graphic that you decide to use can be treated in this manner and used for programming.

References

Anderson, B. (1980) *The Complete Thinker*. Prentice Hall, Trenton, NJ.

Bazeli, MJ and Olle, R,. "Using Visuals to Develop Reading Vocabulary". In *Eyes of the Future: Converging Images, Ideas, and Instruction. Selected Readings from the Annual Conference of the International Visual Literacy Association*, 27th , Chicago, IL October 18-22, 1995.

Berkeley, G. (1710). "A Treastise Concerning Principles of Human Knowledge". In R. N. Hutchins (CD) *Great Books of the World*. London. Encyclopedia Britannica Inc. 1952.

Canning, C ."Visuals in English Language Testing" *TESOL Arabia News*, 5:5. June 1998 P.19.

Canning, C. "Maximizing the Effect of Visuals in Software Programs", *EMCEE*, 4:3, April 1998, P.4.

Canning, C. "Implementing Visual Support in F/SL Materials", *Visual Support in English Language Teaching (VSELT) Newsletter* 1:1, March 1998 P.1-2.

Canning, C. "Visual Support and Language Teaching", *TESOL Arabia News*, Volume 5:4, March 1988, P.18.

Canning, C. "Theoretical use of Visuals". In Christine Canning and Jane Koester's *Illustrated Visual Aids for Academic English*, 1:1, Fall 1997 Pp. 2-5.

Clark, C and Clark D. (1977) *Psychology and Language: An Introduction to Psycholinguistics*, Hardcourt and Brace, NY.

Carrell, P and Eisterhold, J,. (1987) *Schema Theory and ESL Reading Pedagogy*, Longman Press, London.

Fleming, M. "Five Fold Classification Review" In *AV Communication Review*, Fall 1967, Pp. 246-258.

Herschenson, H: (1980) *The Psychology of Perception*, John Wiley and Sons, NY.

Gibson E. and Levin, H. (1975) *The Psychology of Reading*. MIT Press, Cambridge, MA.

McKim, R. *Thinking Visually* . Part of the Life Long Learning Series, Bel Air, CA.

Polyak, D. (1968) *The Human Body*, NY University Press, NY.

Sless, D. (1981) *Learning and Visual Communication*, Division of Croom Helm, John Wiley and Sons, NY.

Sultan, A. and Jones, M.(1995) "The Effects of Computer Visual Appeal on Learners' Motivation". In *Eyes of the Future: Converging Images, Ideas, and Instruction. Selected Readings from the Annual Conference of the International Visual Literacy Association*, 27th , Chicago, IL October 18-22, 1995.

Torrence, R,. (1970) "Domains of the Brain" CAL Tech Projects Studies and Research Reports.

Whorf, B., (1956) *Language, Thought and Reality*. Selected writings of BL Whorf, J Carroll (ed.) NY: Wiley.

Williams, T,. "So what is so different about visuals?" *Journal of Technical Communication*, 40:4, November 1993.

How Different Examination Boards Present Vocabulary

KEITH ALDRED
United Arab Emirates University

In this paper an analysis is made of the American and British approaches to vocabulary, by looking at the TOEFL and the UCLES examinations. A synopsis is taken from "Building Skills for the TOEFL" (King and Stanley, Nelson, 1989). The objectives are reviewed. The range of UCLES examinations from PET to CPE is looked at, with reference to the hand-books, together with sample papers.

The TOEFL (Test of English as a Foreign Language) is for international students planning to enter a university in the USA, at either graduate or undergraduate level. The time scale here discussed is that at May 1998. The emphasis is on vocabulary, consequently the section analyzed is the third section entitled Reading Comprehension and Vocabulary. The remaining sections, Listening Comprehension, Structure and Written Expression as well as the Test of Written English are omitted from the formal presentation, because the element of vocabulary therein is not stressed as evidently as in the Reading Comprehension and Vocabulary.

The Reading Comprehension - Vocabulary Section contains some sixty items divided into two parts, to be completed in forty five minutes. It is not intended that the student re-reads the question. It is assumed that either s/he will know the answer immediately or not at all.

For the vocabulary items, each item consists of one sentence in which a word or phrase is underlined, below which there are four other words or phrases. The student has to choose the answer which is the closest in meaning to the underlined word or phrase.

To become competent, prior to actually taking the TOEFL students are encouraged to increase their vocabulary systematically. As they read, they should look for contextual clues to the meaning of unknown words, as well as notice the grammatical function of the words. The strength of this objective lies in the fact that, with an expanded vocabulary, the student will be better equipped to deal with the test when it comes.

