How do subject specialists construe second language proficiency?[1]

Catherine Elder

1. Differing perceptions of L2 communication

Research on rater behaviour suggests that any reference to "the native speaker" as a criterion for assessment of second language proficiency needs careful qualification. Native speakers vary considerably and unpredictably in their perceptions of foreigner talk with respect to the dimensions along which they evaluate performance and the degree of consistency and tolerance they manifest in their judgements. The findings of Galloway (1977); Ludwig (1982); Barnwell (1989) and Hadden (1990) amongst others, offer evidence that language experts, whether they be teachers or trained language testers, view second language performance differently from other "linguistically naive" native speakers. The implication is that if as raters we are indeed concerned with gauging the impact of second language communication on the wider native speaker population, linguistic expertise may be a liability. There are however few language testing initiatives which give serious consideration to the possibility of using non-language experts as assessors.

This possibility is explored in the context of a research project conducted by the NLIA Language Testing Centre at the University of Melbourne.

2. The research project

The project involves the development of a classroom-based assessment procedure to monitor the English proficiency of graduates from non English-medium universities who are training to be teachers of maths and science in Australian secondary schools. It has arisen out of a concern that substantial numbers of non English-speaking background graduates entering teacher education courses

[1] A version of this paper was presented at the Language Testing Research Colloquium at the University of British Columbia, Vancouver, Canada in Feburary 1992.

Melbourne Papers in Language Testing 1992 Volume 1.1 pp.16-32. The Language Testing Research Centre, The University of Melbourne.

are unable either to function effectively during their school-based teaching practica or ultimately to perform credibly as teachers. Our assessment procedure offers a means for determining the extent to which their difficulties are related to language. It has both a screening and diagnostic purpose: it serves to identify those whose limited competence in English may place them at undue risk of failure in their studies and also to provide information about aspects of their performance which could be improved through supplementary English language support.

The particular pragmatic features and discourse structures of the classroom situation have been amply documented in the research literature (e.g. Sinclair & Brazil (1982); Stubbs (1983); Allwright (1980)). Because of the specificity of the context a performance test is arguably (see for example Bailey (1985); Briggs (1986); Jones (1979); Hinfotis et al, (1981)) the best way of getting to grips with the issues of measurement. Our assessment procedure takes the form of an observation schedule to be applied to the classroom performance of teacher trainees of non English-speaking background during their school-based teaching practica.

The schedule (see Appendix) is designed to be administered by the maths and science teachers and teacher-trainers who are normally involved in the assessment of teacher trainees' performance. This fact has placed practical constraints on its design: to be acceptable to subject-specialists it needs to be formulated in terms which are meaningful to non language experts and must be easy to administer. The current version of the schedule (which has undergone three substantial revisions) can be completed within a fifteen minute time-span, although repeated administrations are essential to ensure adequate sampling of candidates' performance throughout the practicum.

The schedule itemizes those features of language and language-related behaviour which the research literature revealed and our needs analysis confirmed to be crucial for effective classroom performance. These features serve as performance indicators and are grouped under six broad headings: "intelligibility", "fluency", "accuracy", "comprehension" (all of which can be regarded as components of general language proficiency) and "use of subject-specific language" and "use of the language of classroom interaction". (which are specific to the classroom context and fall

within the parameters of Bachman's (1990) definition of strategic and pragmatic competence. These categories function as criteria against which language performance is evaluated. In recognition of the fact that overall level of performance may be more than the sum of the parts, we have also included an "overall communicative effectiveness" criterion. Assessors are asked to produce seven ratings for each candidate by marking with a cross the appropriate point on the scales provided. The scales are defined at four points to distinguish between "highly satisfactory", "acceptable", "at risk" and "unsatisfactory" performance. Scores (from 0 - 8) can be derived by measuring the distance of the cross from the left hand end of the scale[2] .

3. Rater validity

The validity of using subject-specialists as assessors of language ability is, as suggested at the outset, open to question. The remainder of this paper will be dedicated to a consideration of this issue, investigated through the trialling process, which was set up so as to elicit judgments from two groups: language experts (in this case ESL teachers) on the one hand, and subject specialists (maths/science teachers/teacher trainers) on the other. We focus here on two research questions.

