Are Raters' Judgements of Language Teacher Effectiveness Wholly Language Based?

Catherine Elder

1. Introduction

The influence of factors other than language in performance-based assessment has been acknowledged by Jones (1985), Upshur (1979), McNamara (1990), Wesche (1992) and others. While some writers take the Hymesian view that these non-linguistic factors (sensitivity to audience, interactive skill, personal style etc.) are part and parcel of communicative competence, others see them as beyond the scope of language testing or a source of what Messick (1992) describes as 'construct-irrelevant variance'. In discussing this issue McNamara(1990) makes an interesting distinction between 'strong' performance tests, in which test tasks are the target of the assessment with language being treated as a necessary but insufficient condition for their successful execution, and 'weak' performance tests, in which language proficiency is assessed independently of other factors involved in test performance and tasks serve merely as vehicles for eliciting a relevant language sample. An example of a 'strong' performance test is the British PLAB Test for doctors (Alderson et al. 1986) in which language skills are assessed in conjunction with clinical competence. An example of a 'weak' test is the Australian Occupational English Test (McNamara 1990) in which candidates perform mock medical role plays and are assessed for linguistic rather than professional expertise. In practice, however, this weak/strong distinction may not always be clear cut and the adoption of one or other approach to performance testing may depend less upon principled decisions made at the design stage than upon the particular orientation of the raters involved in the assessment process. Brown (1993) considers the effect of involving members of the tourist industry alongside language teacher experts in assessing oral skills on a test of Japanese for tour guides. She finds that teachers are harsher than industry raters in their assessment of grammar, but that the industry group are more demanding than language experts in assessing task fulfilment in areas which they perceive to be particularly crucial for professional effectiveness. The influence of rater background is also dealt with in a study by Elder (1993) on classroom-based

Melbourne Papers in Language Testing 1994 Volume 3.2 pp. 41-61. The Language Testing Research Centre, The University of Melbourne.

assessment which shows that in assessing the English language proficiency of non-native teachers of mathematics and science, subject-specialists (i.e. maths/science method lecturers) come close to adopting a 'strong' approach to performance testing in that they are less concerned with language control than with aspects of classroom methodology. Language experts, on the other hand, take a 'weaker' view of task performance and focus more closely on what they are equipped to assess, and that is the quality of the language sample produced in the classroom context.

While in the two tests mentioned above a balance of language and other key factors determining communicative effectiveness in the relevant occupational area can be achieved by pairing the two types of rater, this dual perspective is built in to the assessment process when the raters involved have both linguistic and occupational expertise. For example, in proficiency tests for second/foreign language teachers, raters are typically trained teachers who have a good knowledge of the target language as well as substantial professional experience. Both areas of expertise can have a bearing on their assessments of task performance. Comments such as "she made quite a few mistakes, but she's got what it takes to be a teacher" are not uncommon amongst raters. These comments suggest that, even in the unclassroom-like environment of a test, raters, rightly or wrongly, are prone to passing judgement on aspects of professional competence.

Although performance-based testing of teacher competence is widespread (see for example Bailey 1985, Briggs 1986, Hinofotis et al. 1981, Viete 1993) few studies have considered the influence of occupation-specific factors on the assessment of test performance. In this paper we look at the extent to which these factors are separable from linguistic ones and explore the feasibility of assessing candidates on both dimensions of performance concurrently.

2. Context for of the study

The context for the study is an Italian language proficiency test, developed at the NLLIA Language Testing Centre for the purpose of determining whether teachers who lack the requisite foreign language qualifications (in most cases a post-Year 12 major sequence of study in the relevant language) have adequate skills to perform

their professional role effectively. The test is located towards the 'strong' end of the performance test continuum in that, somewhat ambitiously, it invites judgements from raters about both the linguistic and 'teacherly' qualities of candidate performance.

