The Role of Assessment by Teachers in School

Glenn Fulcher

1. Introduction

It has been common for students of English as a Foreign Language who wish to take some form of external examination to be tested in what may only be described as a 'formal' manner. This may involve simply reading and writing, as in the University of London (syllabus 161B) overseas paper, or in all skills including speaking, as in the University of Cambridge Local Examinations Syndicate First Certificate in English and Proficiency in English. In the latter case, oral ability is tested by the means of an oral interview. Indeed, from only a cursory glance at the range of proficiency tests in English as a Foreign Language (EFL) on the market, all testing methods are *examination-board* oriented, and not *school* oriented (See Alderson, Kranke and Stansfield, 1987).

What is meant by this? A school or institute in which students are following courses which lead to an external qualification may assess their own students in a number of ways, including formal testing, but *only* in order to place them in the appropriate class (Placement Testing) or to decide if they are making progress in their studies (Achievement Testing). In no case can the class teacher's assessment of the abilities of an individual student be recorded and used as part of the final grade in the external qualification.

This sharply contrasts with the situation in countries where the teacher's assessment of her own individual students does play a role in making up the grades on the final leaving certificate which is awarded. Even in the United Kingdom the Task Group on Assessment and Testing (1988) recommended assessment based in the school through the use of Standard Assessment Tasks in all subjects, whilst the introduction of the General Certificate of Secondary Education (GCSE) places great emphasis on course work and internal assessment, with the appropriate safeguards of moderation procedures. As a result of these developments, the University of Cambridge has introduced the International General Certificate of Education (IGCSE) which is aimed at overseas educational institutions. One of the syllabuses available is English as a Second Language (ESL), and this contains an element of coursework and internal assessment.

As coursework and teacher assessment of students is likely to become more widespread in the EFL world, it is the appropriate time to begin to ask questions about the nature of the assessment procedures used. Is assessment carried out by teachers *different* from the assessment in traditional examinations? That is, does it provide extra information to external examinations, does it overlap, or does it provide exactly the same information? If the latter turned out to be true, then there would be little justification for the expenditure of vast amounts of time, energy and money in training assessors and operating complex moderation systems. This is essentially a question of the *validity* of using teacher assessment, which would have practical consequences for the ways in which we choose to assess students.

No less important is the question of *reliability*. Are teachers capable of reliably assessing their own students? This essentially means following students' progress, using standard assessment tasks, and converting their observations into a number or a letter which represents the ability of any given student.

It was with these questions in mind that two studies were devised in order to analyse a number of assessment techniques. The studies were carried out in a particular context which may well be relevant to the findings, and indicate to the reader where these findings may be limited in terms of application to their own teaching situation. The English Institute in Cyprus has used the University of London GCE (syllabus 161B) overseas examination for many years as an 'end-product' examination for students after nine or ten years of study. With changes in teaching methodology and examinations within the United Kingdom it was decided that the IGCSE ESL examination may very well be appropriate for our use. In order to investigate how the internal and external assessment of students at this Institute were operating, all assessment techniques were analyzed in two studies, one during the academic year 1988/89, and one during 1989/90.

The first study was designed to answer the question of whether the internal assessment provided by the teachers provided insights into the ability of students which was not provided by any external examination. The second study was designed to confirm these findings and, further, to investigate the issue of the reliability of teacher assessment.

2.1. First study

During 1988/1989, 114 students studying at the English Institute received six assessments. Of these, four were externally marked examinations, one was an internally marked examination, and one was a simple letter grade representing the teachers' assessment of student abilities. These are summarised in Table 1.

Letter	Grade from	Code
A	University of Cambridge First Certificate in English	FCE
B	Internal Teacher Assessment of Abilities	TASS
C	English Language Testing Service: Grammar	GRAM
D	English Language Testing Service: Listening	LIST
E	Internal Mock GCE Examination	MOCK
F	University of London GCE English (Overseas 161B)	GCE

Table 1: Assessments used in the first study

The analysis of student results should provide a way in to the study of teacher-based assessment within the school. Of the assessment techniques listed in Table 1, it was hypothesised that B represented teacher-based assessment, with the possibility of E also having some relationship to the ways in which teachers grade internally. The internal teacher assessment of abilities (B) was carried out within the school at the end of the academic year, and took the form of a global mark on a scale of A to E with plus and minus points between the two extreme grades, thus giving 13 possible grade points. Teachers were asked to assign a grade on the basis of their impression of the achievement of each student throughout the school year in relationship to the syllabus which was being followed. As such, it represents a subjective and personal impression which was not controlled by any system of moderation or carefully constructed grade descriptors to aid the teachers in their task. The internal GCE Mock examination (E) was written and marked by teachers, although before marking papers a co-ordination meeting was held to ensure that work was being marked in a similar way by all members of staff. No teacher marked the work of a student whom he or she had taught during the year. The two English Language Testing Service Modules were pilot versions of the new International English Language Testing Service test (see Criper and Davies, 1988 and Hughes, Porter and Weir, 1988 for information on why changes were suggested, and The British Council Information, 1989, for the current format). However, the modules which have now come into service do not differ greatly from the pilot modules.

2.2. Analysis

As a first step in the investigation, all data from all measures were correlated (Table 2).

