Who should be the judge? The use of non-native speakers as raters on a test of English as an international language[1]

Kathryn Hill, NLLIA-LTRC, University of Melbourne

1. Abstract

the features of Standard English in any variety, native speaker or non-native, are not what any outsider thinks they should be. (Lowenberg 1994: 59)

A great deal of debate has centred on how native speaker competence in a language should be defined (eg. Davies 1991, 1994; Paikeday 1985). Nevertheless, performance on language tests is typically judged, explicitly or implicitly, with reference to a native speaker ideal (Hamilton, et al 1993). By extension, examiners for 'high stakes' tests such as IELTS or TOEFL are usually required either to be native speakers of the target language or to have 'native-like'competence.

However, a number of authors have challenged the notion that the elusive native speaker represents an appropriate model of English for language teaching and assessment beyond the 'Inner Circle', ie. countries like Australia, the UK, the USA and Canada (Kachru 1994, Lowenberg 1994)[2]. This paper reports on a test of teacher proficiency in English, developed for use within Indonesia, where interaction with native speakers of English is the exception rather than the rule. A feature of this test is, not only that it promotes the local (ie. Indonesian) variety of English as its model, but that it uses Indonesian, in preference to native, speakers of English as raters. Just as only native speakers have a strong claim to understanding

[1] An earlier version of this paper was presented at the 1996 Language Testing Research Colloquium, 31 July–2 August, Tampere, Finland.

[2] "According to Kachru (1988:2)... the Outer Circle... is composed of "the former British colonies and American protectorates." Here English 'may be an official language used in educational, commercial and trade institutions, but it is generally not the native language of the citizens' (op cit). In the Expanding Circle, including China, Indonesia and Japan, English is not an official language and is viewed as a foreign language learned primarily for international communication." in Vavrus (1991:192)

Melbourne Papers in Language Testing 1996 Volume 5.2 pp. 29-49. The Language Testing Research Centre, The University of Melbourne.

what it is to be 'native-like' in English, here it was felt that only Indonesian speakers of English could define an acceptable level of performance on the test for its purposes. Decisions to use exclusively native or non-native speaker raters contain an assumption that one or the other group would rate the same performance differently.

2. Literature review

Studies comparing the behaviour of native and non-native speaker raters, whilst they differ in a number of respects, including the language being assessed and whether raters are trained or 'naive', have tended to find that non-native speakers judge test performance more harshly. In a study of reactions to native English speakers and native Spanish speakers to tapes of Puerto Rican speakers of English, Fayer and Karsinski (1987) found that the Spanish (ie. non-native speaker) group were less tolerant. Although the two groups ranked students in the same way, the Spanish group gave significantly lower ratings on some linguistic criteria and were more likely to react unfavourably to some features. It should, however, be noted that these judgements were made by naive (ie. untrained) raters and without descriptive criteria upon which to base their judgements. Sheorey (1985) also found non-native speakers of English to be less tolerant of their peers than were native speaker assessors. Ross (1979), comparing judgements of the correctness of a set of sentences in English, found the non-native speakers tended to reject more sentences and to make fuller use of the rating scale than the native speakers (see Davies, this volume). Finally, Santos (1988) asked a group of professors to rate the academic writing of two non-native speaking students using an analytic scale. All had been briefed in the use of the scales. Once again the nonnative-speakers were found to be more severe in their judgements.

In constrast to these studes, Barnwell (1989) found the 'naive' native speakers in his study tended to be harsher than an ACTFL trained non-native speaking assessor on a test of oral proficiency in Spanish. However, as in the Fayer and Karsinski (1987) study, it is unclear whether it was their 'nativeness' or the lack of training which accounted for the differences. Furthermore, although they ranked candidates in the same way, there were considerable differences between the 14 native-speaker raters in the levels actually assigned to individual candidates. As only a single rater was used for comparison, the reliability of the trained (ie. non-native) rater

cannot be assessed. The fact that two different scales were used (ie. the original and a translation into Spanish) is also problematic.