However, it must be stressed that in the vocabulary items themselves, there are no contextual or grammar clues, let alone stem or visual clues. This in short means, as said earlier, either the student knows the answer or not. In such a situation some familiarity with lexical items is only to be expected. It may help to focus on the underlined words and the four possible answers, looking for the most exact synonym. Sample questions are as follows:

Practice Test I, Section 3

1. Soaring rates of interest have recently made it difficult for young couples to buy their own homes.

   (A) rapidly rising
   (B) very expensive
   (C) slowly rising
   (D) extremely painful

2. Many companies have commented on the government's gratuitously complex labeling requirements for all canned food.

(A) insistently
(B) thankfully
(C) freely
(D) unnecessarily

3. When Lee Iacocca took over the Chrysler Corporation, he insisted that the changes he would introduce would not be merely cosmetic.

(A) fanciful
(B) structural
(C) superficial
(D) invented

4. Meteorologists are at odds over the workings of tornadoes.

(A) mystified
(B) in disagreement
(C) up in arms
(D) in disarray

"Building Skills for the TOEFL" p.484

The British approach to vocabulary will now be considered. In this respect the series of examinations offered by U.C.L.E.S. is investigated. The first examination under consideration is the PET (Preliminary English Test). The assessment aims of PET are designed to ensure that the test reflects the use of language in real life.

Vocabulary is mentioned in the outline for the Reading Test. Although there will be a large amount of unfamiliar vocabulary, the candidates will not be required to understand such vocabulary in order to answer the question set. In addition, there is a short text containing numbered blanks, with a multiple choice question for each blank. The blanks are designed to test vocabulary and grammatical points. An example of this type of question follows:

TRAVELLING IN THE LAKE DISTRICT

The Lake District is (0) .... popular for holidays all year round. Roads leading into the area have been improved in (26) .... years. Within the area, however, many roads are (27) .... and winding with steep hills, and it may not be safe to drive (28) .... roads like this when they are (29) .... in ice.

For the mountain walker a word of warning - every season visitors (30) .... lost or are injured and (31) .... to be rescued by the Mountain Rescue teams. This kind of problem can be (32) .... by following a few simple safety rules. When exploring the mountains, wear warm (33) .... and strong boots, and take a map and a small (34) .... of food. Don't go off alone and always tell someone where you (35) .... to go.

0	A	very	B	lots	C	much	D	many
26	A	recent	B	next	C	last	D	close
27	A	thin	B	slim	C	narrow	D	shallow
28	A	on	B	above	C	by	D	in

"PET Handbook" p.10

The second examination under consideration is the FCE (First Certificate in English). As with the preceding test, this one and the following ones are analyzed as of May 1998. Vocabulary is emphasized in the Use of English paper, throughout the different parts, as illustrated in the following format:

Part	Task Type and Focus	Number of Questions	Task Format
1	Multiple choice cloze An emphasis on vocabulary	15	A modified cloze text containing 15 gaps and followed by 15 four-option multiple choice questions.
2	Open cloze Grammar and vocabulary	15	A modified cloze text containing 15 gaps.
3	'Key' word transformations Grammar and vocabulary	10	Discrete items with a lead-in sentence and a gapped response to complete using a given word.

4	Error correction An emphasis on grammar	15	A text containing errors. Some lines of the text are correct, other lines contain an extra and unnecessary word which must be identified.
5	Word formation Vocabulary	10	A text containing 10 gaps. Each gap corresponds to a word. The 'stems' of the missing words are given beside the text and must be transformed to provide the missing word.

"FCE Handbook" p. 48

Part I is not only a test of meaning, but the answer must also fit in with the grammar of the sentence. Collocations and phrasal verbs are also tested. Examples follow illustrating the various methods of testing vocabulary at this level:

Part 1

For Questions 1-15, read the text below and decide which answer A, B, C or D best fits each space. There is an example at the beginning (0).

Write your answers on the separate answer sheet.

Example:

0 A hour B minute C time D day

	A	B	C	D
0	□	▨	□	□

OUR SHIP SETS SAIL

There was so much to do at the last (0) .... that there was not time to be nervous. All of us wanted to be on our way (1) .... to sea, but there was still one more problem to overcome: the anchor became (2) .... under a rock. We (3) .... it a mighty pull, but it would not come loose. In the end a fishing boat had to pull us (4)................

1	A	out	B	across	C	down	D	over
2	A	stuck	B	fixed	C	attached	D	held
3	A	took	B	put	C	gave	D	let
4	A	clean	B	free	C	safe	D	secure

"FCE Handbook" pp. 52/53

Part 2

For Questions 16-30, read the text below and think of the word which best fits each space. Use only one word in each space. There is an example at the beginning (**0**). Write your answers on the separate answer sheet.

Example:

0	*than*

NICOLAS-FRANCOIS APPERT (1749-1841)

Tinned food and drink is big business: every day more (0) .... 175,000 million tins are sold throughout the world. The process by (16) .... food products can be preserved in tins was invented in 1810 by a Frenchman (17) .... Nicolas-Francois Appert.

