Rating scales and native speaker performance on a communicatively oriented EAP test[1]

Jan Hamilton, Marilyn Lopes, Tim McNamara and Eileen Sheridan

Abstract

Explicit or implicit references to the performance of native speakers are to be found in rating scale descriptors in communicatively oriented tests. But the use of the native speaker as a reference point derives from a pre-communicative tradition, and the performance of native speakers on cognitively demanding communicative tests has not been carefully investigated. This paper reports on three studies of the performance of native speakers of varying educational backgrounds on a test of reading and writing skills in English for Academic Purposes contexts. The results show that reference to native speaker performance in rating scales is unwarranted, and help us to understand the nature of the skills being measured in such tests.

1. Introduction

For some years now, the question of the performance of native speakers on language proficiency tests designed for non-native speakers has been the subject of research and debate. On the one hand, a frequent practice has been for test developers to examine the performance of native speakers on test items as part of test development; items which they cannot manage easily are excluded. On the other hand, the assumption of homogeneity of native speaker performance behind this practice has been questioned by researchers such as Alderson (1980) and Bachman (1990).

[1] Earlier versions of parts of this paper were given as McNamara and Hamilton (1991) and McNamara, Hamilton and Sheridan (1992). We are grateful to Greg Deakin of IDP Australia for permission to use the exemplar version of the IELTS reading and writing tests in this research, and to the Editing Committee of IELTS chaired by Professor Chris Candlin for permission to publish the research. Aspects of the research reported here were done in order to fulfill in part the requirements of the degree of M.A. in Applied Linguistics at the University of Melbourne. Readers are referred to Hamilton (1991), Sheridan (1991) and Lopes (1992) for fuller discussion and presentation of data than could be included in a short paper.

Melbourne Papers in Language Testing 1993 Volume 2.1 pp. 1-23. The Language Testing Research Centre, The University of Melbourne.

Previous studies have considered mainly the performance of native speakers on discrete point tests of grammar, vocabulary, on tests of relatively decontextualized listening and reading skills, or on integrative tests of vocabulary and syntax such as cloze tests. To date, little consideration has been given to native speaker performance on performance tests, that is the class of tests involving test tasks which simulate in the test situation the tasks facing test takers in real life.

The reporting of performance on such tests and on other communicatively oriented tests of receptive and productive skills frequently involve the using of rating scales. The terms in which the rating scales descriptors are couched are important, because they constitute implicit definitions of the construct on which the test is based. Frequently, such scales make specific reference to the performance of native speakers in describing the top level of the scale; this practice dates from the earliest oral proficiency interview test, the Foreign Service Institute (FSI) Oral Proficiency Interview, and has survived (with some cosmetic modifications) in many related rating scales. Little empirical evidence has been adduced to support the location of native speaker performance at the top of the scale.

In this paper, a series of studies examining data from native and non-native speaker performance on a test of reading and writing English for Academic Purposes (EAP) will be reported. The implications of our findings for the construct validity of performance tests will be considered, with particular reference to the way in which scalar descriptions reflect such a construct. The implications of implicit or explicit reference to native speaker performance in such descriptors will be a particular focus of the discussion. Issues of equity will also be raised. The discussion will make reference to a distinction between a *strong* and a *weak* sense of the term performance test, which depends on the focus of assessment. To the extent that performance on the test involves factors other than straight second language proficiency, and these factors are included in the assessment, then we may expect there to be an overlap in the performance of native and non-native speakers; and the performance of native speakers will be highly variable, as we will demonstrate from the data.

2. The performance of native speakers on foreign language proficiency tests

The performance of non-native speakers on language tests used for purposes of academic selection has long been an object of study. Studies can be divided into those involving non-communicative tests such as the TOEFL, and communicative EAP tests such as the TEEP (Weir, 1988a), ELTS (Alderson and Hughes, 1981; Weir, 1998b), IELTS (British Council/UCLES 1989a,b) and others.

In the former category of tests, the performance of native speakers has been assumed to be relatively homogeneous, and at the top of the range of possible test scores, on the assumption that language proficiency is something that native speakers possess, and possess uniformly well; this distinguishes them from non-native speakers. A number of studies have concluded that TOEFL does discriminate between native and non-native speakers in this way (Angoff and Sharon, 1971; Johnson, 1977; Clark, 1977), although native speaker performance was less homogeneous, and relatively lower, in the sub-tests of Reading Comprehension and Writing Ability than in the other sub-tests (Listening Comprehension, English Structure and Vocabulary).