Brown (1995) used multi-faceted Rasch measurement to compare, inter alia, the ratings of native and non-native speakers of Japanese on a test of speaking for tour guides. Both groups had undergone training in the use of an analytic scale with descriptive criteria. She concluded that there were no significant differences between the native and non-native Japanese speakers in terms of rater harshness and the ranking of candidates. However, she found that the non-natives demonstrated less variation in their harshness estimates (ie. they were more in agreement than the native speaking group) and that there were some differences between the two groups in their use of individual assessment criteria. Finally, whilst there was no difference between the groups in terms of misfit, a number of non-native raters were found to be 'overfitting' (ie. they demonstrated less variability in their assessments than expected by the Rasch model). On the basis of these findings, Brown concluded that there was "little evidence that native speakers are more suitable than non-native speakers [as raters]" (13).

In all of these studies test performance has been judged with reference to a native-speaker ideal. In contrast, this study will compare native and non-native speaker raters where the local (ie. Indonesian) variety of English is the criterion.

3. Background to the test

The status of English as a language for communication within and between South East Asian countries has risen dramatically in recent years. In Indonesia, all high school students must study English language for 4 to 6 hours per week. English has been taught in years 5 and 6 in a small proportion of primary schools since 1992 and, it is intended, the study of English as a compulsory subject from primary school will eventually become the norm throughout Indonesia.

However, the demand this has created for teachers of English has given rise to concerns about teachers' language proficiency. This, in

turn, has highlighted the need for some means of assessing teachers' English language proficiency[3].

4. Description of the test

The English Proficiency Test for Indonesia (EPTI)[4] is a specific purpose test designed to assess English language proficiency as relevant to classroom teachers. It comprises two integrated tests: Reading/Writing (2 hours) and Listening/Speaking (1 hour). Its purpose is to assess teachers at the end of training and/or to determine eligibility for in-service training for practising teachers. However, it is also hoped that the communicative orientation[5] of the test will have a positive influence on the way teachers are trained.

Barnwell (1989) claims that "the native speaker is the target of communicative efforts [for foreign language learners]" (154). However, this is only true for the very small number of Indonesians who have the opportunity to study abroad. In reality the majority of Indonesian learners will use English to communicate with other non-native speakers within South East Asia[6]. For this reason it was decided the test should emphasize the ability to communicate effectively in English as it is used in the region, rather than relate proficiency to the norms of America, Britain or Australia. This approach is a pragmatic one, to the extent that demand for teachers is increasing and that the majority of local teachers will rarely achieve native-like proficiency. However, this approach also aims to recognise the Indonesian variety of English both as an

[3]Existing international tests were found to be unsuitable both practically (in terms of costs and turnaround time) and, being designed primarily for students intending to study abroad, in terms of their cultural orientation. Existing local assessment procedures, which had not been professionally developed, posed problems for reliability (Brown & Lumley 1994).

[4]The test has been developed by the NLLIA-Language Testing Research Centre (LTRC) in collaboration with the South East Asian Ministers of Education Regional Language Centre (SEAMEO-RELC) and a number of state teacher training institutions (IKIP) in Indonesia with funding from AusAid.

[5]This test is 'communicative' to the extent that it includes authentic texts as well as writing and speaking components and that it emphasises the effectiveness of the performance in contrast to the focus on grammar and mechanics typical of language classrooms in Indonesia.

[6]Kachru claims that "an overwhelming majority of the users of English in the outer circle (e.g. Malaysia, Singapore, India, Sri Lanka, Nigeria) only minimally interact with the native speakers"(1994: 5)

appropriate model to be provided by teachers and as a valid target for learners (Brown & Lumley 1994).