"FCE Handbook" p.54

Part 3

For Questions 31-40, complete the second sentence so that it has a similar meaning to the first sentence, using the word given. Do not change the word given. You must use between two and five words, including the word given.· There is an example at the beginning (0).

Write only the missing words on the separate answer sheet.

Example:

0 You must do exactly what the manager tells you.

carry

You must .... instructions exactly.

The gap can be filled by the words 'carry out the manager's' so you write:

0	*carry out the manager's*

"FCE Handbook" p.55

Part 5

For Questions 56-65, read the text below. Use the word given in capitals at the end of each line to form a word that fits in the space in the same line. There is an example at the beginning (0). Write your answers on the separate answer sheet.

Example:

0	*traditional*

PUPPET SHOWS

Puppets are dolls representing (0) .... or modern characters in stories. **TRADITION**
They are a popular form of (56) .... for both children and adults. **ENTERTAIN**
Some puppets seen in Europe today were (57) .... created **ORIGIN**
in Italy in the 1500s. A puppet show was an (58) ..... way to **EXPENSIVE**

"FCE Handbook" p. 58

It would undoubtedly help the candidate to learn words and expressions in context for answering Part Two; whereas in Part Three an awareness of parallel and synonyms expressions helps. In Part 5 the student needs to adopt a systematic, methodical approach to different types of word formation.

The third British examination under discussion is the CAE (Certificate in Advanced English). In the Reading Paper unknown lexis may be tested if it can reasonably be expected that meaning can be deduced from context. Paper 3 "English in Use" tests control of lexis amongst other things. There is a cloze text modified to place emphasis on lexical words. Recognition and correction of errors within a text may include errors of lexis. Recognition and manipulation of items of vocabulary to complete a text so that it is stylistically appropriate is also tested. Examples follow:

CLOZE TEXT 1
SECTION A

CRIME - REVERSING THE TREND

Crime, as we are all (**0**) ...., has been a growing problem all over the world in the last thirty years. But we are not (**1**) .... against crime. Much is being done - and more can be done to reverse the trend. You can play a part in it.

0	[✓ A]	aware	B	conscious	C	informed	D	known
1	A	unprepared	B	hopeless	C	powerless	D	weak

"CAE Handbook" p. 54/55

A further cloze text puts an emphasis on structural words.

SECTION B

RECOGNITION/CORRECTION OF ERRORS

In most lines of the following text, there is either a spelling or a punctuation error. For each numbered line 31 -47, write the correctly spelled word(s) or show the correct punctuation in the boxes on your answer sheet. Some lines are correct. Indicate these lines with a tick (_) in the box. The exercise begins with two examples (0).

0 When Deansgate was a narrow street and the sight of
0 Central Station was a squalid slum, Wood Street Mission was
31 founded. In 1869, according to a contempory police
32 officer, the neighbourbood was 'the rendezvous of thieves,
33 the worst haunt of vice'. In the Mission building hundred's
34 of meals were served and thousands of pears of shoes
35 given away. At Christmas four hundred tramps, and
36 criminals came to a meal and a service; in the
37 summer hundreds of children queud to be taken out
38 for a day at the seaside. Every night the streets
39 were searched for homeless boys sleeping in door ways
40 and under market stalls. They were given beds in
41 Wood Street jobs were found for them and many were
42 sent to live in canada. The new Superintendent
43 in 1892 was an excriminal and he founded a
44 holiday camp at St. Anne's-on-Sea which could
45 accommodate a hundred and twenty children. Many
46 local residence still remember happy holidays there. The
47 Mission still provides about a thousand familys a year
with clothing and helps or advises many more.

31	contemporary	**39**	doorways
32	_	**40**	_
33	hundreds	**41**	Street, jobs/Street. Jobs/Street; jobs
34	pairs	**42**	Canada
35	tamps and	**43**	ex-criminal
36	_	**44**	which
37	queued	**46**	residents
38	_	**47**	families

"CAE Handbook" p.57 + 63

The final examination under consideration is the CPE (Certificate of Proficiency in English). In the Reading Comprehension paper the focus includes lexical appropriacy. Semantic sets and collocations, phrasal verbs and use of grammatical rules and constraints all have lexical connections. In the Use of English paper vocabulary is tested at word, sentence and paragraph level. Modified cloze, discrete sentence transformations and discrete gapped sentences are used. Examples follow:

MODIFIED CLOZE

Fill each of the numbered blanks in the passage with one *suitable word.*

Gardens

Redesigning a garden can be a fascinating experience. Both the first-timer and the (1) .... gardener confronted with (2)

.... a task are (3) .... of bright hopes and grand ideas. Realising them is (4) .... matter altogether, but great expectations certainly (5) .... the basis of many fine creations.

1 experienced/expert/professional/seasoned

"CPE Handbook" p.p. 44 + 51

DISCRETE SENTENCE TRANSFORMATIONS

Finish each of the following sentences in such a way that it is as similar as possible in meaning to the sentence printed before it.