1. Do language experts differ from subject specialists in the way they construe classroom language proficiency?

2. Do these differences (if they indeed exist) jeopardize the reliability and validity of our assessment procedure?

4. Test trials

Trialling of our procedure has involved the application of the schedule to the viewing of a number of videoed segments of classroom interaction, or simulated classroom interaction, as well as to observations of actual performance in the classroom.

4.1 Video trials

[2] The decision to use scores rather than categorical ratings when tabulating data was determined by assessors' reluctance to assign their ratings to the defined points on the scale.

Since there are obvious practical constraints on the numbers of raters who can view a single teaching performance simultaneously, it was decided in the first instance to use video segments of classroom performance as a means of validating our procedure. Six videoed teaching segments (three maths and three science lessons each of approximately 8 minutes in length) conducted by teachers and teacher-trainees from a range of non-English speaking backgrounds were used as the basis for these trials. (The videos had been selected from a larger sample to represent performance at a range of proficiency levels.)

At the viewing session participants were asked to use our observation schedule to rate the various dimensions of communicative competence on the scales provided.

The level of agreement between groups with respect to both global and analytical scores was calculated with an intra-class correlation *(r_I)* statistic (Bartko, 1966).[3] Intragroup correlations are presented in Table 1 below.

Criteria	r_I
Intelligibility	0.92
Fluency	0.94
Accuracy	0.96
Comprehension	0.85
Subject-Specific language	0.73
Interaction language	0.96
Overall communication	0.87

Table 1: Inter-Group Reliability — Intra-class correlations between ratings assigned by ESL teachers and Subject-Specialist teachers/teacher trainers to observations of 6 videoed performances.

[3] This statistic is computed by applying a one-way analysis of variance to the data with each subject constituting a group. The intra-class correlation is derived from the F value with the following formula.

$$r_I = \frac{F-1}{F+m-1}$$

where m denotes raters. This statistic was chosen in favour of Pearson's r which is a measure of linearity rather than agreement.

These correlations, derived from a comparison between each group's mean ratings on the seven criteria, show that the extent of agreement between ESL and subject specialists is high on all but one criterion. No substantial claims can be made for this apparently high level of agreement since individual differences have been ironed out through the averaging process. On the other hand, some explanation needs to be offered for the lack of agreement between the two groups in their rating of "subject specific language use", a disagreement which remains in spite of the averaging of individual differences. The most obvious interpretation of this low correlation is that subject specialists and ESL teachers are interpreting this dimension of communicative competence differently. Douglas & Selinker (1990), in their comments on the trialling of Maths Speak, a test of the ability to talk about mathematics in English, have raised the possibility that trained second langage raters, in assessing candidates' presentations on mathematical topics, could assign high ratings for responses

> "which are well-pronounced, grammatically fluent and comprehensible but which are at the same time illogical, poorly organized and just plain wrong"(p.12)

The implication is that non-trained raters might be more concerned about the rightness or truthfulness of subject content. This was in fact evident in a comments from one of the subject specialist raters *"my judgements of his language ability are clouded by the way he presents the topic You just don't teach maths like that."* It is quite conceivable that in assessing use of subject specific language the ESL teachers are focusing on the lexis, grammar and internal cohesion of the presentation while the subject specialists are more concerned about the way in which subject content is conceptualized.

However it would be unwise to attach too much importance to this one low correlation (which may not in fact be significantly lower than those for the other dimensions). The suggestion that the two groups of raters behave differently because of different notions about what they are assessing needs to be tested on a larger number of subjects. We are in the process of collecting further video recordings for this purpose.

Of relevance to the practical question of whether subject specialists can be entrusted with the assessment of language are the intragroup

reliability figures for overall communicative effectiveness reported in Table 2 below.

	r_I
ESL teachers (n=7)	0.71
Subject specialists (n=8)	0.46

Table 2: Intragroup Reliability Indices — Intra-class correlations between ratings assigned by individual group members on the overall communicative effectiveness category

Although again it is not clear with such a small N-size that these intragroup reliability indices are significantly different from one another, the ESL teachers on the strength of this evidence appear to be the more closely aligned in their global assessments than are the subject specialists The variation amongst subject specialists may be an indication of a greater uncertainty among this group in assigning overall ratings, perhaps because of limited experience in assessing language performance. Similar findings were reported in Barnwell's study (1990) involving "linguistically naive" assessors of second language proficiency in Spanish.