	FCE	TASS	GRAM	LIST	MOCK
TASS	.49				
GRAM	.68	.33			
LIST	.54	.36	.48		
MOCK	.69	.64	.63	.44	
GCE	.58	.34	.55	.29	.55

Table 2: Correlation matrix of assessments (Study 1)

The question posed was whether or not teacher assessment offers something which is valuable to the overall assessment which other measures do not tap. In order to explore this Principal Components Analysis (PCA) with Varimax rotation was used to further explore the correlation matrix.[1] The results are presented in Table 3.

	Factor 1	Factor 2	Factor 3
FCE	.658	.364	.461
TASS	.146	.946	.156
GRAM	.749	.120	.452
LIST	.160	.193	.928
MOCK	.574	.643	.262
GCE	.883	.187	-.002
Variance explained by rotated components (Eigen values)			
	2.149	1.529	1.371
Percent of total variance			
	35.825	25.476	22.848

Table 3: Rotating loadings on Factors from the PCA (Study 1)

It must be stressed that this study was exploratory. Principal Components Analysis has, in the past, been used in EFL studies in order to draw conclusions regarding the structure of language ability (Oller and Perkins, 1980), but many dangers lie in basing conclusions upon this type of analysis (Woods, 1983). One of the dangers will become clear in what follows.

It is the role of the investigator to label the factors which are produced from the analysis, primarily by using the measures which load highly on particular factors. It is then hoped that tentative hypotheses may be generated about the nature of assessment which may be followed up with a study using other techniques.

In this data, it may be observed that TASS loads very highly on Factor 2. This is followed by MOCK, which was the internally (teacher) assessed written examination. Factor 2 may therefore be labelled 'Teacher Assessment'. LIST loads most heavily on Factor 3, followed by FCE (which contains a listening and oral test), and GRAM. To label this factor we must assume that grammatical knowledge influences scores on aural/oral tests. This does not seem to be an unreasonable assumption. Factor 3 may therefore be called 'Aural/Oral Skills'.

GCE loads most highly on Factor 1, followed by GRAM, FCE and MOCK. The GCE examination primarily contains writing, reading and grammatical exercises; hence, we may assume, the unimportant loading on Factor 3. Grammatical accuracy would seem to be an important factor in the measurement, as indeed it is in the written components of the First Certificate examination. This factor may tentatively be called 'Writing/Reading/Grammatical skills'. This large category cannot be broken down further from this data, but the labelling may be supported by the fact that Factor 1 accounts for 35.82% of variance across measures, much higher than other factors.

In order to examine the relationship between the various measures on the factors isolated it is useful to use Factor Plots. Figures 1 to 3 show the relationship between measures on the three Factors generated by the analysis.

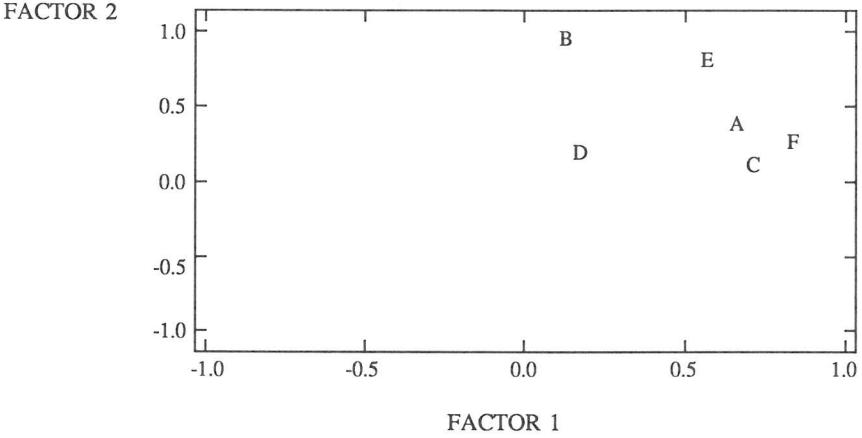

Figure 1: Factor plot for Factors 1 and 2 (Study 1).

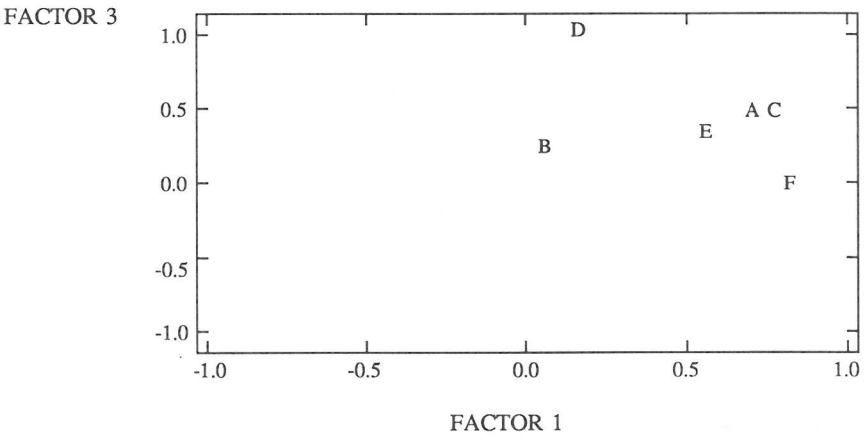

Figure 2: Factor plot for Factors 1 and 3 (Study 1).