5. The Indonesian variety

Lowenberg (1994) defines the standard model of a variety of English in any country as "the accepted language usage for official, journalistic and academic writing; for public speaking before an audience or on radio or television; and for use as a medium and/or subject of instruction in the schools."(7)

In the context of developing this test, there was no attempt to define Indonesian English or what distinguishes it from the standard (ie. American, British or Australian) varieties. Instead, the local (ie. Indonesian) variety was recognised firstly, by selecting the reading and listening texts from authentic local English language materials on topics relevant to Indonesia and, secondly, by having all assessment carried out by trained IKIP[7] lecturers.

According to Lowenberg (1994),

> *Often...deviations [from native-speaker English] reflect non-native norms that are appropriate in many English educated speech communities, and to assume without any further information that such features reveal deficiencies in English is to perform an invalid assessment of their user's proficiency.* (1994: 63)

The advantage of using trained IKIP lecturers as raters is, therefore, that they alone can distinguish differences from deviations. Furthermore, as teacher trainers, they are uniquely placed to determine whether a candidate provides an acceptable model of language use for learners of English in Indonesia.

However, the success of this approach very much depends upon raters' acceptance both of the communicative orientation of the test and of the Indonesian variety of English as an appropriate model. In an earlier paper on this project it was claimed, "the rating scales and the model implied in them were accepted by the raters and

[7] state institutes of teacher training

they had no more difficulty in using them than would a comparable native speaker group" (Brown & Lumley 1994: 127). The rest of this article is devoted to investigating this claim.

6. Research questions:

1. Do native (Australian) and non-native (Indonesian) speaker raters rate writing performance in English differently?

2. What do these results tell us about:

(i) the suitability of non-native speakers as raters;

(ii) the extent to which raters have accepted the communicative orientation of the test.

7. Methodology

Data for the present study come from the extended writing task on the Reading/Writing sub-test of EPTI. The task requires candidates to comment on an article from an English language teaching journal. Candidates are allowed 30 minutes for the task and must write at least 100 words. The writing scripts used in this study came from a trial administration of the test conducted in Indonesia in April, 1994. The 100 trial candidates were all local students in their final year of teacher training.

13 English lecturers from two teacher training institutes in Indonesia (IKIP Malang and IKIP Surabaya) and 10 experienced English-speaking background raters in Australia participated in the study. Each group underwent a half day of training before rating commenced, the Indonesian group at IKIP Malang in April, 1994 and the Australian group at the University of Melbourne in August, 1994. Rater training was conducted by the researcher on each occasion.

Rater training is considered essential for subjectively scored tests in order to reduce the amount of variability between examiners and to ensure than individuals are rating consistently (Lumley & McNamara 1993). As the approach was very new to them, rater training also provided a means for ensuring that the Indonesian raters understood and accepted the communicative orientation of the test.

Both groups of raters were asked not to employ the idealised native speaker as a reference point. Instead they were asked to think of what would constitute a good model for the purposes of teaching English in Indonesian high schools, where the majority of students would ultimately use English to communicate with other Asians. Naturally it was difficult for the Australian raters, who were unfamiliar with the Indonesian school system, let alone the Indonesian variety of English, to conceptualise this criterion. Therefore, the Australian group were asked to think of a level of performance which would be considered acceptable for a foreign language teacher in Australian classrooms. In each training session it was left to the group to define acceptable levels of performance.

100 scripts were each double marked by 13 Indonesian raters (ie. a total of 200 ratings). Of this 100, 30 scripts were selected at random and these were double marked by 10 Australian raters (ie. a total of 60 ratings). Scripts were rated using a six point analytic scale. The five categories for assessment were 'overall impression', 'content', 'vocabulary', 'coherence & cohesion' and 'control of linguistic features'. An average score of 4 was chosen as the cut-off for 'adequate', a threshold which is explicit in the descriptors for 'overall impression'(below).