While firm conclusions cannot be drawn from findings based on such a limited set of ratings, this lack of consistency amongst the subject specialists constitutes a threat to the validity of our using them as assessors. Further attention was therefore paid to this issue by looking at the level of inter- and intragroup agreement as to whether candidates' performance was either satisfactory or satisfactory. (A score of 4, the mid-point on our scale, was used as the cut-off since this yielded the highest level of agreement).

Findings presented in Table 3 below show that in spite of the disturbingly low inter-rater reliability indices based on actual scores assigned to candidates, there is intergroup agreement as to the overall status ascribed to five of the of the six videoed performances. Differences are still evident amongst members of each group but these differences diminish at the extremes of the proficiency continuum i.e. most assessors agree about instances of performance which are clearly satisfactory or unsatisfactory. Mechanisms which compensate for discrepancies amongst raters are proposed later in this paper.

	ESL Teachers (n=7)		Subject Specialists (n= 8)		Consensus Rating	
S	No of Satisfac-tory ratings	No of Not satisfac-tory ratings	No of Satisfac-tory ratings	No of Not satisfac-tory ratings	**ESL teachers**	**Subject specialists**
A	3	3	6	2	Borderline Mean = 4	Satisfactory Mean = 3.1
B	2	5	3	5	Not satisfactory Mean = 4.7	Not satisfactory Mean = 4.9
C	7	0	5	3	Satisfactory Mean = 3.8	Satisfactory Mean = 4
D	2	5	2	6	Not satisfactory Mean = 4.9	Not satisfactory Mean = 4
E	0	7	0	8	Not satisfactory Mean = 7.4	Not satisfactory Mean = 6.2
F	7	0	6	2	Satisfactory Mean = 2.3	Satisfactory Mean = 3

Table 3 Categorical Ratings Assigned By ESL Raters & Subject Specialists

4.2 School trials

Classroom trials, which are continuing, involve independent assessments by two parties: an ESL expert on the one hand, and one (or sometimes two) subject-specialists (maths/science teachers/teacher trainers) on the other. For control purposes the same ESL rater is involved in each of the observations, while subject specialists necessarily vary according to the school in which the trainee has been placed for teaching practice sessions. The ESL rater who also participated in the video trials, has been chosen for his reliability (the extent of his agreement with other raters averaged at r_I=.89). He can therefore be regarded as typical of this group.

Results obtained so far from the school-based trials are reported below. Given that they are based on an N size of only 19, it is readily acknowledged that trends identified may not hold good when further data becomes available. The difficulty of gaining access to subjects in the complex situation of real teaching practica has slowed the data gathering more than we had anticipated.

Table 4 shows the correlations between ESL and subject specialist raters on each of the criteria included on the schedule.

Criteria	Intra-class correlation r_I
Intelligibility	.69
Fluency	.60
Accuracy	.58
Comprehension	.58
Subject-specific language	.69
Interactive language	.83
Overall communication	.78

Table 4: Inter-Rater Reliability — Intra-class correlations between ratings assigned by ESL teacher and subject-specialist teachers/teacher trainers

While correlations for interactive language use and for overall communication are acceptable, there is considerable divergence between ESL and subject specialists in the way they rate candidates on all other categories. The possible explanation of this finding is that, by virtue of his training, the ESL teacher is more adept than the subject specialists in assessing the more traditional features of language proficiency. An examination of the distribution of scores shows that for accuracy the subject specialists appear to overrate low accuracy subjects and underrate high accuracy subjects when compared with the ESL rater — perhaps a further indication of uncertainty. Conversely, for comprehension the ESL rater tends to give a moderate comprehension score when the subject specialists gives a high (i.e. severe) one. For comprehension there proved in fact to be a significant difference ($t=2.24$ $p = 0.034$ two-tailed)[4] between mean scores of the ESL and subject specialist raters. The main reason for this difference is that the ESL rater has refrained in some instances from making a judgement about comprehension and annotated the procedure with comments such as "very little evidence", while the same candidates have been rated "unsatisfactory" by subject specialists. It may be that subject specialists are equating lack of classroom interaction (e.g. the teacher's tendency to hold the floor, non-response to student

[4] A two sample test was used. This allowed us to include the missing values for ESL raters when comparing mean scores.

questions) with inability to understand. This is borne out by a high level of agreement (r_I = .85) between comprehension and interaction scores assigned by subject specialists compared to a relatively low correlation (r_I = .52) between these two dimensions as rated by the ESL teacher.