A more recent study in the same tradition is that of Oscarson (1986), who investigated the construct validity of a national test of English as a foreign language in Sweden by comparing the performance of Swedish upper secondary level students with a group of English subjects matched for age. Sub-tests of vocabulary, phrases, grammar, reading and listening comprehension were found overall to reveal significant differences between the two groups, but there were differences among the sub-tests. In particular, on the longer of two reading passages, the non-native speaker mean score in fact narrowly exceeded that of the native speakers. The native speaker mean for the whole test was relatively high (83.4%), but by no means perfect. Oscarson argued nevertheless that the significant difference between the two groups in their results on the test overall was evidence of the construct validity of the test.

It is interesting that in both the TOEFL studies and in the Swedish study the argument for the distinctiveness of the performance of the two contrasting groups is weakest on the most communicative parts of the tests (those focusing on whole skills such as reading or listening).

EAP tests in the communicative tradition have been less certain about the competence of the native speaker as a reference point, both in the empirical validation of tests and in the wording of rating scales defining levels of performance on the tests. We can distinguish those who appear to be recommending and using the performance of native speakers as a reference point in test development and validation, those who caution about such an approach, and a number of fence sitters. A number of writers appear to favour reference to native speaker performance. Cziko (1983: 294) suggests the use of native speaker performance as a reference point in criterion-referenced assessment in general. In the specific context of EAP tests, Weir appears to have been a strong advocate of reference to the native speaker, both in his own development of the associated Examining Board's Test of English for Educational Purposes (TEEP), and in advice about the development of the IELTS test. For example, Weir (1988a) eliminated items in the reading comprehension sub-test of the TEEP test if native speakers found them difficult. Nevertheless the performance of native speakers was found not to be homogeneous: while there were clear differences between the performance of native and non-native speakers, the native speakers achieved scores ranging from 76% to 88% for reading and for writing from 65% to 88%, that is, less than perfect results. The TEEP itself uses an analytical assessment scale of six levels for 'listening comprehension', 'accent', 'formal accuracy', 'referential adequacy', 'sociocultural appropriateness' and 'fluency'. In each case, the top level is described in terms of the user displaying 'native speaker' competence (Emmett, 1985: 145–148).

On the other hand, empirical findings, mainly in the context of semi-direct tests such as cloze, have led to the assumptions implicit this position being questioned. Alderson (1980) found that although there were significant differences between the scores of native and non-native speakers on the cloze test, with the native speakers performing better, the difference was not very great and there was a considerable degree of overlap, with some non-native speaker scores exceeding those of native speakers. Nor did native speaker performance appear to be uniform. Oller and Conrad (1971) had similarly found variability in native speaker performance related to educational level, with significant overlap between advanced non-native speaker performance and that of some categories of native speakers. Bachman (1990) has also criticized the assumption of homogeneity in native speaker performance.

A number of scale developers, perhaps aware of these findings, have been cautious in referring to native speaker performance in scalar descriptions. Among the fence-sitters, Hughes (1988) used the native speaker as a reference point in defining performance levels in an EAP test for a Turkish University (levels include 'Educated Native Speaker Standard' and 'Very Close to Native Speaker Standard'). Nevertheless, elsewhere, in advising the constructors of rating scales, he points out (Hughes, 1989: 110) that the use of a native speaker standard to judge non-native performance has come in for criticism. Nevertheless, such reference frequently implicitly remains. Barnwell (1989) speaks of reference to the native speaker as 'hovering in the background' of the ACTFL Oral Proficiency Interview assessment scale. This is because of the origins of the ACTFL scale in the first and very influential rating scale, the FSI scale, where the highest level (5) is defined as follows:

> *Native or Bilingual Proficiency: Speaking proficiency equivalent to that of an educated native speaker.*
>
> Clark and Clifford (1987: 131)

Other rating scales, for example the ASLPR (Ingram, 1984), and, as we have seen, the one used in the TEEP, are not so cautious. The ASLPR defines its highest level in terms of the performance of native speakers in all four macroskills. For example for writing, the level is defined as follows:

> *W:5 NATIVE-LIKE PROFICIENCY*
>
> *Written proficiency equivalent to that of a native speaker of the same socio-cultural variety. The learner's written language in all its forms is fully accepted by such native speakers in all its features including formal accuracy, structural variation, word choice, idiom, colloquialisms, register appropriateness, discourse structure (including thought sequence and coherence), subtlety of meaning and cultural references. Deviations from educated native speaker forms, special register features, or stylistic conventions will only be those recognizable as native speaker variants.*

> *Can perform as effectively as a native speaker in all writing tasks normally encountered and has native-like flexibility in mastering new ones.*

The test used in this study is the exemplar version of the reading and writing sub-tests of the IELTS test (British Council/UCLES 1989a,b). This test, developed jointly by a British and Australian research team to replace the earlier British ELTS test, is a performance-based EAP test in which the reading and writing tasks simulate those encountered by students in university settings, either as graduates or undergraduates, and by students in more general training contexts. Four versions or modules of the reading and writing sub-tests are available, depending on the candidate's broad area of intended study.

Weir (1988b) called for native speaker performance to be considered in the IELTS test development process, and to a limited extent this was done, using three groups of Sixth Form College students; the results of these trials are available in Clapham and Alderson (forthcoming), and will be discussed below. Further more extensive trails are currently under way. Evans (1990) produced some evidence that native speaker performance on the tests was far from uniform and far from perfect: her native speakers subjects (N=16) at a tertiary institute in Melbourne scored only in the middle range on the IELTS Exemplar Reading Test, at or just below the level required for entrance by foreign students to the institution concerned.

The supposed performance of native speakers is used directly and indirectly in the interpretation of the performance of non-native speakers on the tests. While the reporting scales (Band Scales) for IELTS do not refer explicitly to native speaker performance, avoiding doing so because of the cautions suggested by Alderson, Hughes and Bachman quoted above, the native speaker makes a covert but unmistakable reappearance in the highest Band Scale in the guise of the Expert User, defined as follows: 'has fully operational command of the language; appropriate, accurate and fluent with complete understanding' (British Council/UCLES/IDP, 1989: 14). It is clear, then, that the native speaker 'hovers' over IELTS; and Alderson (personal communication) has stated that the FSI and other rating scales were examined and consulted carefully in the drawing up of the IELTS band descriptors. Reference to native speaker performance is also made in the IELTS reading test specification document for item writers (British Council/UCLES, 1989a: 3), which says that native speakers in the first term of their study should be able to complete the tasks in this test successfully.

In what follows, native speaker performance on the IELTS exemplar test is reported in a series of related studies, two on the reading sub-test, one on the writing sub-test. The performance of educated native speakers of varying levels of post-secondary educational achievement is investigated.

3. Native speaker performance on an EAP reading test

As mentioned above, some trialling of the IELTS reading sub-test was carried out as part of the development of the test, and more is planned. Clapham and Alderson (forthcoming) report the results obtained in the earlier trials. The subjects were just completing the first of a two year course at a sixth form college in Cambridge, England, where they were preparing for the English A-level examinations. The three academic modules of the first live version of the reading sub-test were used. Subjects took the module which was most appropriate for their field of study. Results were as set out in Table 1.

Module	*n*	*raw score mean*	*raw score s.d.*
1 (Arts and Social Sciences)	29	27.52	3.42
2 (Life and Medical Sciences)	16	32.25	4.39
3 (Physical Science and Technology)	10	27.5	3.5

Table 1. Results of Cambridge trials, IELTS reading sub-test (from Clapham and Alderson, forthcoming)

In the studies to be reported below, the Arts and Social Sciences Module is used, so it is worth looking at the data from this module in a little more detail. Table 2 reports frequency data for each raw score level, together with information on how this converts into the Band scale score used for reporting purposes.

None of the native speakers taking this module could be classified as 'Expert User' (Band 9, corresponding to a score of 37); the mean score was barely at the level at which foreign students are admitted to English-medium universities.

In order to investigate these matters further, two studies were carried out at the University of Melbourne with groups of native speakers contrasting in their educational levels, as this had been found to be an important variable by Oller and Conrad (1971) on performance on the cloze test. In broad terms, the first study dealt with students studying at post-secondary level outside the University setting, while the second study examined the performance of graduates. These groups represent groups of native speakers with educational levels respectively lower (Study 1) and higher (Study 2) than the subjects used in the Cambridge trials.