3. Test format

While the test measures proficiency in all four skills, it is the performance-based speaking component which will be discussed here. Six of the seven phases of the speaking test are classroom-specific i.e. they are designed to simulate teaching tasks typically required of teachers in the second language classroom. The identification of suitable tasks was based on a needs analysis conducted in a number of primary schools where Italian was both medium and object of instruction. The content validation process is described in some detail in another paper (Elder 1994). The test takes the form of a face-to-face interview between the candidate and a trained interlocutor. The pilot version, which is the basis for this study, has seven phases and lasts 30 minutes. The first phase is warm-up conversation and is not assessed. From Phase 2 onwards candidates are required to simulate the role of a foreign language teacher. In Phase 2 candidates are asked to read aloud a children's story; in Phase 3 they retell the same story in their own words, as if to a group of children; in Phase 4 they give instructions to a class about how to make a model or to play a game; in Phase 5 they set up and perform a role play, assigning one of the roles to their interlocutor (as they might do with a classroom learner), in Phase 6 they give a brief presentation on a culture-related topic and in Phase 7 they identify and offer explanations for errors in a text produced by a school-age L2 learner.

4. Assessment criteria

Assessment criteria (see Appendix A), which were developed in consultation with language teacher experts, are of two types. First, the linguistic criteria, which are applied task by task, assess pronunciation, grammatical accuracy, resources of expression, fluency and comprehension. There is also a metalanguage category which relates to Phase 7 of the interview in which learner errors are explained. Assessments for each of the above categories are made at least once, and in most cases twice, during the course of the interview. Descriptions of performance at six levels of ability are

provided for each rating category. Classroom competence criteria, on the other hand, invite judgements about the 'teacherliness' of task performance or, in other words, its suitability for the classroom, since this was felt by teacher experts to be a key consideration in making decisions about teacher readiness. To assist with these classroom competence assessments raters are provided with a checklist of points to consider such as "Was the style and tone of delivery appropriate for the classroom?" "Did the candidate tailor her language in such as way as to make it intelligible to second language learners?" "Were instructions issued in a clear and convincing manner?" These judgements are thus concerned with aspects of candidates' pragmatic competence. Classroom competence assessments are made three times during the test and ratings are recorded on a four-point scale. At the end of the speaking test two further holistic assessments are elicited: one for global language proficiency, which is a summation of the various linguistic judgments, and another for overall level of performance in which raters are required to evaluate the whole performance in relation to the requirements of teaching. These final assessments are recorded on a sliding scale defined at four points. The variation between 6 point, 4 point and sliding scales is aimed at reducing the likelihood of a halo effect which could obscure distinctions between the different assessment categories.

5. The trial population

The pilot version of the speaking test was trialled on a sample of 75 subjects including 5 native speakers who had acquired their Italian outside Australia, 42 undergraduate Italian language students in the second or third year of post-secondary study and 28 Diploma of Education students, who had undertaken variable amounts of formal study in Italian and were training as Italian language teachers. The majority of the trial subjects were second- or third-generation immigrants of Italian language background.

6. The assessment process

Each candidate's test performance was videoed for retrospective assessment. After a preliminary training session the videotapes were distributed amongst 15 raters all but two of whom were trained and experienced Italian language teachers with native or near-native proficiency in the target language. Each tape was assessed at

least twice (by both a native and a non-native speaker) and 10 of the tapes were assessed by all fifteen raters following an initial briefing session.

7. Research questions

Drawing on data derived from these raters' assessments answers were sought the following questions:

1. Do the two kinds of assessment items (ie language and classroom competence) fit together to define a single measurement trait?

2. Are ability estimates based on the classroom competence criteria identical to those derived from linguistic rating categories?

3. Do the two sets of criteria together produce orderly (i.e. "fitting") measures of candidate ability?

Since the rating process has been set up to elicit judgements of both linguistic and classroom competence, the issue of whether these aspects of performance are indeed distinguishable from one another is an important one. The reliability of the ability estimates produced by combining the two sets of scores is also a matter of concern, since results on this test may determine whether or not candidates are accepted into the teaching profession.