EXAMPLE: Immediately after his arrival things went wrong.
ANSWER: No sooner *had he arrived than things went wrong*

(a) The rescue team will try again to find the missing seamen tomorrow morning.

Another
....
....

....
....
(Marks for each portion as shown; some variations in answers allowed)

(a) ...attempt (1) will/is going to be made/take place/is to take place
tomorrow morning (1)

"CPE Handbook" pp 45 + 51

DISCRETE GAPPED SENTENCES

Fill each of the blanks with a suitable word or phrase.

EXAMPLE: He doesn't mind one way or the other; it makes *no difference to* him.

(a) Lavinia didn't tell Charles the truth .... upsetting
him.

(b) It was so quiet one .... a pin drop.

(1 mark for each item; some variations in answers allowed)

(a) ... for fear of/because she was/so as/in order to avoid

(b) ... could hear/would/could have heard

"CPE Handbook" pp. 46 + 52

DISCRETE SENTENCES

For each of the sentences below, write a new sentence as similar as possible in meaning to the original sentence, but using the word given. This word must not be altered in any way.

EXAMPLE: Not many people attended the meeting.
turnout

ANSWER: There was a poor turnout for the meeting.

1. Some customers' payments are in arrears.
behind

Answer: Some customers are/have fallen behind with their payments

"CPE Handbook" pp. 47 + 52

The range of this paper is wide. Omissions about TOEFL's extent of dealing with vocabulary were clarified by a member of the audience with practical experience in the field. The author is grateful for the participation of the audience during the presentation. Things are changing - TOFEL is shortly to be computerized and the CPE is set for revision. Computer-based testing is also being followed up in the U.K.

Why vocabulary is tested in the ways outlined is a question to which we would all like to know the answer. Does testing in such ways reflect good teaching?

APPENDIX

The Second Annual CTELT Conference Abstracts

Date: Wednesday, May 6, 1998

Keynote Speaker
(9:30-10:30 Conference Room)

The Impact of High Stakes Testing on Teaching and Learning

DIANNE WALL
Lancaster University

'Impact' refers to any of the effects that a test may have on individuals, policies or practices - within the classroom, schools, the educational system or society as a whole. The notion that important tests will have important effects has been discussed in the general education literature for many years, but it was not until this decade that language educators began to pay serious attention to whether tests were as powerful as was generally accepted, where their supposed power came from, what kinds of effects they really had on teaching and learning, and what other factors in the educational context might influence what happened in the classroom. This presentation will review the functions of testing in modern society, ideas from general education and language education concerning the concept of impact, and what recent research findings suggest about using tests to change classroom practice. Special attention will be paid to the relevance of innovation theory to the study of test impact.

Dianne Wall is a lecturer at the Institute for English Language Education, Lancaster University, United Kingdom. She specializes in language testing and has conducted seminars and consultancies in many countries, most recently serving as adviser for national test development projects in Russia and the Baltic States. She is co-author, with Charles Alderson and Caroline Clapham, of Language Test Construction and Evaluation (Cambridge University Press, 1995), and is co-editor of Language Testing Update, the official newsletter of the International Association of Language Testers.

Dianne Wall can be contacted at:

Dianne Wall
Institute for English Language Education
George Fox Building
Lancaster University LA6 4YJ
United Kingdom

E-mail d.wall@lancaster.ac.uk
Tel 44 1524 592436
Fax 44 1524 594149

Program and Schedule

8:30 - 9:15 CONFERENCE ROOM	**Registration and coffee**
9:15 - 9:30 CONFERENCE ROOM	**Opening Ceremonies**
9:30 - 10:30 CONFERENCE ROOM	**Keynote Address** The Impact of High Stakes Testing on Teaching and Learning. Dianne Wall, Lancaster University
10:40 - 11:25	**Concurrent Presentations 1**
ADMINISTRATION CONFERENCE ROOM	Assessment at the KFUPM Orientation English Program Mark S. Algren, King Fahd University of Petroleum and Minerals.
SKILLS ROOM 202	Using Written Communication Descriptors to Assess Varied Task Types. Nicola Marsden, Higher Colleges of Technology
ALTERNATIVE ASSESSMENT ROOM 203	Altemative Assessment in a Content Process Program. Bahia ASSESSMENT Diefenbach and Donal Huriey, UAE University
RESEARCH ROOM 217	Developing a Testing Sense: A Case Study in ESP. Molly Kirk, UAE University
PRE-TERTIARY ROOM218	Ministry of Education Textbooks and Assessment in Focus Ali Abdel-Fattah, Sally Ali and Adrianna Sutherland, UAE University
TECHNOLOGY LAB 213	Results of Computer Adaptive Testing Experiments at LAB UGRU. Chris Head and Royal Hansen, UAE University
11:35 - 12:20	**Concurrent Presentations 2**
ADMINI- STRATION ROOM219	Testing Myths and the Icons of PET, TOEFL, and IELTS Graeme Tennent, UAE University
SKILLS ROOM 202	How Different Exam Boards Present Vocabulary. Keith Aldred, UAE University
ADMINI- STRATION ROOM 203	Developing a New Testing Program for the University of Sharjah. Allison Curtis, University of Sharjah
RESEARCH ROOM217	The Creation of Rapport, Interlocutor Consistency, and Features of Co-Construction in Oral Proficiency Interviews: A Pilot Study. John Pollard, Saudi Development and Training