6.	Outstanding model for learners of English.
5.	A good model for learners of English
4.	An adequate model for learners of English
3.	An inadequate model for learners of English
2.	A poor model for learners of English
1.	Minimal answer - insufficient language to assess

Table 1. Descriptors for 'Overall Impression'

8. Analysis

Data were analysed using FACETS (Linacre 1989-1994), an extended Rasch model. The facets investigated were candidate, rater, type

and item. FACETS provides an estimate of item difficulty[8] and rater severity, and adjusts candidates' ability estimates accordingly. This ability estimate (expressed as a logit) is expressed as the probability of a candidate obtaining a certain score given the ability of the candidate, the difficulty of the item and the harshness of the rater (Linacre 1992). As the output from FACETS provides information on rater consistency ('fit') as well as severity, researchers have been able to use FACETS to compare different types of raters (e.g. Brown 1995; Lumley 1995) as well as to monitor the performance of individual raters (Wigglesworth, 1993; Wigglesworth, Morton & Williams, forthcoming).

Three separate data sets were analysed using FACETS. The first combined the Australian and Indonesian ratings. In this data set all 100 candidates had been double rated and 30 of these had been rated 4 times (ie. twice by each group). The second data set comprised ratings from the Australians only (30 candidates, each double rated) and the third was of the Indonesian ratings only (100 candidates, each double rated). In all analyses, 'candidate' was the non-centred facet.

9. Results and Discussion

1. Do native (Australian) and non-native (Indonesian) speaker raters rate writing performance in English differently?

FACETS was used to determine whether the the two groups of raters were comparable in terms of:

(i) consistency (fit)

(ii) harshness

(iii) use of the assessment criteria

(i) How consistent are individual raters within the respective groups?

[8] in this case, the items are the five individual assessment criteria for the writing task

FACETS provides information regarding rater consistency as well as rater harshness. Raters whose pattern of scoring is inconsistent with the overall trend for the other raters, items or candidates or who are internally inconsistent in their ratings are identified as misfitting by the model. Those whose ratings show too little variation or lack of independence are termed 'overfitting'. Generally speaking, overfitting raters are acceptable whereas misfitting raters are not (Brown 1995: 7).

An initial analysis of the combined data set showed that two raters, one Indonesian and one Australian, were misfitting. In addition one Australian rater was overfitting. Data for the two misfitting raters were removed from all subsequent analyses.

(ii) Does one group rate more harshly than the other?

Comparison of rater severity was approached, first of all, by comparing the rater harshness estimates and use of the rating scale for each group. Next, the number of candidates placed above and below the cut-off by each group were compared.

Comparison of harshness estimates

With subjective assessment it is normal to expect some differences between individual raters (rater training notwithstanding). In the second analysis of the combined data set (ie. with data from the misfitting raters removed) rater harshness estimates ranged from -1.63 to 2.45 (a range of 4.08 logits). The separation index of 4.9 shows that individual raters within the combined (ie. Australian/ Indonesian) group can be reliably separated into 5 different levels of severity. As Brown (1995) has pointed out, this situation underscores the importance of multiple ratings and moderation, as it is unlikely that, when the test is in use, statistical programs such as FACETS will be available to compensate for the inevitable differences in rater harshness.

	Score	Estimate	Error	Fit	Rater
Indonesians	297	-1.52	0.22	1.1	22
	289	0.08	0.19	0.9	23
	196	-1.63	0.25	1.0	24
	367	-0.42	0.18	1.0	25
	272	0.13	0.19	0.8	26
	261	0.69	0.20	1.3	27
	330	-1.42	0.20	1.1	28
	290	0.12	0.19	1.0	29
	311	-0.47	0.19	0.9	30
	308	-1.67	0.20	1.0	31
	303	-0.68	0.19	0.8	32
	241	0.80	0.21	1.0	33
Australians	180	0.24	0.23	0.9	101
	185	0.76	0.23	0.8	103
	180	1.23	0.23	0.9	104
	192	-0.19	0.23	1.3	105
	160	1.84	0.23	1.1	106
	208	-0.49	0.23	0.9	107
	187	0.68	0.23	1.0	108
	151	2.45	0.23	1.0	109
	202	-0.53	0.23	0.9	110
Separation 4.90; Reliability 0.96					