8. Data analysis

Data was analysed with a multifaceted Rasch programme, Facets (Linacre 1989, 1990 which has the capacity to model and adjust for the variability which occurs in different aspects of the test situation. In this study there were three facets in the data matrix - the candidates, the raters and the items (assessment criteria).

The Rasch analysis allows the hypothesis of unidimensionality which was posited in the research questions above, to be tested in relation to the data. For comparison purposes three data runs were performed with this programme: the first incorporating scores on all test items, the second involving ratings assigned against the classroom competence criteria, and the third consisting of ratings for linguistic competence (whether global or analytic). The overall teacher readiness category, on which final decisions about teacher

readiness are based, was the only common element in the three sets of data. Correlational statistics were used to supplement the findings of the Rasch analysis.

9. Results

9.1. Item fit

Tables 1, 2 and 3 below show the fit statistics for test items derived from each of the three analyses referred to above. Items are listed in order of their occurrence on the test. The fit statistics indicate the probability of particular pattern of responses to an item given an assumption of unidimensionality in the data. Extreme negative values (ie with a standardised infit meansquare of -2 or more) indicate that the item accords too closely with the measurement model and extreme positive values (of 2 or above) indicate that the item does not fit within the measurement dimension defined by the other items on the test.

Item/rating category	Infit MnSq	Infit std
Fluency 1	1.1	0
Pronunciation	0.8-	-1
Resources 1	0.8	-1
Accuracy 1	0.6-	-2
Classroom Competence 1	1.3	2
Comprehension 1	0.6-	-2
Accuracy 2	0.6-	-2
Classroom Competence 2	1.3	2
Fluency 2	0.8	-1
Resources of Expression 3	0.7-	-2
Metalanguage	1.5	3
Classroom Competence 3	1.2	1
Comprehension	1.0	0
Global language proficiency	1.2	1
Overall level of performance	1.2	1

Table 1: Fit statistics for all rating categories on the test

nb. Light shading indicates underfit, dark shading indicates overfit.

The figures in Table 1, which are based on scores assigned on all test items, show that there are two overfitting items (with extreme negative values): accuracy and resources of expression. This is a sign that these more traditional components of language proficiency account for a large portion of the overall ability measure. The predominant influence of grammar and linguistic resources on assessments of speaking ability is commonly reported in the language testing literature (see Wilds 1979, Raffaldini 1988, McNamara 1990) and is defensible on a test of this kind. There are certainly grounds for arguing that well-formedness and breadth of range are crucial aspects of foreign language teacher competence, given that the teacher may be the only source of target language input available to learners.

More relevant to this enquiry are the underfitting ratings which in Table 1 (marked with an asterisk) indicating, 'noise' or, in other words, the presence of factors which are not captured in the overall construct of ability defined by the analysis. The item which sits least comfortably with the others is Metalanguage. This is not surprising since the skills involved in giving explanations in Italian about grammar and phonology are cognitively complex and are likely to depend as much on formal knowledge as on communicative competence. The other two categories which do not fit with the rest of the data are Classroom Competence 1 and Classroom Competence 2. Both of these are designed to assess the 'teacherliness' of communication which may, thinking back to the points on the rater checklist, be more a matter of pragmatic competence than of actual linguistic control. As anticipated at the outset, there is a distinction between what is being measured on linguistic and classroom competence criteria in early ratings of performance on the test. This is not however the case with the third classroom competence category which fits with all the other (linguistic) items. It may be that, at this late stage of the interview, raters are aggregating previous impressions and hence conflating linguistic and 'teacherly' considerations in their assessment of this item.

A clearer picture is obtained by removing linguistic items from the analysis (see Table 2 below). Ratings for each of the classroom competence categories (including the third one) now appear to fit perfectly with one another (standardized infit of 0). This confirms what was suggested above and that is, that the third classroom competence category incorporates both language and occupation-

related considerations, but that the latter become more salient when the whole set of classroom competence items are grouped together. The same appears to be true of the overall level of performance category according to which final determinations are made. Here it fits with the classroom competence data, whereas, in the previous analysis, it was aligned with the other linguistic rating categories. It can therefore be assumed that both linguistic and occupation-specific factors are contributing to this final assessment of teacher readiness.