The possibility that the subject specialist raters differ from the ESL teacher in the weighting of different categories in relation to their assessment of overall communicative effectiveness was explored by examining the relationship between analytical and global (overall communicative effectiveness) scores. The intraclass correlations shown in Table 5 below are interesting in two ways. First, it is somewhat surprising to note that accuracy is for both parties the lowest ranking criterion in relation to global assessment (even lower for the subject specialists than for the ESL teacher). This is at odds with the findings of Wilds (1975) Raffaldini (1988) and McNamara (1990), which point to the centrality of grammar in the assignment of second language oral proficiency ratings, and also contradicts the observations of Criper & Davies (1988) about subject specialists' obsession with the formal aspects of linguistic proficiency. Second, the figures suggest that interaction has by far the most powerful bearing on subject-specialists' overall judgements, whereas for the ESL rater this aspect of performance is less important.

Criteria	r_I (subject specialists)	r_i (ESL teacher)
Intelligibility	0.84	0.85
Fluency	0.74	0.80
Accuracy	0.59	0.70
Comprehension	0.82	0.75
Subject-Specific	0.73	0.85
Interaction	0.95	0.78

Table 5: Correlation Between Global And Analytic Scores — Intra-class correlations between ratings of communicative effectiveness and other categories as assigned by ESL teacher and subject specialist teachers/teacher trainers.

The extent to which each rater's overall communicative effectiveness scores could be predicted by one or other of the

analytical ratings was explored further by performing a stepwise regression on each data set. With the ESL data, subject specific language emerges as the first variable and comprehension as the second. In contrast, the subject specialist data selects only interaction.

Whereas both "comprehension" and "use of subject specific language" are generally accepted to be features of linguistic ability, the "interaction" section of our procedure is concerned solely with features of strategic competence which sit less comfortably with the commonly held view of what constitutes proficiency. Items included in this section such as "poses questions to check understanding of previously learned material", "grades questions appropriately for students and learning task.", "deals effectively with wrong answers". "adopts appropriate level of formality" have less to do with language resources per se than with the ability to use these resources effectively to accomplish communicative goals. Subject specialists thus appear to be more concerned with these classroom applications than the ESL rater who focuses more on the traditional components of language proficiency in making his global assessment. While the results of the stepwise regression must be interpreted a little carefully since they are based on only 19 cases and measure *linear* relationship rather than agreement, they give further support to the notion that subject specialists' conceptualize classroom language proficiency differently from ESL teachers.

ESL raters

Step	Category	R^2	Change in R^2	t
1.	Subject specific	69.53	69.53	5.45**
2.	Comprehension	82.54	13.01	2.99**

Subject specialist raters

Step	Category	R^2	Change in R^2	t
1.	Interaction	94.57	94.57	16.69**

**p = < .01

Table 6: Relationship Between Analytical & Global Scores — Stepwise Regression (n=19)

It remains to be considered whether these different orientations of ESL and subject specialist raters have practical implications as far as the reliability of our assessment procedure is concerned. The divergence between subject specialists' and the language expert ratings on the various dimensions of general language proficiency do not give grounds for confidence in the diagnostic capacity of the procedure.

There is on the other hand a better level of agreement ($r_I = 78$) between global scores assigned by each party, although the effect that any discrepancies in rater assessment are likely to have on overall satisfactory/unsatisfactory determinations warrants further attention. If we again, as for the video trial data, use a 4 rating (the mid-point on the scale) as our cut-off between satisfactory and unsatisfactory performance there is agreement between raters about the status of all but 4 of the 19 candidates observed so far (see Table 7 below).