Score	*Frequency*	*Band*
21	1	5.5
22	1	5.5
23	3	6
24	1	6
25	3	6
26	3	6.5
27	1	6.5
28	3	7
29	4	7
30	3	7.5
31	3	7.5
33	3	8

Table 2. Cambridge trials, raw score frequency and band scale equivalent, Arts and Social Sciences module (reading) (data from Clapham and Alderson, forthcoming)

STUDY 1

Hamilton (1991) investigated the performance of 84 native speakers of English on the IELTS reading sub-test. The subjects were all enrolled in post-secondary courses in Colleges of Technical and Further Education (TAFE) in Melbourne, Australia. Such colleges offer a range of non-degree vocational training courses, usually requiring the successful completion of the final year of secondary education, and are the Australian equivalent of British Further Education (FE) Colleges and American Community Colleges. The groups tested were as follows:

Group 1.1	Advanced Certificate of Marketing (N=23)
Group 1.2	Traineeship Certificate in Clerical Skills (N=20)
Group 1.3	Advanced Certificate of Secretarial Studies (N=41)

Entry to such vocational courses usually involves satisfactory completion of the final year of secondary education (Year 12), although students' scores would not in most cases guarantee them entry to a University place directly. Instead, TAFE courses articulate with University degree courses, with credit transfers after successful completion of part of the TAFE course. We can thus say that although these subjects are not in the first rank academically, they have satisfied entry requirements for post-secondary education. The relevance of educational background to performance on the tests is an issue that emerges in the study. The entrance requirements for the courses taken by the groups in this study differ: a higher Year 12 score (Year 12 is the final year of secondary education) is required for entry into the Marketing course (Group 1.1) than into the Secretarial Studies course (Group 1.3); admission to the Clerical Skills course (Group 2) does not require successful completion of Year 12, so the academic standard is likely to be lower again.

As explained earlier, the IELTS reading sub-test comes in four versions, depending on the type of course the candidate is applying for. There are three Academic Modules in three broad subject areas, and a fourth General Training Module, designed particularly to cover the situation of shorter term job-related training attachments in an English speaking country (for example, firemen might come to the UK for updating on modern work methods). There has been considerable debate in Australia over which module is appropriate for the more academically oriented (as against technically oriented) TAFE courses, of which the above courses are examples. The problem is that the choices available are more suited to the British than to the Australian situation. In the latter, government policy is to increase retention in higher education and training substantially, and to blur the academic/non-academic divide, for example by allowing transfer with credits from TAFE courses to degree courses.[2] Recently, IELTS policy has been clarified so that students in professionally oriented TAFE courses (of the kind involved here) should take the relevant Academic Module, rather than the General Training Module, as previously. Inspection of the kinds of texts read by the students in the courses concerned here suggested that they were like those found in the Academic Module. The subjects thus took the exemplar version of Academic Module C, suitable for students in arts, law, business and the social sciences.. This Module is specifically recommended for (among others) candidates seeking entry into marketing/clerical studies programmes (British Council/UCLES, 1989: 1).

The means and standard deviations of scores for the three groups are reported in Table 3. The maximum possible score on the test was 37.

Group	N	M	sd
1.1	23	22.0	5.74
1.2	20	18.5	5.33
1.3	41	17.88	6.28

Table 3. Scores of three groups of native speakers on the IELTS exemplar reading test

The data in Table 3 reveal that the scores of the native speakers were neither homogeneous, nor high. The groups managed to score approximately half marks, the first group a little above that, with

[2] This will increasingly be the case in the UK also, given current UK Government post-school education policy.

a fairly broad range of scores. Moreover, there was a significant overall difference between the group means, as revealed by an ANOVA F (2,81) = 3.73, $p < .05$. Inspection of the means suggested that the mean for Group 1 was contributing most to this difference; this was confirmed by a further t-test, which revealed that the means for Groups 2 and 3 were not significantly different.

STUDY 2

In view of the results of Study 1, and given that the native speakers concerned were of a lower academic level than many students taking the IELTS test, it was decided to replicate the study using a number of highly educated groups. Lopes (1992) investigated the performance of 73 native speakers on the same IELTS Academic Module C exemplar reading sub-test. The subjects were all University graduates with a minimum of one degree. They comprised:

Group 2.1 23 postgraduate students at a teacher training institute, Melbourne, Australia.

Mean age 34.4 (S.D. 7.25) — M 5 F 18.

Group 2.2 30 members of the academic staff at the same institute.

Mean age 48.4 (S.D. 6.35) — M 23 F 7.