PRE-TERTIARY ROOM 218	Curriculum Evaluation in Oman Loring Taylor, Sultan Qaboos University
TECHNOLOGY LAB 213	The Effect of Computer Generated Graphics on Reading Scores. Lisa Barlow, Cecilia Kawar and Christine Canning UAE University
12:30 - 1:30 CONFERENCE ROOM	**Lunch**
1:40 - 2:25	**Concurrent Presentations 3**
ADMINI- STRATION CONFERENCE ROOM	Testing Issues in a Start-Up Intensive English Program. John Shannon, American University of Sharjah
SKILLS ROOM 202	Oral Assessment Bryan Davis, UAE University
LEGAL ISSUES ROOM 203	What Test Writers Need to Know about Copyright Law. Christine Canning, UAE University
ALTERNATIVE ASSESSMENT ROOM217	Student Designed Tests: "Why teacher? That's your job!" Phil Cozens, Higher Colleges of Technology
PRE-TERTIARY ROCBM 218	Q&A Session on Teacher Training, Testing, and Evaluation. Salah Troudi, Hedi Guefrachi, and George Murdoch, UAE University
TECHNOLOGY LAB 109	Computer-Based Listening Assessment Christine Coombe, Chris Head, Nancy Hubley and Jon Kinney, UAE University
2:35 - 3:20	**Concurrent Presentations 4**
ADMINI- STRATION CONFERENCE ROOM	Performance-Based Language Assessment in a Vocational Context. Elizabeth Howell, Higher Colleges of Technology
SKILLS ROOM 202	Pitfalls to Avoid in Listening Tests Paul Houghton, Higher Colleges of Technology
RESEARCH ROOM 217	C-Testing: A Theory Yet to be Proved Neil McBeath, Royal Air Force of Oman
PRE-TERTIARY ROOM 218	Grade Inflation: Who is to Blame? Abdullah Soliman, UAE Ministry of Education
POSTER SESSION ROOM 110	Proficiency Testing John Pollard, Saudi Development and Training
ESP	Assessment in the Medical Faculty of UAE University

ASSESSMENT LAB 211	Peter Ridding, Angela Kiwan and Jo Sanders, UAE University
3:20 - 3:40 ROOM11O	**Refreshments**
3:45 - 4:30 CONFERENCE ROOM	**Closing Panel Discussion** Placement Testing

Concurrent Presentations

First Session: 10:40 - 11:25

Assessment at the KFUPM Orientation English Program

MARK S. ALGREN
King Fahd University of Petroleum and Minerals

Assessment in the OEP is based on common exams three times per term. Maintaining exam integrity is paramount since grades are based on a curve. The presenter will review exam structure, preparation, administration and grading procedures with particular emphasis on procedures aimed at maintaining integrity.

Mark S. Algren has taught in Saudi Arabia for 10 years in both public and private sector organizations. Prior to completing his MA in ESL at Southern Illinois University, he taught in Hong Kong. On leave from the Applied English Center at the University of Kansas, he is the Director of the Orientation English Program at King Fahd University of Petroleum and Minerals.

Using Written Communication Descriptors to Assess Varied Task Types

NICOLA MARSDEN
Higher Colleges of Technology

This paper describes the written communication descriptors created by Academic Services, Higher Colleges of Technology, for use in the if assessment of writing in tertiary EFL programs. It defines a variety of task types in use in the HCT's different programs. It also addresses the issue of I combining task fulfillment measures with linguistic descriptors.

Nicola Marsden is Coordinator of the EFL Curriculum in the Certificate and Diploma Program of the HCT. She has taught EFL in the Middle East and Bangladesh and has organized several testing prograrr.s.

Altemative Assessment in a Content Process Program

BAHIA DIEFENBACH AND DONAL HURLEY
United Arab Emirates University

This presentation compares traditional and alternative assessment modes. Some disadvantages of traditional testing, mainly validity issues, are outlined. Then. administration of alternative assessment, its development and practicality are discussed with particular reference to how alternative assessment can solve problems posed by traditional modes of assessment. Examples will be drawn from recent curriculum development work .

Bahla Diefenbach and **Donal Hurley** are lecturers in the UGRU Mathematics and Computer Program at UAE University. They recently developed an alternative assessment component for an integrated mathematics/information technology/ science curriculum for the UAE's new Zayed University.