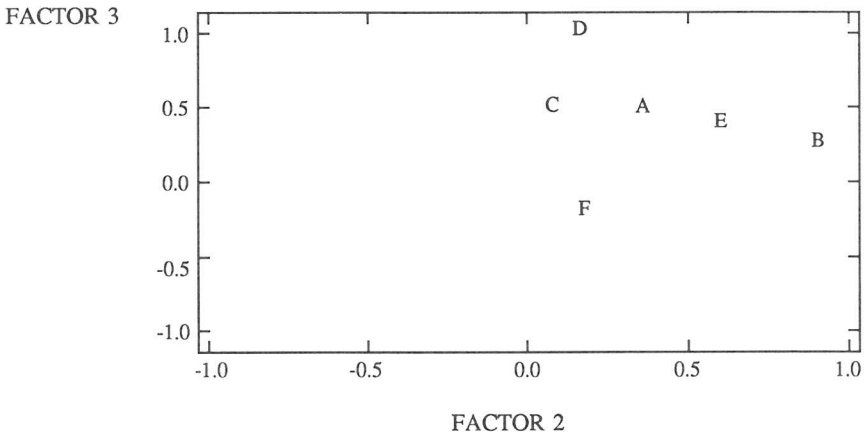

Figure 3: Factor plot for Factors 2 and 3 (Study 1).

2.3. Discussion

This type of exploratory analysis is, almost by definition, circular. For once factors are labelled, the labels are then used to interpret the loadings! However, placing this warning to one side, it does seem that teacher assessment (TASS)

is very different from other measures. It consistently stands alone on the Factor Plots.

The data suggest the hypothesis that teacher assessment is qualitatively different from other modes of assessment.[2] The only measure which loads on all three factors is FCE, which is a multi-skill examination, but recorded here as a global grade. It is also highly likely that the multi-skill ELTS examination taps all factors, but this is hidden here, as GRAM and LIST are reported separately, and the students in the sample did not sit the other components of the test.

15.85% of the variance in these measures was not accounted for in this analysis. It is possible that this is error variance, or may indicate the presence of other factors too small in their influence to be susceptible of analysis. No further explanation can be offered.

It may now be possible to begin to suggest an answer to the first of the questions posed through this exploratory analysis. Teacher assessment of students does seem to provide information not supplied by traditional written examinations, or aural/oral tests in language examinations. However, it would also appear equally true that teacher assessment alone cannot substitute for the information provided by these examinations.

It may be claimed that teachers are sensitive to aspects of student performance which should be assessed, but that this assessment should be balanced by external examinations with a broad skills base, such as the ELTS or the UCLES examinations, rather than a narrow skills base such as the GCE which is still used in many countries such as Cyprus. The IGCSE may therefore be considered as providing an alternative for overseas students at the age of 16. However, the results of the study do not tell us anything about how teachers assess students or whether their assessment is reliable.

With these tentative suggestions in hand from the first study, further questions remained: could the study be reproduced with a separate sample of students, and is teacher assessment a reliable measure of student abilities?

3. Second study

In the second study into the nature of teacher assessment 122 students were used, again from the English Institute, in the academic year 1989/90. However, some of the measures used in the study differed.

Once again, all students took a mock examination which was prepared and marked in exactly the same way as the previous year, the FCE and the GCE. Unfortunately, no ELTS grades were available, but changes were carried out in the way teachers assessed their students. This latter development allows further investigation into the nature of teacher assessment.

The single letter grade method which was employed prior to 1989/90 was a holistic approach which asked the teacher to give a global assessment of student abilities. The new system provided for componential grading in skill areas together with an assessment of the amount and quality of work produced by students outside class (homework). The components of assessment were: reading ability, writing ability, accuracy of spelling, grammatical accuracy and homework. Each component was to be graded on a Likert-type scale of 1 to 5, with 5 being very good and 1 being very poor. Listening and speaking ability were omitted due to the nature of the course which the students were following: these skills were not tested in the end-product examination. The measures used are summarised in Table 4.

Letter	Grade from	Code
A	University of London GCE (Overseas 161b)	GCE
B	Internal Mock GCE Examination	MOCK
C	Internally assessed Reading ability	READ
D	Internally assessed Writing ability	WRITE
E	Internally assessed Spelling ability	SPELL
F	Internally assessed knowledge of/ability in Grammar	GRAM
G	Internally assessed quality/quantity of homework	HOME
H	University of Cambridge First Certificate in English	FCE

Table 4: Assessments used in the second study

The issues to be investigated here are complex, and it was difficult to decide on the techniques which would shed most light on the questions which had been posed. Initially, although undoubtedly unsatisfactory in itself, it was decided to conduct another Principal Components Analysis on the data in order to discover if similar factor patterns emerged. This in itself, with a different sample, would at least tend to suggest whether or not the initial exploratory study was merely an artefact of the sample or not. Secondly, it was decided to regress the components of teacher assessment onto the GCE scores for the 122 students. On the basis of the results of the exploratory study it would be expected that the multiple correlation coefficient would not be exceptionally high, showing relative independence of teacher assessment from external examination results. These two techniques and the results are presented in Section 4.