Group 2.3 20 Junior Barristers of the Victorian Bar[3], Australia.

Mean age 33.4 (S.D. 6.06) — M15 F5.

The means and standard deviations of the total sample of this study were compared to the combined group of native speakers used in Study 1.

[3] The Australian legal system, like its British source, has two broad categories of lawyers: barristers and solicitors. Barristers represent clients in court, and are briefed for their appearances by solicitors.

	N	*Mean*	*S.D.*
Study 1 (Groups 1.1, 1.2 and 1.3)	84	19.15	6.11
Study 2 (Groups 2.1, 2.2 and 2.3)	73	29.85	4.37

Table 4 . Comparison of performance of native speaker groups in Studies 1 and 2

This comparison indicates clearly that native speakers with higher educational qualifications (i.e. Groups 2.1, 2.2 and 2.3) on average perform better on this test than native speakers with weaker educational qualifications (Groups 1.1, 1.2 and 1.3) and that their scores do not vary to the same extent. No statistical test was carried out as the difference in the means was very substantial and clearly statistically significant.

The performances of the three highly educated groups were then compared (Table 5).

	N	*Mean*	*S.D.*
Group 2.1	23	27.565	4.326
Group 2.2	30	29.6	4.328
Group 2.3	20	32.85	2.477

Table 5. Comparison of Performance of Groups 2.1, 2.2 and 2.3

The data in Table 2 reveal that the means were progressively higher for each group. The variance of scores of the postgraduate student group (Group 2.1) and the lecturer group (Group 2.2) was almost identical even though the mean of the latter group was higher. The Junior Barristers (Group 2.3) had a higher mean again and, differently, a narrower dispersion of scores.

A one-way ANOVA was used to examine the differences between the means of Groups 2.1, 2.2 and 2.3 (Table 6)

Source	*DF*	*Sum Squares*	*Mean Square*	*F-test*	*p*
Between groups	2	301.94	150.97	9.864	.0002
Within groups	70	1071.402	15.306		
Total	72	1373.342			

Table 6. Analysis of Variance table for data from 3 highly educated groups

The above results show that there was a significant difference between the three groups: *F* (2,70) = 9.864, p < .01.

A variance ratio test between Group 2.1 and Group 2.3 was carried out and found to be significant: *F* (22,19) = 3.05, p < .01, thus confirming the difference in the variances in the groups. It is acknowledged that ANOVA requires equal variances of the samples and we have shown that this assumption is violated in this instance; however, ANOVA is known to be robust to violations of this assumption (Hatch and Lazaraton, 1991: 352)

A *post hoc* comparison of means (the Scheffé test) was used to indicate where the differences occurred (Table 7).

	Difference in means	*Scheffé test*
Groups 2.1 & 2.2	-2.035	1.761
Groups 2.2 & 2.3	-3.25	4.141*
Groups 2.1 & 2.3	-5.285	9.76*

* p<.05

Table 7. Comparison between groups of highly educated native speakers

The Scheffé test revealed significance at the 5% level between institute lecturers (group 2.2) and Junior Barristers (group 2.3) as well as between postgraduate students (group 2.1) and Junior Barristers (group 2.3), demonstrating a significant difference in these two populations while confirming that the postgraduate student group (2.1) did not differ significantly from the lecturer group (2.2). Thus, significant differences were fund between the performances of even very highly educated native speakers.

These findings clearly indicate that performance by native speakers on this test is far from uniform and is significantly related to educational level and work experience. We come closest to finding examples of the 'Expert User' in the Junior Barrister group, although by no means all of this group could be so classified; the Expert User is indeed an elusive creature. It is not surprising in view of the close reading and analysis of texts that is characteristic of the work of barristers that they had most success with the reading tasks on the test in question.

4. Native speaker performance on an EAP writing test[4]

Sheridan (1991), in a companion study to that reported in Study 1 above, administered the exemplar version of the IELTS writing subtest, using Academic Module C, as before. The subjects were TAFE students undertaking similar courses to those reported in Study 1 above, that is, students enrolled in courses leading to the Advanced Certificate of Marketing and the Advanced Certificate of Secretarial Studies (cf Groups 1.1 and 1.3 above; there was some, but not complete overlap between the subjects used in each study). 84 students sat for the test but only 62 completed it. Of these, 14 were found to be non-native speakers and 48 native speakers of English; of the native speakers, 32 came from a bilingual background (having exposure to one of eight immigrant languages in the home) and 16 from a monolingual background. All subjects classified as native speakers had completed at least the whole of their secondary education in Australia; most were Australian born; only 4 students had attended school overseas, usually for one or two years at lower primary level.