Developing a Testing Sense: A Case Study in ESP

MOLLY KIRK
United Arab Emirates University

This case study reviews testing practices during three consecutive midterm examinations for the ESP-Education course. Although the content and examination criteria remained similar, the test results fluctuated significantly with attempts to improve the exam. Possible explanations as to why this happened will be discussed. Findings suggest that test writing is part art, part science, but a "sixth sense" for good test writing can also be developed.

Molly Kirk is Lead Teacher for the ESP-Education Unit at UAE University. She holds two MAs in both TEFL and Chinese and is currently enrolled in a Ph.D. program. She has taught in the US, the Far East and the Middle East.

Ministry of Education Textbooks and Assessment in Focus

ALI ABDEL-FATTAH, SALLY ALI AND ADRIANNA SUTHERLAND
United Arab Emirates University

This presentation aims to examine the testing formats currently used in the EFL textbooks and exam papers of the UAE government schools. Alternative testing formats which may enhance learners' comprehension, progress and achievement in the skill areas will be suggested.

Ali Abdel-Fattah, Sally Ali and **Adrianna Sutherland** are all lecturers in the UGRU English Program at UAE University. Ali has a Ph.D. in TEFL from Indiana University of PA. Adrianna has an MA in ESL from the University of Minnesota. Sally has a Ph.D. in TESOL from Georgetown University.

Results of Computer Adaptive Testing Experiments at UGRU

CHRIS HEAD AND ROYAL HANSEN
United Arab Emirates University

Computer Adaptive Testing (CAT) is an extension of traditional Computer Based Testing (CBT), both of which can be implemented with a simple and intuitive student interface. In the UAE University General Requirements Unit, CBT has been used for many years and has recently been improved with a Windows version. Experiments using an adaptive testing algorithm to drive the program have demonstrated the positive effects that adaptive testing can have on both drill-andpractice and assessment. The presentation includes experiment results as well as explanations of the adaptive process.

Chris Head is Assistant Coordinator of the Math/Computer Program in the UAE University General Requirements Unit. Chris led the development team for the Computer Generated Examination software. Royal Hansen is a graduate in Computer Science from Yale University. He is currently a United States Fulbright Fellow conducting research on CAT at UAE University.

Second Session: 11:35- 12:20

Testing Myths and the Icons of PET, TOEFL, and IELTS

GRAEME TENNENT
United Arab Emirates University

This paper sets out to question, not demolish, certain tendencies in testing in the Gulf. These can be defined as the processes of seeking validation and broad acceptability through the use of external examinations such as the three in the title. The questions are: how valid are these examinations in terms of the learner population here and, perhaps more importantly, what is it which drives testers to these tests?

Graeme Tennent is Coordinator of Language Skills in the DeDartment of English Language and Literature at UAE University. In a previous incarnation. he coordinated testing in the foundation course at UAE University. Other lives have been passed in Libya, Borneo, Yemen, and Sudan.

How Different Exam Boards Present Vocabulary

KEITH ALDRED
United Arab Emirates University

Vocabulary is an important component of standardized tests both in the United States and Britain. An analysis is made of the American and the British approaches to assessing vocabulary by reviewing the TOEFL and UCLES examination formats. Objectives are reviewed and examples are presented and discussed.

Keith Aldred is a lecturer and member of the Testing and Measurements Committee in the United Arab Emirates University General Requirements Unit. He has an MA in TEFL from the University of Kent.

Developing a New Testing Program for the University of Sharjah

ALISON CURTIS
University of Sharjah

This paper is an analysis of statistics concerning the progress of the University of Sharjah's first intake of students. with particular reference to English. Its purpose is to determine the accuracy of the current placement process and provide information for curriculum evaluation. Students will be ranked by school leaving scores, internal placement test results, January exam results, and May TOEFL scores. Comparisons will be made by gender. career and faculty choice, and private or government school background. Local and regional implications of these results will be examined.

Allison Curtis is a lecturer at the English Center of the University of Sharjah. She has an MA in Applied Linguistics and TESOL from Leicester University. She has taught at the Higher Colleges of Technology and in Saudi Arabia. Allison is also the TESOL Arabia joint Sharjah Representative.

The Creation of Rapport, Interlocutor Consistency, and Features of Co-Construction in Oral Proficiency Interviews: A Pilot Study

JOHN POLLARD
Saudi Development and Training

Recent research has questioned the validity and reliability of OPIs vis-a-vis discourse structure, with raters applying individual standards to test designs. One suggestion has been the training of raters-as-interlocutors. This pilot study begins to examine this issue by focusing on the relative rapport by different interlocutors.

John Pollard has an MA in TESOL/TEFL. He has 16 years' experience of working in language testing in a variety of contexts. He has worked for the British Council, the British ODA, and UNESCO. He has published a number of papers on teaching methodology and second language testing.