Item/rating category	Infit MnSq	Infit std
Classroom Competence 1	1.0	0
Classroom Competence 2	1.0-	0
Classroom Competence 3	1.0	0
Overall level of performance	0.9	0

Table 2: Fit statistics for occupation-related categories

The fit statistics in Table 3 below show what is left after the classroom-oriented categories have been removed from the equation. The linguistic items (fluency accuracy etc.) are perfectly consistent with one another in terms of what they are measuring, whereas the global language proficiency category and the overall level of performance category are underfitting (i.e. they cannot be defined in terms of the same measurement trait). The 'whole' is clearly more than the sum of the linguistic parts. This provides further support for the notion that there are two factors which have a bearing on the assessment of teacher ability: on the one hand a narrowly-focussed linguistic factor and on the other a more general and diffuse classroom competence factor which may be language-related, but which discriminates amongst candidates in a different way.

Item/rating category	Infit MnSq	Infit Std
Fluency 1	1.2	1
Pronunciation	0.9-	0
Resources 1	0.8	-1
Accuracy 1	0.7	-2
Resources 2	0.8	-2
Accuracy 2	0.8	-1
Fluency 2	0.9	-0
Resources of Expression 3	0.8	-1
Comprehension	1.0	0
Global language proficiency	1.2	2
Overall level of performance	1.2	3

nb. Light shading indicates underfit, dark shading indicates overfit

Table 3: Fit statistics for linguistic categories

To supplement these findings a classical analysis involving stepwise regression was carried out with the same data set. The regression procedure was performed on the correlation matrix derived from the average of all raters scores on each test item (seeTable 4 below). Scores in the overall level performance category were treated as the dependent variable and scores in the other categories as the independent variables.

Parameter	Value	Std. Err.	Std. Val.	F to remove
Intercept	-.179			
Classroom Competence 3	.329	.081	.305	16.561**
Global Language	.897	.084	.796	112.85**

** p = < .01

Table 4: Stepwise regression analysis — Italian speaking test

From this analysis it emerges that it is global language proficiency which makes the most powerful contribution to the overall determinations about teacher readiness but that assessments made against the third classroom competence category also have a significant part to play. The fact that these steps have been selected over and above other variables entered in the equation is not surprising since, as we demonstrated in the previous analysis, both categories subsume features of both linguistic and classroom competence. The findings of the stepwise regression analysis can be taken as confirmation that both linguistic and occupation-related considerations are combining in their contribution to overall assessments of candidate ability.

9.2. Relationship between ability estimates

This question was investigated by correlating person ability measures derived from the linguistic items[1] and those based on classroom competence assessments. The ability estimates yielded by the Facets programme are expressed as logits or probability units which adjust for error in raw scores by taking into account both the harshness of the raters and the difficulty of test tasks. The two sets of logit values were compared using the Spearman correlation statistic and the results (rho = 0.73) show that while there is a significant relationship between the two sets of measures, it is not a particularly strong one. Classroom competence is clearly not synonymous with linguistic competence and the variance between the two sets of ability estimates must be explained by factors other than language. Because of these differences in what is being measured it would seem appropriate to treat the whole test as two separate entities and report each element of performance on a separate scale.

9.3. Estimates of person ability

The consequences for candidates of either combining or separating these somewhat disparate elements of performance will now be discussed. Table 5 below shows summary data from the candidate measurement report consisting of separation indices, which are

[1]Global language was not included as a linguistic item in this analysis, because as demonstrated earlier in the paper, it subsumes aspects of both linguistic and classroom competence.

measure of scalability and their corresponding reliability coefficients. Figures are derived from the whole data set and from the separate analyses of linguistic and classroom competence data respectively.