With regard to reliability, two techniques were used. Firstly, each component of teacher assessment was treated as if it were a separate 'test item', providing a five item test. In this way it is possible to treat teacher assessment as a single test of 25 marks (5 for each component) and calculate Cronbach's alpha as a measure of internal consistency. Further, each item or component may be

analysed separately in terms of reliability. This technique is preferable in this context to calculating inter-rater and intra-rater reliability, as the assessment is the judgement of the class teacher of a student's ability and work over a whole year. Reproducing exactly the same experience of each student with more than one teacher is not practically feasible, and hence the assessment given by more than one person potentially subject to invalidity. This would appear to be a strange argument for another reason: traditional approaches to reliability of judgement rely totally on the ability of two or more persons agreeing in their assessment (see Krzanowski and Woods, 1984), or correlating teacher assessments with some criterion test.[3] But with teacher assessment of her own students in the intimacy of the classroom over an extended period of time, it is suggested that the teacher possesses an understanding and depth of insight into the students' abilities which cannot be easily replicated.

Secondly, what little research has been done into the nature of teacher assessment has suggested that sources of unreliability stem from the effects of sex and age (Jasman, 1987). In the case of this sample all students were of the same age, and so this factor cannot be assumed to be a source of unreliability. However, it was decided to investigate the effect of sex through the use of Analysis of Variance. This allows the researcher to look at the potential effect of the sex of the student, the sex of the teacher, and any possible interaction effect between sex of the teacher and the sex of the student (do male teachers tend to favour boys or girls in assessment, and the same for female teachers). Of course, there are many variables which can affect scores (see Bachman, 1990:119 for examples), but it is often difficult to isolate all potentially confounding factors, let alone design a study which will take them into account. It is for the reader to judge whether or not other variables which have not been considered may be so important as to render this study of marginal value.

If, given the above assumptions, assessment is reliable, one would expect to see a high Cronbach alpha, reasonable item reliability statistics, and no effect of sex on scores. The investigation into reliability is presented in Section 5.

4.1. Principal Components Analysis

Once again, as a first step, Pearson correlation coefficients were calculated for all data (Table 5). From these figures attention should initially be drawn to the relationship between teacher assessment of writing and spelling with the external examination results on the GCE. This may lead us to think that certain aspects of teacher assessment do relate to formal examination results, but that other components tap other aspects of student ability, as suggested by the exploratory study.

	GCE	MOCK	READ	WRITE	SPELL	GRAM	HOME
MOCK	.71						
READ	.37	.44					
WRITE	.61	.77	.44				
SPELL	.61	.69	.38	.70			
GRAM	.49	.55	.24	.57	.60		
HOME	.42	.58	.23	.62	.54	.51	
FCE	.66	.64	.41	.57	.60	.43	.46

Table 5: Correlation matrix of assessments (Study 2)

Turning to the Principal Components Analysis, it was decided to retain a solution with three components. In the exploratory study three components were retained on the criterion that the Eigen value for each component should be greater than 1, but in this solution it was necessary to discover if the factor pattern was similar, even though listening was not assessed. It transpired, nevertheless, that the Eigen values associated with each component were indeed significant. Table 6 provides the results of the analysis.

	Factor 1	Factor 2	Factor 3
GCE	.285	.127	.854
MOCK	.555	.284	.632
READ	.119	.954	.222
WRITE	.675	.342	.463
SPELL	.615	.207	.544
GRAM	.763	.015	.292
HOME	.844	.091	.168
FCE	.250	.199	.821
Variance explained by rotated components (Eigen values)			
	2.593	1.204	2.476
Percent of total variance explained			
	32.414	15.045	30.947

Table 6: Rotating loadings on Factors from the PCA (Study 2)

McNemar (1951, quoted in Child, 1970:12) was among those who were highly critical of Factor Analysis and Principal Components Analysis, arguing with a large degree of fairness that 'when interpreting factors all factorists struggle and struggle and struggle in trying to fit the factors to their initial hypotheses.' Hence my initial comment that this approach to replication of results is not entirely satisfactory.

This caveat to one side, a fairly clear factor pattern does emerge from this study. Firstly, all formal examinations load more highly on Factor 3 than do any other forms of assessment. However, the teacher assessment of spelling and writing do load to some extent on this factor also, confirming the initial examination of the correlation matrix. The teacher assessment of reading loads highly on Factor 2 and on no other factors. An interpretation of this will be offered shortly. All forms of teacher assessment load most highly on Factor 1, followed by MOCK, which was internally assessed. No aural/oral factor emerged, but this is not surprising as the only component containing those elements was the FCE which, once again, was represented by a global grade including other skills.

This clear factor pattern removes some of the hesitation which the researcher may have in interpreting the results given the problems which exist in factor interpretation, and we may conclude that the results of the Principal Components Analysis do generally confirm the exploratory hypothesis that teacher assessment validly taps aspects of student abilities and performance to which external examinations are not sensitive, with the important rider that some aspects of teacher assessment do coincide with external examination results – in this case spelling and writing, as the GCE external examination does in fact place 70% of all marks on writing abilities.

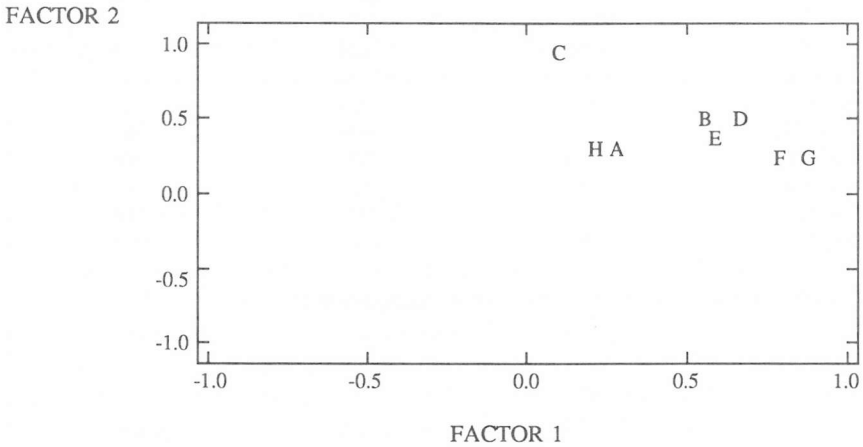

Figure 4: Factor plot for Factors 1 and 2 (Study 2).