Curriculum Evaluation in Oman

LORING TAYLOR
Sultan Qaboos University

The present paper examines the techniques and results of a major curriculum review project in Oman. From 1974 to 1980, a new curriculum was designed specifically for Omani public schools. Five years later, a review of this curriculum was initiated to determine the extent to which the educational objectives of the curriculum were being realized. Conclusions from the primary, preparatory, and secondary levels have been incorporated into a 25year plan to restructure the entire Omani educational system.

Loring Taylor is an Associate Professor and Deputy Chair of the English Department at Sultan Qaboos University. He holds a Ph.D. from UC Santa Barbara and has taught in the U.S., Romania, Yemen, Oman, and Jordan.

The Effect of Computer Generated Graphics on Reading Scores

LISA BARLOW, CECILIA KAWAR AND CHRISTINE CANNING
United Arab Emirates University

This paper examines the effects of computer generated visual prompts on reading comprehension scores of students in the UAE University UGRU English Program. Research was conducted to investigate whether the presence of supporting graphics influenced comprehension scores. Results from the study will be disseminated.

Lisa Barlow, Cecilia Kawar and **Christine Canning** are lecturers in the UAE University UGRU English Program. Lisa has an MA from the University of Chicago and is a member of the Testing and Measurements Committee. Cecilia is an artist specializing in computer graphics and a member of the Educational Technology Committee. Christine chairs the Media Graphic Visual Aids Committee.

Third Session: 1:40 - 2:25

Testing Issues in a Start-Up Intensive English Program

JOHN SHANNON
American University of Sharjah

One of the most critical issues facing faculty and administration at the American University of Sharjah IEP is student placement into and exit from the program. This session will discuss some of the testing issues that have confronted us as we have worked toward establishing the IEP. Our purpose is not to provide answers to testing problems but to elicit input on questions that may be relevant to many EFL professionals in the Gulf region and beyond.

John Shannon is the Chair of the English Department and the Acting Director of the IEP at the American University of Sharjah.

Oral Assessment

BRYAN DAVIS
United Arab Emirates University

The presenter will briefly review the oral testing literature as well as discuss a variety of issues pertaining to the successful administration of oral examinations. Examples of these issues include the delivery and administration of oral exams, grading and collating scores, assessor training and the use of technology in oral testing.

Bryan Davis is a lecturer in the UGRU English Program at the UAE University. He has an MA in Applied Linguistics from the University of Dublin. He has taught in the United Kingdom, Botswana. and Japan.

What Test Writers Need to Know about Copyright Law

CHRISTINE CANNING
United Arab Emirates University

This paper will focus on copyrighting testing materials and other related works under United States and international copyright VOWS. Twelve of the most frequently asked questions by test writers will be discussed. Information related to obtaining copyrights and patents will be disseminated.

Christine Canning is a lecturer at the UAE University where she chairs the Media Graphic Visual Aids Committee. She has taken law courses and presented her work on copyright law at the University of South Florida. Christine is the TESOL Arabia Al Ain Representative.

Student Designed Tests: "Why teacher? That's your job!"

PHIL COZENS
Higher Colleges of Technology

The paper discusses the reactions of students and teachers after they were asked to prepare the reading component of their second progress test. The rationale behind student-designed tests and their possible motivational effect on students will be discussed.

Phil Cozens teaches English language and Computers at the Higher Colleges of Technology in Ras Al Khaimah. He has an MA in Linguistics from the University of Surrey. He has taught in Hong Kong as well as the Middle East.

Q&A Session on Teacher Training, Testing, and Evaluation

SALAH TROUDI, HEDI GUEFRACHI AND GEORGE MURDOCH
United Arab Emirates University

This session provides a forum for discussing issues in pre-service teacher training as well as in-service professional development in testing. Teachers and administrators are encouraged to come with their questions and concerns.

Salah Troudi, Hedi Guefrachi, and **George Murdoch** are faculty members in the UGRU English program at UAE University. They are experienced teacher trainers and active in outreach programs with the UAE Ministry of Education. Salah chairs the TESOL Arabia Teacher Training SIG, while George is the TESOL Arabia Executive Secretary. Hedi supervises the UGRU English Distance Learning Program.

Computer-Based Listening Assessment

CHRISTINE COOMBE, CHRIS HEAD, NANCY HUBLEY AND JON KINNEY
United Arab Emirates University

The UGRU English Program at the UAE University has been using computerbased testing (CBT) since 1993 to assess reading, vocabulary, and language use. To date, technical constraints have precluded testing listening by CBT. This session demonstrates a CBT prototype based on current research in listening assessment.

Christine Coombe and **Jon Kinney** have conducted and published research in academic listening comprehension and assessment. Chris Head and Nancy Hubley have developed an English language computer based testing program at UAE University. Christine and Nancy co-chair the TESOL Arabia Testing SIG, while Jon chairs the Research SIG.