Data set	No of items	Separation Index	Reliability
Whole test	15	7.23	0.98
Linguistic competence	126.62	6.62	0.98
Classroom competence	3	2.59	0.87

Table 5: Test reliability indices

Separation indices and reliability coefficients derived from separate analyses of a) the whole test and b) the linguistic competence items on their own are high, showing that either version of the test discriminates effectively amongst candidates[2] The reliability index for classroom competence items is considerably lower but this may be a product of test length rather than an indication of poor item discriminability. There were only three classroom competence items in the set, which, it is acknowledged, is a defect in the research design. The scalability of scores derived from the classroom competence rating categories could presumably be boosted by increasing the number of assessments from three to six, with a rating being assigned after each task, rather than after every two tasks as was the case for the pilot version. Once such a revision has been made it would appear that there are minimal disadvantages for the majority of candidates in reporting classroom competence and linguistic assessments on two independent scales since each of them is likely to produce reliable estimates of ability.

Turning now to individual candidates, an analysis of person ability measures produced by the analysis reveals that there are several

[2]Separation indices indicated that at least 7 discrete levels of ability (in the case of the whole test) or 6 levels (in the case of the shorter test based on linguistic items only) can be identified.

ability estimates which do not fit with the overall pattern of responses. Table 6 below shows these misfitting estimates which emerge from the analysis of the whole data set (in the left hand column) as well as those yielded by a separate analysis of data runs involving linguistic and classroom competence rating categories respectively.

	Whole test Infit statistics		Linguistic items Infit statistics		CR competence items Infit statistics	
Cand no.	MnSq	Std.	MnSq	Std.	MnSq	Std.
7	0.6	-4				
10	2.4	4	2.4	3		
14					0.1	-2
16	2.0	2	1.7	2		
17	1.7	2				
18					0.1	-2
23	1.7	2				
27	1.8	2				
61			0.4	-2		
63	0.7	-3				
74					0.1	-2
85	0.4	-2	0.2	-2		
91					0.3	-2
95	0.4	-2				
Total no of misfitting candidates	9		3		4	

Table 6: Misfitting Estimates of Candidate Ability

When all items are grouped together there are nine misfitting candidates (ie approximately 12% of the sample). This is a clear indication that when performance on all items is aggregated, it is difficult to obtain orderly estimates of candidate ability for a disturbingly large number of candidates in the sample. On the other

hand, when scores for the two different sets of rating categories are analysed separately, the number of misfitting candidates is lower: there are three cases of misfit on the linguistic criteria and four on the classroom competence ones. Looking only at the underfitting estimates, which are more problematic[3], separation of the two scales leads to a reduction from five instances of mismeasurement on the whole test to one (on the linguistic rating categories). In other words, the probability of being candidates' being poorly measured by the test diminishes when classroom competence and language proficiency are treated as separate entities. It is also interesting to note that, while there is some commonality in the incidence of misfit on the linguistic items and on the whole test, there are no cases of misfit across all parts of the test.

In an attempt to identify the causes of these high numbers of misfitting ability estimates, the raw scores of those candidates with misfitting scores on the whole test, but not on its component parts, were scrutinized. Table 7 shows the raw scores assigned by one or more raters to candidats 7, 16, 17, 23, 63 and 95. These scores have been selected from the larger data set because of their peculiar characteristics. each set of figures has at least one shaded value - shaded because it is unpredictable in one of the following ways:

a) performance on the metalanguage category (which because of its high degreee of misfit was not included in either linguistic or classroom competence data sets) is rated much higher (see caniddates 7 & 16) than would be expected given the scores assigned for other rating categories;

b) performance on some or all of the classroom communicative competence categories is lower (see candidates 16,17, 23 and 63) than would be expected given all the other scores.

c) the overall level of perforance and/or global language scores are somewhat lower (candidates 7 and 16) than what would be expected from the average of other test items.

---

[3]Overfitting estimates simply indicate that candidates' performance is unusually consistent across rating categories, Underfitting estimates on the other hand are a sign that the candidate, as perceived by the rater, displays characteristics which are not typical of the cohort and which do not fit within the construct of ability measured by the test.