Factor Plots are once again provided to allow the reader to conceptualise the relationships between measures on the three factors retained in the solution (Figures 4 - 6).

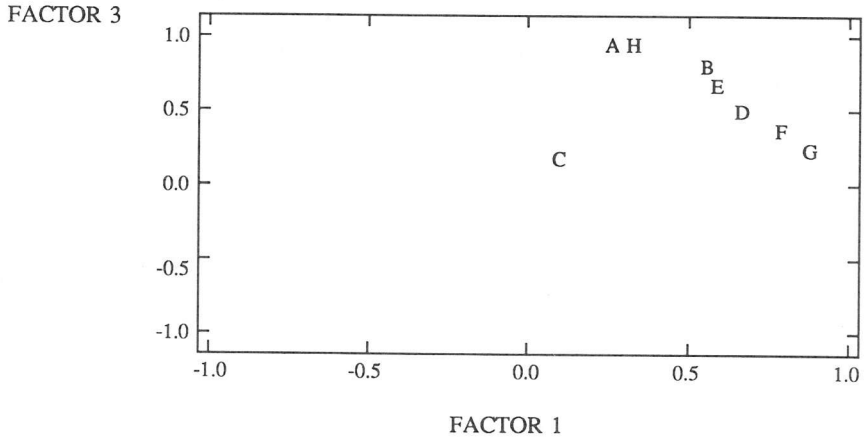

Figure 5: Factor plot for Factors 1 and 3 (Study 2).

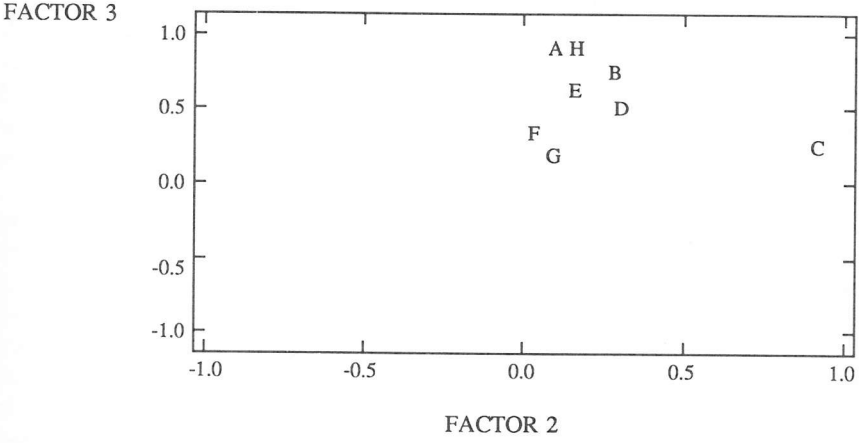

Figure 6: Factor plot for Factors 2 and 3 (Study 2).

4.2. Multiple regression

Each component of teacher assessment was regressed on the GCE score with the hypothesis that prediction from teacher assessment to external examination results would not be extremely high. The results of this part of the study are in fact extremely interesting, in that they confirm the conclusions drawn from the Principal Components Analysis. The results are presented in Table 7.

Variable	Coefficient	Std Error	Beta	Tolerance	T	P
READ	.182	.517	.112	.784	1.420	.093
WRITE	.419	.165	.278	.408	2.540	.012
SPELL	.432	.181	.244	.464	2.380	.019
GRAM	.200	.165	.112	.566	1.212	.228
HOME	.102	.128	.074	.537	.802	.424

N: 122 Multiple R: .660 Squared Multiple R: .435
Adjusted Squared Multiple R: .411 Standard Error of Estimate: 1.038

Table 7: Multiple regression of teacher assessment on GCE scores

It will be noted that although the multiple correlation coefficient is significant at .66 the variance shared with the external examination result is only .43, hardly large enough for accurate prediction of examination results. The adjusted squared multiple correlation is what may be expected in terms of prediction for any future sample of students drawn from the same population, and this is a lower .41. This tends to confirm the hypothesis, once again, that teacher assessment is offering something of value and validity in addition to the external examination results.

The final column in the table (P) provides the degree of significance of the correlation between the predictor and the examination results. Only writing and spelling are significant at $P < .05$, whilst the assessment of the amount and quality of work done outside the classroom is least significant in predicting results. We may be fairly confident of these findings as the Tolerance figures for all of the predictor variables do not approach zero, indicating that each of the independent variables is not so highly correlated with any other variable that we would suspect problems of collinearity: each of the variables does tap something unique as well as the variance it shares with other variables.

In 4.1 and 4.2 similar conclusions may be drawn, even though two separate techniques have been used. Teacher assessment of students does add something important to the overall assessment of students in addition to information provided by external examinations results.