[4] This section of the paper is based on Sheridan (1991).

The writing sub-test of IELTS presents two tasks: Task 1 involves information transfer or reprocessing (15 mins allowed) and Task 2 requires candidates to draw on information from the passages in the reading sub-test, together with their own experience, to present an argument or to suggest a solution to a problem. The scripts for both tasks were marked by trained and approved IELTS raters. Results are expressed on a 9 point scale in terms of defined band levels.

Before presenting the results, it is worth noting that because of the absence of surface grammatical errors the identity of the writers as native speakers would have been apparent to the raters, who may have applied different criteria or interpreted the set criteria differently as a result; it is possible that a kind of norm referencing may have been going on, despite the fact that raters were ostensibly using the level descriptions in the rating scales. Additionally, as in Study 1, the performance of the NS group may not have been optimal, because of factors relating to test familiarity and motivation. Usually, candidates have had some test preparation, and their performance on the test has relatively serious consequences for them; neither of these things is true for the NS group. It is indicative, for example, that 22 of the native speakers did not complete the writing test through fatigue or lack of motivation, and although their data have not been included, the performance of those who were included may have been affected by these factors.

Part of test	*NS group (N=48)*	
	M	*sd*
Task 1	6.69	1.26
Task 2	6.56	1.37
Overall	6.65	1.36

Table 8. Performance of a group of NS on the IELTS exemplar writing test

The results are presented in Table 8. Mean band levels and standard deviations are given for each task, and for the test overall (there is a procedure for determining the final band score when the scores on the two sub-tests differ).

The mean band score is around Band 6.5, a crucial point on the scale in terms of entry decisions for many courses.

5. Discussion

The results of the study reported here reveal that the performance of native speakers on two IELTS sub-tests was far from homogeneous.

On the *reading* sub-test, differences were found between two broad types of native speaker, those enrolled in non-University post-secondary courses (Study 1) and graduates (study 2). Moreover, within these broad types, significant differences were found in each study between subgroups characterized in terms of educational background (Study 1) and educational and professional background (Study 2). For example, performance on the test was related to the entrance requirements for the courses taken by the groups in question: a higher year 12 score is required for entry into the Marketing course (Group 1.1) than into the Secretarial (Group 1.3) and Clerical Skills courses (Group 1.2). This is reflected in the scores of the three groups, as Group 1.1's score was significantly higher than that of the other two groups. It appears from Study 1 then that educational achievement, as measured in terms of final year secondary school results, is related to performance on this test. In Study 2, differences were found between the performance of Junior Barristers, an academic elite whose reading skills are honed in their training and professional practice, and other graduates.

Looking at the native speaker groups in Study 1 as a whole (N=84), the mean score for the native speakers represented a mark of about 60%. When this is converted into an approximate band score using the conversion table calculated for scores on the 'live' version of the test, this corresponds to a Band 5. A person receiving this score is described as a 'modest user', that is, slightly more than 'limited' (Band 4) and less than 'competent' (Band 6). A more detailed description of the Band is given as follows:

> *Has partial command of the language, coping with overall meaning in most situations, though is likely to make mistakes. Should be able to handle basic communication in own field.*
>
> (British Council/UCLES, 1989: A1)

These results differ from the results of studies of native speaker performance on the TOEFL, where there is a relatively narrow spread of scores among native speakers, who perform significantly better than non-native speakers, and at the top of the scoring range. The reason for this is most likely to be that IELTS tests language skills in context, in performance situations, and TOEFL does not. (Alternative explanations in terms of test-wiseness and motivation for the poor performance of the native speakers should not of course be discounted).

The Swedish study by Oscarson (1986) has been mentioned above. As stated earlier, despite a generally superior performance by native speakers overall, native and non-native speaker performance could not be distinguished on the reading proficiency sub-test. In his discussion of this point, Oscarson (1986: 104) suggested that certain 'non-language-specific variables' such as deductive ability, background knowledge related to the topic, associative memory and reasoning are more important in reading than in other language activities and therefore a smaller difference between native and non-native speaker performance on a reading task would be expected.