Table 2. Rater harshness estimates

To see if there were differences between raters at the group level, rater harshness estimates (from the same analysis) were compared. The Australians harshness estimates had a range of 2.9 logits (-.53

to 2.45) compared to 2.4 logits (-1.63 to 0.8) for the Indonesians. Given that the Australian raters were more experienced than their Indonesian counterparts, it is surprising that there were greater differences between the Australians in terms of their harshness estimates than within the Indonesian group. However, this finding is similar to Brown's study where non-native speakers of Japanese were found to be more in agreement and "more likely to prove to be within the limits of acceptability (ie. they produce a narrower range of harshness variability)" (1995: 9).

A t-test was then used to investigate whether the difference in overall harshness estimates between the two groups was significant. Table 2 shows that the logit for the Australian group was significantly higher than that for the Indonesian group. In other words the Australians were significantly harsher.

	N	Mean Logit	STD	SE Mean
Indonesians	12	-0.50	0.89	0.26
Australians	9	0.67	1.03	0.34
(t = -2.7, df = 15, p = 0.016)				

Table 3. Comparison of rater harshness estimates

Use of the rating scale

A further analysis of the combined data set was performed to investigate whether the two groups use the rating scale in the same way. In order to do this a separate rating scale structure was specified for each group and the facet 'rater type' was anchored at zero. The output from this analysis provided information on the difficulty of each of the six levels of the scale by indicating the candidate ability estimate at which each level on the rating scale becomes the most likely to be awarded for each group of raters. Table 3 gives a graphic representation of this. The numbers (1 to 6) in the table are located at the ability estimates where that level on the scale commences to be the most probable score awarded for each group (Linacre 1992).

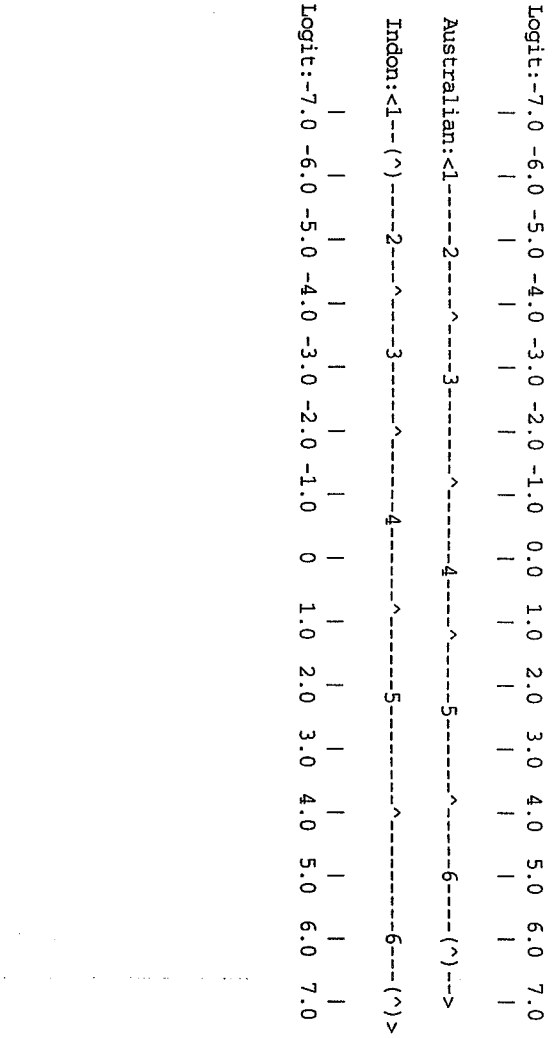

Table 4. Use of the scale

This indicates that the Australian group rated more harshly at the cut-off (ie. 4) than the Indonesians. That is, to be judged 'adequate', candidates would, on average, need an ability estimate (logit) of 0.24 from the Australian group whereas they would only need an ability of -0.44 to get the same score from the Indonesian raters (a

difference of 0.68 logits). However, it appears the Indonesians were more reluctant to use the top of the scale than the Australians. That is, to get a score of six from the Australian group a candidate would, on average, only need an ability estimate of 5 whilst they would need an ability estimate of 6 to get the same score from the Indonesians.