S	*ESL rater global score*	*Subject specialists' global score*	*ESL rater's overall determination*	*Subject specialist's overall determination*
1	2.4	2.9	Satisfactory	Satisfactory
2	6.7	7.0	Not satisfactory	Not satisfactory
3	3.5	4.0	Satisfactory	Satisfactory
4	3.5	2.8	Satisfactory	Satisfactory
5	6	6.7	Not satisfactory	Not satisfactory
6	3.4	1.6	Satisfactory	Satisfactory
7	3	2.9	Satisfactory	Satisfactory
8	0.9	1.0	Satisfactory	Satisfactory
9	5.3	5.3	Not satisfactory	Not satisfactory
10	3.0	2.5	Satisfactory	Satisfactory
11	2.3	3.0	Satisfactory	Satisfactory
12	6.7	3.8	Not satisfactory	Satisfactory
13	5.5	5.5	Not satisfactory	Not satisfactory
14	7.7	6.7	Not satisfactory	Not satisfactory
15	2.6	5.5	Satisfactory	Not satisfactory
16	1.6	3.4	Satisfactory	Satisfactory
17	2.4	4.5	Satisfactory	Not satisfactory
18	4	2.5	Not satisfactory	Satisfactory
19	0.7	0	Satisfactory	Satisfactory

Table 7 — Satisfactory/unsatisfactory ratings assigned by the ESL teacher as against those of the subject specialists.

While the consensus level is not perfect it is sufficient to allow us to maintain that subject specialists can be used to make placement decisions, provided that certain safeguards are set in place.

5. Practical solutions to limited rater reliability

Since differences in overall determinations may effect the life chances of candidates our procedure is accompanied by a set of recommendations as to appropriate strategies for resolving such differences. They are as follows:

a) the supervising teacher should apply the schedule repeatedly in observing trainees' performance to ensure that a complete picture of his/her language ability is obtained and that improvement over the course of the practicum is taken into account;

b) determination of candidates' language proficiency status (satisfactory/unsatisfactory) on conclusion of the teaching practicum should be reached through consensus between at least 2 assessors (the visiting subject specialist lecturer/s and the supervising teacher/s) on the basis of independent applications of our schedule. Candidates whose performance is classed as unsatisfactory should be targeted for extra English support before undertaking further teaching practice, and may in extreme cases, be invited to withdraw from their studies;

c) where consensus is not reached this fact should be noted since it is likely that disagreement is an indication of "borderline" language proficiency which could be improved by additional ESL support. Candidates in this category as well as being offered on-course language support, should ideally be visited by an ESL teacher on the subsequent teaching practicum;

d) mechanisms should be set up to record results of classroom language proficiency assessments by all parties over the course of the academic year. This will assist courses administrators in making their final determinations about readiness to teach.

6. Conclusion

In this paper we have offered some very limited empirical support for the notion that subject specialists when assessing second

language proficiency in the context of classroom performance behave differently from language experts. The small size of our data set, and the fact that we are dealing with dependent rather than independent samples makes it difficult to determine whether such differences are significant. The trends in our data nonetheless allow us to posit (very tentatively) that subject specialists, in emphasizing interactive strategies above all else, are taking a Hymesian view of communicative competence by considering language proficiency in terms of real world criteria (i.e. are teachers creating the necessary conditions for classroom learning to take place?) which as subject teachers they feel well-qualified to assess. In behaving thus they come closer to what McNamara (1990) defines as the 'strong' approach to performance testing whereby language is assessed in terms of successful task completion, with all that that entails. Language experts on the other hand veer towards the 'weaker' (and arguably more conservative) approach by focusing more closely on what they are trained to assess, and that is the quality of the language sample elicited through the teaching tasks. The subject specialists' approach is in a sense invited in a performance test such as ours where there is no artificial manipulation of tasks and the only constraints placed on raters are the assessment criteria. In a less direct procedure, where occupation-specific performance is simulated rather than observed in a real-life setting, the test task is more obviously a pretext for assessing language.

Whether these weak and strong approaches to assessment make a difference to determinations arrived at through the application of our observation schedule is still uncertain and needs to be examined further with a larger sample. If it does not matter our procedure can be said to have accommodated both views. Analysis of the data gathered so far suggests that it does matter for diagnosis, but less so for placement as long as safeguards are put in place to compensate for discrepancies amongst raters.

While it is generally accepted that subject specialists should be consulted during the needs analysis phase of specific-purpose language test development, their role in the actual assessment process is seldom considered. Our findings to date suggest that this is an issue worth pursuing. Indeed, if we accept that there are instances of language performance where the formulation of an acceptable and intelligible message depends on discipline- or

occupation-specific knowledge, the involvement of subject specialists as assessors (notwithstanding the strain that this may place on reliability) could be regarded as a condition of test validity.