Candidate no.	**7**		**16**	**17**	**23**	**63**		**95**
Rater no.	**4**	**9**	**14**	**11**	**3**	**13**	**5**	**16**
Item 1 Fluency 1	-	4	4	3	5	3	4	2
Item 2 Pronunciation	5	4	4	3	5	4	4	3
Item 3 Resources 1	4	3	2	2	5	4	4	3
Item 4 Accuracy 1	4	3	2	3	4	4	4	2
Item 5 Classroom Competence	2	2	2	1	2	2	2	3
Item 6 Resources 2	4	3	2	3	5	4	4	2
Item 7 Accuracy	4	3	2	3	5	4	4	2
Item 8 Classroom Competence	2	2	1	2	2	2	3	2
Item 9 Fluency 2	4	2	2	3	5	4	4	2
Item 10 Resources of Expression 3	4	2	2	2	5	4	4	2
Item 11 Metalanguage	6	2	6	2	3	4	5	2
Item 12 Classroom Competence	3	2	2	1	1	3	2	2
Item 13 Comprehension	3	3	4	-	4	4	4	3
Item 14 Global language	4	1	1	5	4	4	4	1
Item 15 Overall teacher readiness	4	1	1	5	4	4	4	2

Table 7: Raw scores assigned by particular raters to 'misfitting' candidates

Only one of these aberrant scores (the metalanguage score produced for Candidate 16 by Rater 14) shows up in the 'misfitting responses' table produced by the Facets analysis. This table serves to identify those values which, given what is known about rater harshness on each item, are markedly uncharacteristic of particular raters. Apart from this one instance, therefore, it seems probable that these aberrant scores are not a product of random behaviour from

raters, but rather of 'deviant' behaviour from the candidates. The general mismatch between candidates' scores on classroom communicative competence criteria and those on the linguistic rating categories suggests that although both dimensions of performance are language based, they may sometimes be acting at cross purposes with one another. Candidate 63 for example is perceived by both raters to be linguistically adequate (with an average score of 4 on both analytic and global criteria) but unsatisfactory as far as classroom language behaviour is concerned (obtaining an average of between 2 and 2.5 on each classroom competence rating). There may be something about this candidate's style of delivery (eg excessive speed/complexity of utterances or lack of clarity/animation) which raters perceive to be inappropriate for the classroom. Furthermore, the discrepancies between the analytic and global raw scores of candidates 7 (rater 9), 11 and 16 suggest that there is something about their performance which cannot be described in holistic terms.

10. Summary of findings

The above analysis has shown that in assessing performance on an Italian oral proficiency test for foreign language teachers, which requires candidates to simulate the teacher role, raters make a definite separation between linguistic and occupation-related criteria. While these dimensions of performance have been shown to be related (since they are both based on language behaviour) there appear to be problems with combining the two sets of scores to produce a single estimate of ability. It also seems that linguistic and occupation-related criteria sometimes work in opposition to one another. The consequences are that over 10% of the candidates in the sample are mismeasured by the test instrument. On the other hand, when separate scaling of linguistic and occupation-related items is undertaken, the problem of inconsistent measurement is alleviated. For this reason it has been proposed that the test should be treated as two independent tests, with candidates' performance on each dimension reported separately.

11. Discussion

It was suggested above that there may be a conflict between the skills assessed on the classroom competence and the linguistic criteria. Feedback elicited from two of the raters immediately after

the rating exercise throws some light on this issue. The raters suggested that those candidates who took the teacher role simulation seriously and attempted to produce comprehensible input for an imaginary semi-proficient L2 audience placed themselves at a disadvantage in linguistic terms by deliberately simplifying their speech and slowing their rate of delivery. This points to what may be a fundamental incompatibility between the assessment of language proficiency, which assumes a developmental continuum involving an incremental increase in range and complexity of language use as proficiency progresses, and certain kinds of occupation-specific proficiency where certain features of pragmatic or strategic competence such as simplicity and clarity may be valued over and above elaborateness. In a test such as this one it is conceivable that native or near-native speakers, in an attempt to 'show off' their level of linguistic sophistication, may rate poorly on the classroom competence criteria because of an apparent insensitivity to the simplification strategies required in the classroom situation. Conversely, the less proficient speakers, who achieve low scores for language proficiency, may outperform native speakers on the classroom competence criteria precisely because they avoid complex elaborate forms. A practical solution to this problem would be to include on the revised version of the Italian test at least one task which does not require candidates to simulate classroom performance, so that any conflict between the desire to display linguistic sophistication and the need to demonstrate sensitivity to a classroom audience is resolved.