Although the descriptors for a score of 6 for each of the assessment categories (below) deliberately avoid reference to a native speaker ideal, these results indicate that the Indonesians may still have been unconsciously applying a native speaker standard.

Overall

Outstanding model for learners of English.

Content

An excellent answer. Ideas are plentiful, well developed, and relevant to the task. The purpose of the letter is clear.

Coherence & cohesion

The letter reads fluently. Ideas are presented clearly and logically, with good control of cohesive devices.

Control of linguistic features

The writing demonstrates complete control of linguistic features with no noticeable errors.

Vocabulary

An extensive vocabulary is used accurately and effectively.

Descriptor for '6' for each assessment category

The cut-off

Candidate ability estimates from analysis of each of the two separate data sets were compared to see if there was any difference in the proportion of candidates each group placed above and below the cut-off. This was done by looking at the fair average (rather than the logit) which represents the rating of the candidate by a mean rater of the group on a mean item (Linacre 1992). The facets included were candidate, rater and item.

Not surprisingly, Table 4 shows that fewer candidates (n=19) were judged to be 'adequate' (ie. their 'fair average' score was 4 or above) by the native speakers than by the non-native speakers (n=26). An agreement coefficient was calculated as 0.7, indicating that the two groups classified the candidates as passing or failing with only 70% agreement.

		INDONESIANS		
		Pass	Fail	Total
AUSTRALIANS	Pass	18	1	19
	Fail	8	3	11
	Total	26	4	30

Table 5. Pass rate for each group (fair average > 4)

These results indicate that the two groups are interpreting the threshold somewhat differently, ie. they have a different concept of what level of performance is 'adequate' for the purposes of the test. This result is perhaps not suprising given that the Australians were divorced from the Indonesian context. Furthermore, many of the Australian raters also examine a test used to screen non-English speaking background applicants for teacher training courses in Australia. It may be, therefore, that these raters were unconsciously

applying the same standards (which are higher than what would normally be expected of a foreign language teacher) to the Indonesian candidates. An alternative explanation is that the standard expected of foreign languages teachers in Australia is more than is expected of their Indonesian counterparts.

(iii) Do the two groups use the assessment criteria in the same way?

The five assessment criteria were selected to reflect the communicative orientation of the test and the descriptors designed to avoid any reference, either implicit or explicit, to a native speaker ideal. However, according to McNamara, rater training may not always succeed in instilling the desired orientation amongst raters. Rather, he suggests, raters may unconsciously reinterpret the criteria according to pre-existing notions of acceptability (1990: 68). On this basis interrater reliability may merely indicate agreement amongst the group on the salient features of performance, independent of training. McNamara found "perceptions of grammatical and lexical accuracy... played a crucial role in determining the candidate's total score" (ibid: 63). Such a finding for the present study would indicate that the raters were not using the scale as intended and, as such, their behaviour would constitute a threat to the construct validity of the test.

In order to compare use of items across the two groups, the first data set (ie. the combined Australian and Indonesian ratings) was modified so that items were numbered 1-5 for the Australian ratings and 6-10 for the Indonesian ratings and the mean for each set of items was set to zero. The two sets of item difficulty estimates produced by this analysis were then examined, firstly, to find out whether the two groups were using the assessment criteria in the same way and secondly, for evidence that raters had accepted the communicative orientation of the assessment criteria.

The item (ie. assessment category) with the highest logit is the one that it is hardest to get a high score on and vice versa. Table 5 shows that both groups were harshest on 'Overall Impression', ie. the category where raters made an explicit judgement about the acceptability of the candidate as a classroom model. 'Control of Linguistic features ' was the next most difficult category on which to gain a high score for both groups. After adjustment for error the only difference between the two groups was for 'Coherence and Cohesion',

with the Indonesians judging this assessment category more harshly than the Australians.