References

Allwright, R. (1980) Turns, topics and tasks: patterns of participation in language learning and teaching. In Larsen-Freeman, D. (ed.), *Discourse Analysis in Second Language Research.* Rowley, Mass.: Newbury House, Inc.

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

Bailey, K.M. (1985) "If I had Known Then What I Know Now: Performance Testing of Foreign Teaching Assistants" in Hauptman, C., R. Leblanc. & M. Wesche (eds.), *Second Language Performance Testing.* University of Ottawa Press, Ottawa, Canada. 153-180

Barnwell, D. (1989) "'Naive' native speakers and judgements of oral proficiency in Spanish" *Language Testing* 6:2, 1989. 152-163.

Bartko, J.J. (1966) The Intraclass Correlation Coefficient as a Measure of Reliability, *Psychol. Rep.*, 19, 3-11.

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

Criper, C. & A. Davies (1988) *ELTS Validation Report London:* The British Council/Cambridge: University of Cambridge Local Examinations Syndicate.

Douglas, D. and L. Selinker. (1990) *Performance on a General Versus a Field Specific Test of Speaking Proficiency By International Teaching Assistants.* Paper presented at the 1990 Language Testing Research Colloquium.

Galloway, V. (1980) Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal* 64, 428-33.

Hadden, B.L. (1991) Teacher and Nonteacher Perceptions of Second-Language Communication. *Language Learning* 41: 1. 1-24

Hinofotis, F. B., K.M. Bailey, & S.L. Stern. (1981) Assessing the Oral Proficiency of Prospective FTAs: Instrument Development. In Palmer A.S., P.J. Groot, G.H. Trosper (eds.) *The Construct Validation of Tests of Communicative Competence* Washington Dc TESOl 1981.106-126

Jones, R. (1979). Performance Testing of Second Language Proficiency. In Briere, E. and F. Hinofotis (eds.) *Concepts in Language Testing: Some Recent Studies.* Washington D.C.:TESOL21121

Rounds, P. L. Characterizing Successful Classroom Discourse for NNS Teaching Assistant Training. *TESOL Quarterly,* 21:4 643-671.

Ludwig, J. (1984) Native speaker judgments of second langauge learners' efforts at communication: a review. *Modern Language Journal* 66 274-83.

McNamara, T. F. (1990) *Assessing the language proficiency of health professionals.* PhD thesis, The University of Melbourne

Raffaldini, T. (1988) The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition* 10: 197-216.

Sinclair & Brazil, J. McH and Brail (1982) *Teacher Talk* Oxford University Press.

Stubbs, M. (1983) *Discourse Analysis.* Blackwell, Oxford.

Wesche, M. Second Language Performance Testing: the Ontario Test of ESL as an example. *Language Testing* 4/1 28-84.

Wilds, C.P. (1975) The oral interview test. In Jones, R.L. and Spolsky, B. (eds.). *Testing language proficiency.* Arlington, VA: Centre for Applied Linguistics, 29-44.

Appendix

DRAFT ONLY

PRODUCED BY CATHERINE ELDER & TOM LUMLEY,

LANGUAGE TESTING CENTRE, THE UNIVERSITY OF MELBOURNE, AUGUST 1991

Classroom Language Assessment for Maths and Science Teachers in Training

OBSERVATION SCHEDULE

The procedure is designed to be carried out during a 15-minute teaching segment.

You are asked to consider:

1. 6 general categories of classroom language use

2. Accompanying sets of criteria (ie. specific aspects of performance within these 6 categories), which have two functions:

2.1. They are designed to illustrate these categories and assist observers in making their assessments. They are not claimed to be exhaustive lists, nor are all criteria expected to apply in every lesson. Your overall judgements may of course be more powerfully influenced by some criteria than others.

2.2. They can be used to provide feedback to trainees about their strengths and about areas which need improvement.

INSTRUCTIONS

A. Fill in the following details:

Date: ***Name of trainee:***

Year level of class: ***Subject:***

Your name:

B. Read pages 2 & 3 carefully before observing the trainee's performance.

C. During your observation please follow these steps:

1. Rate trainee's performance in each major area by placing a cross anywhere on the line on the scale provided at the beginning of each section.