5.1. Reliability

Teacher assessment may seem to be valid, but it cannot be valid if it is not reliable (Stevenson, 1981:47; Fulcher, 1987a:289; Bachman, 1990:160). Reliable assessment is a necessary but not sufficient condition for valid assessment. Each of the components of teacher assessment was treated as a single test item with a possible score of 1 to 5, with the total score possible on all items being 25. The reliability of teacher assessment as a global concept may then be assessed regarding reliability, and each component analyzed in terms of item reliability. The results of this investigation are presented in Tables 8 and 9.

Method of Calculation	Reliability Coefficient
Split-Half Correlation	.763
Spearman Brown Coefficient	.866
Guttman (Rulon) Coefficient	.846
Cronbach's Alpha Coefficient	.815

Table 8: Internal Consistency Data on Teacher Assessment as a single test

Label	Mean	Standard Deviation	Item-Total R	Reliability Index	Excluding this Item	
					R	ALPHA
READ	3.131	.829	.591	.490	.380	.840
WRITE	2.877	.829	.868	.774	.765	.726
SPELL	2.943	.761	.821	.624	.714	.749
GRAM	3.303	.756	.751	.568	.615	.777
HOME	3.336	.972	.770	.749	.591	.786

Table 9: Item Reliability Statistics,
with Teacher Assessment Categories treated as individual items

Cronbach's alpha for global teacher assessment is a very acceptable .815, surprisingly high in fact for the nature of the assessment process. Given their detailed knowledge of individual students, teachers do seem to be very consistent in the way they award grades.

In Table 9 we are primarily interested in the final column. This tells us what Cronbach's alpha would have been if this component of the assessment had been omitted. In each case, with the exception of the assessment of reading, the alpha coefficient would have been lower, indicating that total assessment would have been less reliable. Had the assessment of reading been omitted, however, coefficient alpha would have been a higher .84.

The assessment of reading ability by teachers (from those abilities investigated here) appears to be a problem area. This is confirmed by the correlation coefficient between this component and the total grade (.59) and the component reliability figure (.49). This may now be used to throw light on the fact that reading alone loaded on Factor 2 in the Principal Components Analysis. The assessment of reading stands alone because its assessment is essentially unstable. Why should this be so? Reading ability, unlike the other components, is an aspect of proficiency which cannot be directly observed. The classroom teacher may only *infer* the quality of reading ability by other means, such as the success with which the students handle comprehension exercises. It may be hypothesised that the same may be found with the assessment of listening ability. The teacher would need to be 'inside the head' of the student to be able to observe what is actually taking place in the reading process. It may be tentatively concluded that reading ability is best tested through a more formal reading test, while those aspects of ability which have more direct behavioral manifestations are reliably open to teacher assessment.

5.2. The effect of sex as potential bias in teacher assessment.

Using the sex of the student and the sex of the teacher as categorial variables, an Analysis of Variance was performed on each of the components of Teacher Assessment. The results were then analyzed for bias in assessment as a result of the sex of the student, the teacher, and any possible interaction effect. The results are presented in Tables 10 to 14. It is important to notice, when reading Tables 10 to 15, that the ERROR component is extremely large. Usually this is undesirable in an Analysis of Variance, but here it is both expected and desirable. As the Analysis of Variance is attempting to see how much variance is attributable to test bias (measurement error) the residual error is the amount of variance which is attributable to true score once the error stemming from sex bias is removed. A large residual error variance thus indicates that sex bias does not unduly influence measurement negatively.

Source	Sum of Squares	DF	Mean-Square	f-ratio	P
SEXS	.8391	1	.839	1.200	.276
SEXT	.005	1	.005	.008	.931
SEXS*SEXT	.005	1	.005	.008	.931
ERROR	82.566	118	.700		

Table 10: Analysis of Variance using READ as the dependent variable, and the sex of the teacher (SEXT) and sex of the student (SEXS) as categorial variables.

Source	Sum of Squares	DF	Mean-Square	f-ratio	P
SEXS	1.029	1	1.029	1.583	.211
SEXT	14.659	1	14.659	22.554	.000
SEXS*SEXT	0.127	1	.127	.195	.659
ERROR	76.694	118	.650		

Table 11: Analysis of Variance using WRITE as the dependent variable, and the sex of the teacher (SEXT) and sex of the student (SEXS) as categorial variables.

Source	Sum of Squares	DF	Mean-Square	f-ratio	P
SEXS	.224	1	.224	.390	.533
SEXT	2.115	1	2.115	3.690	.057
SEXS*SEXT	.003	1	.003	.005	.943
ERROR	67.630	118	.573		

Table 12: Analysis of Variance using SPELL as the dependent variable, and the sex of the teacher (SEXT) and sex of the student (SEXS) as categorial variables.

Source	Sum of Squares	DF	Mean-Square	f-ratio	P
SEXS	.292	1	.292	.565	.454
SEXT	6.065	1	6.065	11.730	.001
SEXS*SEXT	.167	1	.167	.323	.571
ERROR	61.012	118	.517		

Table 13: Analysis of Variance using GRAM as the dependent variable, and the sex of the teacher (SEXT) and sex of the student (SEXS) as categorial variables.

Source	Sum of Squares	DF	Mean-Square	f-ratio	P
SEXS	2.002	1	2.002	2.330	.130
SEXT	8.691	1	8.691	10.114	.002
SEXS*SEXT	.028	1	.028	.032	.858
ERROR	101.394	118	.859		

Table 14: Analysis of Variance using HOME as the dependent variable, and the sex of the teacher (SEXT) and sex of the student (SEXS) as categorial variables.