Estimate	Error	Fit.	Item
0.88	0.17	0.9	overall (Aust)
0.75	0.13	0.9	overall (Indon)
0.18	0.17	0.7	ling features (Aust)
0.13	0.13	1.0	ling features (Indon)
-0.08	0.17	0.8	vocab (Aust)
-0.39	0.13	0.8	vocab (Indon)
-0.23	0.13	1.0	coherence & c (Indon)
-0.60	0.17	1.0	coherence & c (Aust)
-0.26	0.13	1.2	content (Indon)
-0.38	0.17	1.5	content (Aust)
0.00	0.15	1.0	Mean (Count: 10)
0.47	0.02	0.2	S.D.

Table 6. Assessment categories

The only assessment category demonstrating misfit was 'Content' (Australians only). For all other categories the fit values fall within the acceptable range, indicating that "scores in these categories are making independent contributions to the underlying ability dimension constructed by the analysis" (McNamara 1990: 60). In addition, the absence of overfit on any category for either group indicates that no single assessment category had an undue influence on the determination of the final score.

10. Conclusion

The findings of this study appear to support Brown and Lumley's (1994) claim. That is, there is nothing in these findings to suggest that the Indonesian raters are any less suitable to rate a test of English language proficiency than the native speakers. In fact, all but one of the Indonesians raters were found to be rating consistently and, although the Australian group were more experienced, the

Indonesian group were more in agreement (ie. had a narrower range of harshness estimates).

As to the extent to which raters have accepted the model implied in the test, the fact that no one category exercised undue influence indicates that the Indonesian raters have accepted the communicative orientation of the rating scales. However, the question remains as to whether raters have accepted the non-native ideal proposed by the test. On the one hand, contrary to earlier studies, it was the native speakers who tended to be harsher overall, and the Indonesians proved to be less demanding than the Australians in terms of the minimum level of proficiency expected of intending teachers. On the other hand, the fact that it was harder to get a score of 6 (ie. the maximum score) from the nonnative speakers, suggests that the Indonesians may still have been unconsciously applying a native speaker standard at the top of the scale. Further research is necessary to investigate this particular issue.

For a locally administered test, the decision to use local nonnative raters is essentially a pragmatic one. Nevertheless, given the apparent success of this approach and the frequent difficulty of finding suitably trained native speakers when tests, such as IELTS, are administered in countries like Indonesia, the findings of this study suggest it may be appropriate to employ competent local users of English as raters.

11. References

Abdullah, N. A. (1994). Non-standard forms in Malaysian Written English: Variation or Deviation? In Gill, S. K. (Ed.). *Proceedings of the International English Language Education Conference.* (pp. 289–296). Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

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

Brown, A. & Lumley, T. (1994). How can English proficiency tests be made more culturally appropriate? a case study: The assessment of English teacher proficiency. In Gill, S. K. (Ed.). *Proceedings of the International English Language Education Conference.* (pp. 122–128). Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia.

Brown, A. (1995). The effect of rater variables in the development of an occupation specific language performance test. *Language Testing, 12,* (1), 277–303.

Brown, J. D. (1990). Short cut estimators of criterion-referenced test consistency. *Language Testing, 7,* (1), 77–97.

Davies, A. (1991). *The Native Speaker in Applied Linguistics.* Edinburgh. Edinburgh University Press.

Davies, A. (1994). Native speaker not dead! Alive and well in Standard languages. Paper presented at the Applied Linguistics Association of Australia Conference, July.

Fayer, J. M. & Krasinski, E. (1987). Native and nonnative judgements of intelligibility and irritation. *Language Learning 37,* 313–326.

Galloway (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64,* 428–433.

Hamilton, J., Lopes, M., McNamara, T. & Sheridan, E. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. *Language Testing, 10,* (3), 337–354.