Implicit in the rater feedback alluded to above is the view that there are candidates who do not take their role simulation seriously, which raises the issue of test validity, so far unaddressed in this paper. In the presence of a highly proficient adult interlocutor and in the absence of second language learner audience it seems likely that most candidates will be unwilling or unable to display the kinds of linguistic and interactional adjustments which characterise genuine teacher talk to pupils. It remains unclear, therefore, what exactly is being measured by the classroom competence criteria and whether the so-called 'teacherly' qualities displayed in test performance will in fact serve the candidate in the foreign language classroom.

This issue needs to be resolved if the test's claims to specificity are to count as anything more than face validity. Retrospective

interviews with raters (using video taped samples of performance) might prove useful in defining exactly which aspects of candidates' behavour have earned them high or low marks on the classroom competence categories. The more precisely these criteria are defined the easier it will be to determine their relevance to the teacher role. In addition, a comparison between the test performance of competent Italian teachers with that of foreign language graduates with no prior LOTE teaching experience could serve to determine whether the classroom competence criteria are targetting relevant 'teacherly' skills. Better still, a study which tracks test candidates into the classroom and assesses their professional performance against the same set of rating categories as those used on the test itself might give some indication as to whether raters' perceptions of classroom competence in the test environment have anything to do with communicative behaviour in the corresponding real world situation. This kind of investigation is important not only for the measurement of teacher competence, but for performance tests generally, whose predictive validity claims are generally unproven.

Leaving aside this uncertainty about the validity of classroom competence ratings, one could nevertheless argue for their inclusion on the test simply because they encourage assessments of language to be made independently of other qualities of performance which might otherwise have a contaminating influence. (This paper has shown the risk for candidates when linguistic and classroom competence criteria are conflated). Overall determinations could be based on linguistic scores alone, with performance on the other categories reported, but not used in calculating the final ability estimate. In other words, the inclusion of occupation-related criteria, however dubious their validity in relation to the real world, may have the effect of producing more accurate judgments of language proficiency, thereby increasing test reliability. This issue warrants further investigation.

12. References

Alderson, J.C., C.N Candlin, C.M. Clapham, D.J. Martin and C.J.Weir *Language proficiency testing for migrant professionals: new directions for the Occupational English test.* A report submitted to the Council of Overseas Professional Qualifications: Institute for English Language Education University of Lancaster

Bailey, K.M. (1985) "If I had Known Then What I Know Now: Peformance Testing of Foreign Teaching Assistants" in Hauptman, C.R. Leblanc & M. Wesche (eds.) *Second Language Performance Testing*. University of Ottawa Press. Ottawa, Canada. 153-80

Briggs, S.L. (1986) *Report on FTA Evaluation* 1985-1986 English Language Testing Institute, Testing Division, University of Michigan, Ann Arbor, pp. 1-11.

Brown, A. (1994) *The effect of rater variables in the development of an occupation specific language performance test.* Language Testing

Elder, C (1993) 'How do subject specialists construe classroom language proficiency?' Language Testing

Elder, C. (1994) 'Performance testing as benchmark for foreign language teacher education' Babel: Journal of the Federation of Modern Language Teachers Associations 29,2:9-19

Jones, R. (1979) 'Performance Testing of Second Language Proficiency.' In Briere.E. and Hinofiotis (eds.) Concepts in Language Testing: Some Recent Studies. Washington D.C. TESOL 21121

Linacre, J.M. (1989) *Many-Facet Rasch Measurement.* Chicago: Mesa Press.