With the exception of reading, where there is no bias whatsoever, but which we have seen is an element of potential unreliability in assessment, the results are relatively easy to interpret. In no case does the sex of the student influence the teachers' assessments, nor is there any interaction effect between the sex of the student and the sex of the teacher. Thus, there is no evidence to suggest, for example, that male teachers treat female students preferably, nor is there any evidence to suggest any other combination of preferences which would constitute measurement error. However, in the cases of writing, grammar and homework there is an effect from the sex of the teacher. A post-hoc Tukey HSD test (not presented here as it is not of central relevance) confirms that male teachers awarded higher assessments than did female teachers in these areas.

In the case of writing this bias may only be apparent. As we have seen, the writing variable is the most accurate predictor of external examination results in the multiple regression. An Analysis of Variance using GCE results as the dependent variable and the sex of the teacher as the independent variable (Table 15) shows that the students of male teachers in this sample scored significantly higher than other students. As such, this apparent bias may merely

be an artefact of this sample in which male teachers taught more proficient students and hence gave higher internal assessment grades to their students.

Source	Sum of Squares	DF	Mean-Square	f-ratio	P
SEXS	4.460	1	4.460	2.775	.098
SEXT	18.505	1	18.505	11.513	.001
SEXS*SEXT	2.118	1	2.118	1.361	.246
ERROR	189.666	118	1.607		

Table 15: Analysis of Variance using GCE as the dependent variable, and the sex of the teacher (SEXT) and sex of the student (SEXS) as categorial variables.

In the case of the assessment of the quality and quantity of homework the situation would seem to be much more clear cut. Male teachers are simply more lenient than female teachers.

6. Conclusions

From the two studies reported here, it would not seem unjust to conclude that teacher assessment within the school and classroom setting is valid in that it taps aspects of student abilities to which formal examinations are not sensitive. The experience the classroom teacher has of her students in the learning process should be taken into account. The classroom teacher can carry out the process of assessment reliably, although caution must be recommended in skill areas which are not more directly observable, as is the case with reading ability. Bias due to sex would not appear to be of any great concern at least as far as this sample is concerned.

The questions posed at the beginning of this article have been answered to some satisfactory degree for the specific situation to which they were relevant. The same questions will be relevant to other situations, but the population of teachers and students different. To the extent that they differ this kind of research must be carried out again to examine the issues of reliability and validity.

On the basis of this study, however, it would be recommended that EFL examinations should develop along the principle that externally set papers should be retained, but that continuous assessment by the teacher make up part of the final score or scores (if a profile reporting system is adopted) recorded on the certificate which the student receives. The UCLES IGCSE examination has

begun this process by allowing some limited degree of internal teacher assessment. More experimentation along these lines would be beneficial in the interests of valid assessment.

This article has dealt exclusively with issues of reliability and validity. At the end of the day, however, practical considerations must be taken into account. And it is with these that we conclude. Achieving acceptable levels of reliability and validity in teacher assessment of students depends completely on each individual institution employing a fully professional and highly trained teaching staff. There are institutions in many countries which, given the opportunity of raising their students' grades, would jump at the chance in order to better sell the commercial product which they offer. Teacher assessment is not for them. The Examination Boards must be careful to vet thoroughly each application to run internal assessment schemes. Any mistake would open the abyss into which reliable and valid assessment would disappear. This cannot be emphasised too strongly. If safeguards against those who would exploit the system do not work, then the Examination Board itself would surely end up with a severely damaged reputation among respectable institutions, which could not easily be repaired. In everything which the Examination Boards do, they must observe the ethical rules of the profession as laid down in the *Standards for Educational and Psychological Testing* (American Psychological Association, 1985). Some consequences of not doing so are dealt with in Fulcher (1987b). Secondly, the institutions themselves must make resources available for in-service teacher training. Even with a highly professional staff this will increase levels of reliability. Thirdly, each institution operating internal assessment should have at least one member of staff who is a qualified measurement expert, capable of monitoring the assessment which takes place.

As for the teachers themselves, much groundwork needs to be covered. It needs to be ensured that the amount of extra work required in recording internal assessments does not place too high a burden upon them, given their normal teaching loads. If it does, it will be resented. Given the knowledge that 'their grades count' some teachers may be fearful of giving assessments which are too high or too low. These fears must be overcome in the interests of the students themselves. The management of change is important should a school decide that it wishes to operate internal assessment.

Teacher assessment will undoubtedly become an issue in the world of EFL teaching within the next ten years even though it is not furiously knocking on our door at the present time. The more EFL professionals can discover about it now the better placed we will be to cope with it when the need arises.