Fourth Session: 2:35 to 3:20

Performance-Based Language Assessment in a Vocational Context

ELIZABETH HOWELL
Higher Colleges of Technology

This paper defines some of the issues involved in the integrated testing of language and subject content areas, focusing on the assessment of language competency in a task-based vocational context. It describes some of the tasks and criteria used for assessing work-related tasks in a second language context in business and technical courses of the Certificate Diploma (CD) Program at the Higher Colleges of Technology.

Elizabeth Howell has taught and tested EF/SL in Europe. the Middle East and the Far East and is Manager of General Education at HCT.

Pitfalls to Avoid in Listening Tests

PAUL HOUGHTON
Higher Colleges of Technology

This session will discuss some problems in writing items for listening tests! many of which are also applicable to the testing of other skill areas. Participants will listen to excerpts of several tests in order to identify problematic test items in three main areas: productive-response items, receptive response items, and the applicability of using published listening materials for testing.

Paul Houghton is Program Assessment Editor for the Foundations Program at HCT. He also worked in the U.K., France, Bahrain, and Oman. Paul holds a BA in History and Archaeology and a Diploma in Teaching English Overseas

C-Testing: A Theory Yet to be Proved

NEIL MCBEATH
Royal Air Force of Oman

Although C-tests have been widely endorsed since they were developed 15 years ago by Klein-Braley, this paper suggests that their endorsement may I be premature. Testers have altered the original concept. There is evidence E to suggest that the original research may not be generalizable to EFL.

Nell McBeath, contract Flight Lieutenant in the Royal Air Force of Oman, was the first English Education Officer to receive the Distinguished Service Medal. He holds an MSc. in Teaching English and a Trinity College TESOL Diploma.

Grade Inflation: Who is to Blame?

ABDULLAH SOLIMAN
United Arab Emirates Ministry of Education

We live in an age of economic inflation with skyrocketing prices and an age of grade inflation where many students get higher grades than they deserve. This presentation explores some of the educational and social issues surrounding grade inflation and its impact on teaching and testing.

Abdullah Soliman, a Supervisor for the Al Ain Educational Zone for the Ministry of Education, has worked in the UAE for 25 years. Abdullah frequently presents at regional conferences on issues related to teacher development.

Proficiency Testing: Poster Session

JOHN POLLARD
Saudi Development and Training

This poster presentation depicts tasks from a recent CD ROM proficiency test and highlights a number of innovative features. John will be available to discuss his project in the Heritage Center, Room 110.

John Pollard has an MA in TESOL/TEFL. He has 16 years' experience of working in language testing in a variety of contexts. He has worked for the British Council, the British ODA, and UNESCO. He has published a number of papers on teaching methodology and second language testing.

Assessment in the Medical Faculty of UAE University

PETER RIDDING, ANGELA KIWAN AND JO SANDERS
United Arab Emirates University

The UAE University Faculty of Medicine teaches its curriculum in English. Thus, student language proficiency is of great importance. This presentation will show how internationally validated external tests, subject-integrated tests and in-house task-based assessments are all used to assess English skills. Questions will be raised about whether these assessment instruments are compatible and whether scores on external tests are good indicators of academic success in medical sciences.

Peter Ridding is Coordinator of the ESP Program in the UAE University Faculty of Medicine. He has also taught in Germany, Kuwait and Greece. Angela Kiwan is a lecturer in the ESP Program at the Faculty of Medicine. Her interests include curriculum design and materials writing. Jo Sanders is a lecturer in the Faculty of Medicine at UAE University. She has also taught in Bahrain and Syria.

Closing Panel Discussion

Final Session: 3:45 - 4:30

Current Trends and Issues in English Language Placement Testing

Panel moderator:	Chris Pearson, TESOL Arabia President
Panel participants:	Dianne Wall, Lancaster University Karen Asenavage, UAE University Mark Algren, KFUPM John Shannon, American University of Sharjah Elizabeth Howell, HCT Yahia Ahmed, Kuwait University

This panel of experienced program administrators and testers will address your questions concerning student placement. To submit a question, please fill out the insert form provided in the conference packet and place it in the box on the registration desk. Questions must be received by 3 p.m.

U.S. Department of Education
Office of Educational Research and Improvement
(OERI)
National Library of Education (NLE)
Educational Resources Information Center
(ERIC)

ERIC®

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Current Trends in English Language Testing Proceedin	
Author(s): Coombe, Christine	
Corporate Source:	Publication Date: Oct 98

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 in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	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)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
↑ □	↑ □	↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche onl

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

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: Christino Coombe Testing + Measurements Supervisor	
Organization/Address: UGRU English UAE University P.O. Box 17172 Al Ain UAE	Telephone: 971-3-6194796 E-mail Address: coombe@emirates.net.ae	Fax: 971-3-510195 Date: 13 Mar 1999

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.)

Publisher/Distributor:

Address:

Price:

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:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

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