Lunzer *et al.* (1979) concluded that effective reading comprehension depends upon two conditions: first, comprehension, and second, the application of appropriate study skills, which include such processes as skimming, scanning, receptive and reflective reading. This distinction suggests that language proficiency is only one factor involved in reading comprehension and that there are other non-linguistic factors involved. As long ago as 1917 Thorndike claimed that in reading the individual must 'select, repress, emphasize, correlate and organize, all under the influence of the right mental set or purpose or demand' (1917 [1971]: 431), such processes being similar to those required to solve a mathematical problem: in other words, 'reading is reasoning'.

Turning now to the *writing* sub-test: the native speakers performed neither homogeneously, nor homogeneously well. As with the reading sub-test, the band descriptors seem ill-matched to the performance of native speakers. The scale and sub-scales used in the actual assessment of the writing sub-test consider 'communicative quality', 'arguments, ideas and evidence' and formal aspects ('word choice, form and spelling' and 'sentence structure'). The second of these categories clearly introduces assessment of study skills rather than language proficiency conceived in some narrower sense; and there is no reason to suggest that native speakers may have any particular advantage in this area. The wording for Level 5 for Task 2 for this category reads as follows:

> *The essay introduces ideas although there may not be many of them or they may be insufficiently developed. Arguments are presented but may lack clarity, relevance, consistency or support.*

One way of understanding the results reported here is in terms of the distinction proposed by McNamara (1990) between a strong and a weak sense of the term *performance test*. In the strong sense, knowledge of the second language is a necessary but not sufficient condition for success on the test tasks. Language is a means, but it is not (or not only) the means which is being investigated. Success is measured in terms of performance on the task, not only in terms of knowledge of language. The results suggest that the IELTS reading and writing sub-tests are performance tests in the strong sense of the term. In this, IELTS differs from TOEFL, which does not claim to be a performance test and measures mainly knowledge of language.

We are not attempting in this paper to adjudicate between the claims of the two tests, only to clarify what is being measured in each case. Arguments in favour of contextualization of language proficiency assessment may be quite compelling, particularly in terms of acceptability to test users (face validity) and effect on preparation for the test (washback). It is clear however that reference to the native speaker as some kind of ideal or bench-mark in scalar descriptions of performance on performance tests is not valid. IELTS does not itself do this, as we have seen, but clearly few of the native speakers in this study fell into its category of 'expert user'; and the relation of 'expert users' to 'native speakers' remains undefined.

Validity, as Bachman (1990: 243) reminds us, is currently understood in terms of the

> *inferences that are made on the basis of test scores*

and quotes the definition of the measurement profession, for whom validity is (American Psychological Association, 1985: 9):

> *The appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores.*

In the case of communicatively oriented EAP tests such as IELTS, if a candidate performs poorly on the IELTS test, it is not clear what inference should be drawn. In what terms should such a performance be explained — is it a problem of language proficiency, or other non-linguistic skills required in the performance task? If we are interested in whether or not the candidate can cope with the tasks expected of her/him in the study situation, it may not matter, except for diagnostic purposes. However, if the candidate is being denied access to training on the basis of a performance equivalent to that of a native speaker who has been accepted for training, then the differing bases for acceptance for the two groups needs to be investigated and perhaps, in the interest of equity and access, made equivalent.

6. Conclusion

In this paper we have examined the one aspect of criterion statements in rating scales, that is, the frequent reference to the presumed performance levels of native speakers. The conclusion must be that more research effort must go into the validation of such rating scales, which are central to the construct validity of the instruments with which they are associated. The lack of clarity in discussions of this topic may be explained by the fact that many of the important rating scales now in use ultimately derive from the FSI scale, developed originally in the 1950s at the height of the psychometric-structuralist period, in which a view of second language proficiency and its relation to first language proficiency gave the native speaker an important defining role as a kind of benchmark. Alderson concluded well over a decade ago (Alderson, 1980: 75) that

> *attempts to use native speakers as a criterion for non-native speakers in ... such criterion referenced testing are misguided. Similarly perhaps proficiency tests should not be validated with native speakers on the assumption that native speakers will achieve perfect scores.*

Despite this, the position of native speaker performance has not been consistently examined, a sign that rating scale descriptors themselves have not been sufficiently examined. Instead, the Chomskyan 'ideal native speaker/hearer' remains the unrecognized point of reference. An examination in fact of the performance of native speakers reveals not only how elusive and untypical (of native speakers) this idealized performance is; such a study also has the potential to clarify what is being measured in communicatively oriented tests.