2. Tick the appropriate box for any of the individual criteria on which the trainee shows a definite need for further training .

3. In the space provided you may choose to write a comment about particular strengths and weaknesses in the trainee's performance.

4. When the observation session is over give a global rating of the trainee's current level of performance during the period observed, based on your perceptions of effective language behaviour.

****

1 GENERAL LANGUAGE PROFICIENCY

Rating for **INTELLIGIBILITY** of expression

highly satisfactory | acceptable | at risk | unsatisfactory

	Comment *(strengths & weaknesses)*	Needs work
1.1. projects and pitches voice appropriately		☐
1.2. pronounces words/sounds clearly		☐
1.3. utters sentences clearly (i.e. with suitable rhythm & intonation)		☐
1.4. clearly distinguishes questions, statements and instructions		☐
1.5. stresses important words/ideas (eg says them louder, more slowly, with pauses)		☐
1.6. clearly marks transitions from one idea/lesson stage to the next eg using words such as *so, now, right, we're going to*		☐
1.7. uses appropriate facial expression gesture, body movement		☐

Rating for **FLUENCY & FLEXIBILITY** of expression

highly satisfactory | acceptable | at risk | unsatisfactory

	Comment *(strengths & weaknesses)*	Needs work
1.8. speaks at appropriate speed		☐
1.9. speaks fluently (ie not too much stumbling, hesitation, groping for words)		☐
1.10. can express ideas in different ways (eg by rephrasing, elaborating, summarizing)		☐

Rating for **ACCURACY** of expression

highly satisfactory | acceptable | at risk | unsatisfactory

	Comment *(strengths & weaknesses)*	Needs work
1.11. grammar of spoken and written English is generally accurate		
1.12 . formulates questions clearly		☐
1.13. uses correct spelling & punctuation in boardwork and handouts		☐

Rating for **COMPREHENSION**

highly satisfactory | acceptable | at risk | unsatisfactory

	Comment (strengths & weaknesses)	Needs work
1.14. . demonstrates understanding of student language		☐
1.15. . seeks clarification of student language when necessary (eg. asks them to repeat/rephrase)		☐

2. USING SUBJECT SPECIFIC LANGUAGE

Rating for **SUBJECT-SPECIFIC** language	highly satisfactory	acceptable	at risk	unsatisfactory

		Comment (strengths & weaknesses)	Needs work
2.1	demonstrates knowledge of scientific and mathematical terms		☐
2.2	pronounces specialist terms clearly		☐
2.3	uses specialist terms judiciously (eg grading them and writing them on the board when appropriate)		☐
2.4.	makes clear the connections between ideas (eg stresses link words *if, since, in order to)*		☐
2.5.	explains scientific and mathematical processes/ concepts in ways appropriate to the audience (eg using simple language, familiar/concrete examples)		☐
2.6.	explains diagrams/models/use of equipment clearly		☐
2.7.	description/definition of terms/processes is a usable model for students' written assignments		☐

3. USING THE LANGUAGE OF CLASSROOM INTERACTION

Rating for language **CLASSROOM INTERACTION**	highly satisfactory	acceptable	at risk	unsatisfactory

		Comment (strengths & weaknesses)	Needs work
Involvement of students in class and lesson content			
3.1.	uses variety of forms of address (we, you, us/ student names)		☐
3.2.	poses questions to check understanding of previously learned material/new information		☐
3.3.	grades questions appropriately for students and learning task: simpler to more complex; closed/open		☐
3.4.	offers questions to individuals and whole class		☐
3.5.	clearly signals acceptance/rejection of student response		☐
3.6.	responds appropriately to students' questions, requests for assistance		☐
3.7.	deals effectively with wrong answers, non- response (eg by rephrasing questions/reviewing steps in a process)		☐
Classroom control			
3.8.	adopts appropriate level of formality/firmness		☐
3.9.	gives clear instructions		☐
3.10.	maintains contact with class while dealing with individual demands/using blackboard, etc.		☐

OVERALL COMMUNICATIVE EFFECTIVENESS

Rating for **OVERALL COMMUNICATIVE EFFECTIVENESS**	highly satisfactory	acceptable	at risk	unsatisfactory