Notes

1. For readers unfamiliar with the statistical processes used in this article, and the issues regarding Factor Analysis in particular, see Woods (1983), and Woods, Fletcher and Hughes (1986). A very clear explanation of the conceptual background to the interpretation of Factor loadings may be found in Burroughs (1975:274-279), and a more wide ranging discussion is provided by Child (1970). For detailed information on all the techniques used in this study see the excellent work by Crocker and Algina (1986), and the classic work of Cronbach (1984). Henning (1987) also provides much useful background information. Simpler introductory material may be found in Hatch and Farhady (1982), Butler (1985), Isaach and Michael (1981), and Ferguson (1981). For language teachers with no background in measurement theory, introductory texts such as Hughes (1989) or Baker (1989) may be consulted. Selliger and Shohamy (1989) provide a very clear introduction to basic research techniques in the field of foreign language learning and assessment.
2. The problem with labelling Factor 2 'Teacher Assessment' in the exploratory study is that it is not a skill or ability (or constellation of skills or abilities) possessed by students, as are Factor 1 and Factor 3. This essentially means that whilst teacher assessment is different in kind from other measures, we are unable to say from the first study what teachers are taking into account when they assess. This problem is overcome in the design of the second study.
3. Recent research in the United States reported by Levine and Haus (1987) suggested that assessments of students made by teachers of French and Spanish differed significantly from their scores on standardised ratings of oral proficiency. The recommendations made in their study were that more rigorous teacher training should be introduced so that teacher assessments and standardised test results would coincide. Apart from any other consideration, if the two techniques provided the same information (their definition of reliability) then one technique would be redundant by definition. However, it should be noted that the results of the studies presented in this paper would suggest that the difference noticed by Levine and Haus may be important in its own right and not just the result of unreliable teacher assessment. Levine and Haus should, perhaps, begin to question the nature of that which they purport to be investigating rather than assuming that reliability and validity of teacher assessment can be judged solely on an external criterion using correlational studies.

Bibliography

Alderson, J.C., Krahnke, K.J. and Stansfield. 1987. *Reviews of English Language Proficiency Tests*. Washington, D.C.: Teachers of English to Speakers of Other Languages.

American Psychological Association. 1985. *Standards for Educational and Psychological Testing*. Washington, D.C.: APA.

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

Baker, D. 1989. *Language Testing: A Critical Survey and Practical Guide*. London: Edward Arnold.

British Council. 1989. *Introduction to the IELTS*. London: The British Council.

Burroughs, G.E.R. 1975. *Design and Analysis in Educational Research*. University of Birmingham Faculty of Education: Educational Monograph No. 8.

Butler, C. 1985. *Statistics in Linguistics*. Oxford: Basil Blackwell.
Child, D. 1970. *The Essentials of Factor Analysis*. New York, etc.: Holt, Rinehart and Winston.
Criper, C. and Davies, A. 1988. *ELTS Validation Project Report*. The British Council/ University of Cambridge Local Examinations Syndicate.
Crocker, L. and Algina, J. 1986. *Introduction to Classical and Modern Test Theory*. New York, etc.: Holt, Rinehart and Winston.
Cronbach, L.J. 1984. *Essentials of Psychological Testing*. New York: Harper and Row.
Ferguson, G.A. 1981. *Statistical Analysis in Psychology and Education*. Singapore: McGraw-Hill Book Company.
Fulcher, G. 1987a. 'Tests of oral performance: The need for data-based criteria', *English Language Teaching Journal* 41/4, pp. 287-291.
Fulcher, G. 1987b. 'Measurement or assessment: A fundamental dichotomy and its educational implications', *Education Today* 37/2, pp. 60-65.
Hatch, E. and Farhady, F. 1982. *Research Design and Statistics for Applied Linguistics*. Rowley, Mass: Newbury House.
Henning, G. 1987. *A Guide to Language Testing - Development - Evaluation - Research*. Rowley, Mass: Newbury House.
Hughes, A., Porter, D. and Weir, C. 1988. *Discussion on the ELTS Validation Report*. The British Council/University of Cambridge Local Examinations Syndicate.
Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.
Issac, S. and Michael, W. B. 1981. *Handbook in Research and Evaluation for Educational and the Behavioral Sciences*. San Diego, California: EdITS Publishers.
Jasman, M.A. 1987. *Teacher-based Assessments: A Study of Development, Validity and Reliability of Teachers' Assessments and Associated Structured Activities Devised to Assess Aspects of the Primary Curriculum for the Age Range 8 - 12 Years and the Evaluation of In-service Provision to Facilitate such Teacher-based Assessments*. University of Leicester: Unpublished Ph. D. thesis.
Krzanowski, W.K. and Woods, A.J. 1984. 'Statistical aspects of reliability in language testing', *Language Testing*, 1/1, pp. 1-20.
Leveine, M.G. and Haus, G.J. 1987. 'The accuracy of teacher judgement of the oral proficiency of high school foreign language students', *Foreign Language Annals*, 20/1, pp. 45-50.
McNemar, Q. 1951. 'The factors in factoring behaviour', *Psychometrica* 16, pp. 353-359.
Oller, J.W. and Perkins, K. 1980. *Research in Language Testing*. Rowley, Mass: Newbury House.
Selliger, H.W. and Shohamy, E. 1989. *Second Language Research Methods*. Oxford: Oxford University Press.
Stevenson, D.K. 1981. 'Beyond faith and face validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests', in Palmer, A.S., Groot, P.M.J. and Trosper, G.A. (eds), *The Construct Validation of Tests of Communicative Competence*. Washington, D.C.: Teachers of English to Speakers of Other Languages, pp. 37-61.
Woods, A. 1983. 'Principal components and factor analysis in the investigation of the structure of language proficiency', in Hughes, A. and Porter, D. (eds), *Current Developments in Language Testing*. London: Academic Press, pp. 43-52.
Woods, A., Fletcher, P. and Hughes, A. 1986. *Statistics in Language Studies*. Cambridge: Cambridge University Press.