McNamara, T.F. (1990) *Assessing the Language Proficiency of Health Professionals.* PhD thesis. University of Melbourne.

Messick. S. (1992) *The interplay of Evidence and Consequences in the Validation of Performance Assessments.* Educational Testing Service, Princeton, New Jersey.July 1992.

Raffaldini, T. (1988) 'The use of situation tests as measures of communicative ability.' *Studies in Second Language Acquisition,* 10: 197-216.

Wesche, M.B. (1992) 'Performance testing for work-related second language assessment' in Shohamy, E. and R. Walton *Language Assessment for Feedback: Testing and Other Strategies.* Chapter 7. Washington, D.C.: National Foreign Language Center Publications.

Wilds, C.P. 'The oral interview test.' In Jones, R.J. and Spolsky, B. (eds) *Testing language proficiency*. Arlington, V.A: Centre for Applied Linguistics, 29-44.

13. Appendix

ITALIAN TEACHER TEST (PILOT VERSION): RATING SHEET

CANDIDATE'S NAME/NO ☐

	6	5	4	3	2	1
Phase 1 GENERAL CONVERSATION						
• Fluency	☐	☐	☐	☐	☐	☐
Phase 2A READING ALOUD						
• Pronunciation	☐	☐	☐	☐	☐	☐
Phase 2B STORY RETELLING						
• Resources of expression	☐	☐	☐	☐	☐	☐
• Grammatical accuracy	☐	☐	☐	☐	☐	☐

• Classroom competence (voice quality, pace,suitability for L2 learners)
(please circle appropriate term)Full/Acceptable /Limited /Unsuitable

	6	5	4	3	2	1
Phase 3 GIVING INSTRUCTIONS						
• Resources of expression	☐	☐	☐	☐	☐	☐
Phase 4 ASSIGNING & MODELLING A ROLEPLAY						
• Grammatical accuracy	☐	☐	☐	☐	☐	☐

• Classroom competence (clarity,animation, suitability for L2 learners)
(please circle appropriate term)Full/Acceptable /Limited /Unsuitable

	6	5	4	3	2	1
Phase 5 PRESENTING CULTURAL INFORMATION						
• Fluency(cohesion, hesitation etc.)	☐	☐	☐	☐	☐	☐
• Resources of expression	☐	☐	☐	☐	☐	☐
Phase 6 EXPLAINING ERRORS	6	5	4	3	2	1
• Metalanguage	☐	☐	☐	☐	☐	☐

• Classroom competence (clarity, suitability of content and language for an L2 learner)
(please circle appropriate term)Full/Acceptable /Limited /Unsuitable

	4	3	2	1
WHOLE TEST				
• Comprehension	☐	☐	☐	☐

FINAL ASSESSMENT

You should now enter your final definitive assessments below.

A) RATING FOR GLOBAL LANGUAGE ABILITY

Place a cross at any point on the line below. You should regard this rating as a summary of linguistic performance on the whole test. Your rating should take into account the assessments made on each of the linguistic criteria (ie comprehension, fluency, grammatical accuracy, grammatical knowledge, pronunciation, resources of expression).

native-speaker like advanced functional basic

B) CLASSROOM COMMUNICATIVE COMPETENCE

Transfer the assessments made against this criterion after the relevant phases of the test by writing the letter F(= Full) A (=Acceptable) L (= Limited) or U (= Unsuitable) in the spaces provided.

Phase 2A & 2B _____ Phase 3 &4 _____ Phase 5 & 6 _____

C) OVERALL LEVEL OF PERFORMANCE

You should indicate below how satisfactory you consider the candidate's overall level of performance to be in relation to the requirements of LOTE teaching by placing a cross at any point on the scale below.

highly satisfactory acceptable at risk unsatisfactory

Now write a brief comment about the features of this candidate's language or language related behaviour which most powerfully influenced your rating.

_____
_____
_____
_____

Assessor's name (please print)_____Date_____