Kachru, B. B. (1986). The alchemy of English: the spread functions and models of non-native English. Oxford. Pergamon. Reprinted 1990. Urbana: University of Illinois Press.

Kachru, B. B. (1988). Teaching World Englishes. *ERIC/CLL News Bulletin, 12,* (2–4), 8.

Kachru, B. B. (1994). Teaching World Englishes without Myths. In Gill, S. K. (Ed.). *Proceedings of the International English Language Education Conference.* (pp. 1–19). Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia.

Linacre, J. M. (1987/1994). FACETS Version#2.75

Lowenberg, P. H. (1994). What do ESL tests Test? Issues of Cross-cultural norms. In Gill, S. K. (Ed.). *Proceedings of the International English Language Education Conference.* (pp. 57–65). Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia.

Lowenberg, P. H. (Ed.) (1988). Language Spread and Language Policy: issues, implications, case studies. Washington, D.C., Georgetown University Press. Georgetown University Round Table on Languages & Linguistics (38th 1987 Washington D.C.)

Lumley, T. & McNamara, T. F. (1993). Rater characteristics and rater bias: implications for training. Paper presented at *Language Testing Research Colloquium.* Cambridge, August.

Lumley, T. (1995) Doctors vs language specialists as raters in a test of English for health professionals. *Melbourne Papers in Language Testing, 4,* (1), 74-98

Lunz, M. E. & Stahl, J. (1990). Severity of grading across time periods. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA., April, 1990.

McIntyre, P. (1993). The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples. Unpublished Masters Thesis. University of Melbourne

McNamara, T. & Adams, R. (1993). Exploring rater behaviour with Rasch techniques. Paper presented at the Language Testing Research Colloquium, Princeton, NJ, March [ERIC Document Reproduction Service #ED 345 598]

McNamara, T. (1990). Item Response Theory and the validation of an ESP test for health professionals, *Language Testing, 7,* (1), 52–76.

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

Morton, J., Wigglesworth, G. & Williams, D. (forthcoming). Approaches to the evaluation of interviewer behaviour in oral test. In Wigglesworth, G. & Brindley, G. (Eds.). *access: issues in English language test design and delivery.* Sydney: NCELTR.

Nair, A. B. (1994). Motivation without need: a case for promoting Malaysian English. In Gill, S. K. (Ed.). *Proceedings of the International English Language Education Conference.* (pp. 114–121). Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia.

Paikeday, T. M. (1985). The native speaker is dead! Toronto and New York Paikeday Pub. Co.

Quirk, R. (1988). The question of standards in the international use of English. In Lowenberg, P. H. (Ed.). (1988). *Language Spread and Language Policy: issues, implications, case studies.* Washington, D.C., Georgetown University Press. Georgetown University Round Table on Languages & Linguistics (38th 1987 Washington D.C.)

Ross , J. R. (1979). Where's English? In Fillmore, C. J., Kempler, D. & Wang, W. S-Y (Eds.). *Individual differences in language ability and language behaviour.* (pp. 127–163). Academic Press: New York.

Santos, T. A. (1988). Professors' reactions to the academic writing of non-native speaking students. *TESOL Quarterly, 22,* (1), 69–90.

Sheorey, R. (1985). Goof gravity in ESL: native vs. nonnative perceptions. Paper presented at the 19th annual TESOL convention. New York.

Soenjono, D. (1996). English policies and their classroom impact in some ASEAN/Asian countries. RELC Seminar, Singapore, April.

Stahl, J. & Lunz, M. (1992). Judge Performance Reports. Paper presented at AERA, San Francisco, April.

Vavrus, F. K. (1991). When paradigms clash: the role of institutionalised varieties in language teacher education. *World Englishes, 10,* (2),181–195.

Widdowson, H.G. (1994). Proper English and appropriate English. In Gill, S. K. (Ed.). *Proceedings of the International English Language Education Conference.* (pp. 31–41). Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10,* (3) 305–336.