7. References

Alderson, J.C. (1980) Native and non-native speaker performance on cloze tests. *Language Learning* 30,1: 59–76

Alderson, J.C. and A. Hughes (1981) *Issues in language testing.* ELT Documents 111. London: British Council

Angoff, W.H. and Sharon, A.T. (1971) A comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants to U.S. colleges. *TESOL Quarterly* 5,2: 129–136

Bachman, L.F. (1990) *Fundamental considerations in language testing*. Oxford: OUP

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

British Council/UCLES (1989a) *IELTS specifications.* Cambridge: British Council/UCLES

British Council/UCLES/IDP (1989b) *IELTS users' handbook.* Cambridge: British Council/UCLES

British Council/UCLES (1990) *IELTS specimen materials handbook.* Cambridge: British Council/UCLES

Clapham, C.M. and J.C. Alderson (eds) (forthcoming) *Constructing and trialling the IELTS Test. IELTS Research Report 3.* London: British Council/University of Cambridge Local Examinations Syndicate/International Development Program of Australian Universities and Colleges.

Clark, J.L.D. (1977) *The performance of native speakers of English on the Test of English as a Foreign Language.* Princeton, NJ: ETS

Clark, J.L.D. and R.T. Clifford (1987) Th FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status and needed research. *Studies in Second Language Acquisition* 10,2: 129–147

Emmett, A. (1985) The Associated Examining Board's Test in English for Educational Purposes (TEEP). In P.C. Hauptman, R. LeBlanc and M.B. Wesche (eds) *Second language performance testing.* Ottawa: Ottawa University Press, 131–151.

Evans, R. (1990) *The IELTS: Is it suitable for adult immigrant students who apply for tertiary entrance?* Unpublished MA thesis, University of Melbourne

Hamilton J. (1991) *Native and non-native speaker performance on the IELTS reading test.* Unpublished MA thesis, University of Melbourne

Hughes, A. (ed.) (1988) *Testing English for university study.* London: Modern English Publications/British Council

Hughes, A. (1989) *Testing for language teachers.* Cambridge: CUP

Hughes, A., D. Porter and C. Weir (eds) (1988) *ELTS Validation Project: Proceedings of a conference held to consider the ELTS Validation Project Report. English Language Testing Service Research Report 1 (ii).* London: British Council/University of Cambridge Local Examinations Syndicate

Ingram, D.E. (1984) *Report on the formal trialling of the Australian Second Language Proficiency Ratings.* Canberra: Australian Government PublishingService.

Johnson, D.C. (1977) The TOEFL and domestic students: conclusively inappropriate. *TESOL Quarterly* 11,1: 79–86

Lopes, M. (1992) *Native speaker performance on the IELTS reading test.* Unpublished MA thesis, University of Melbourne.

Lunzer, E., M. Waite and T. Dolan (1979) Comprehension and comprehension tests. In E. Lunzer and K Gardner (eds) *The effective use of reading.* London: Heinemann Educational Books, 37–71

McNamara, T.F. (1990) *Assessing the second language proficiency of health professionals.* Unpublished Ph.D. thesis, University of Melbourne

McNamara, T.F. and J. Hamilton (1991) *EAP tests as performance tests.* Paper presented at the 4th ELICOS Association Educational Conference, Monash University, August.

McNamara, T.F., J. Hamilton and E. Sheridan (1992) *Rating scales and native speaker performance on a communicatively oriented EAP* test. Paper presented at the 14th Language Testing Research Colloquium, Vancouver, February 27th–March 1st

Oller, J.W. Jr and C. Conrad (1971) The cloze technique and ESL proficiency. *Language Learning* 21, 2: 183–195.

Oscarson, M. (1986) *Native and non-native speaker performance on a national test of English for Swedish students. A validation study.* Goethenberg: Goethenberg University, Department of Educational Research

Sheridan, E. (1991) *A comparison of native/non-native speaker performance on a communicative test of writing ability (I.E.L.T.S.)* Unpublished MA thesis, University of Melbourne

Thorndike, E.L. (1917 [1971]) Reading as reasoning: a study of mistakes in paragraph reading. *Journal of Educational*

Psychology 8: 323–332. Reprinted in *Reading Research Quarterly* 6,4: 425–434 (1971)

Weir, C. (1988a) The specification, realization and validation of an English language proficiency test. In A. Hughes (ed.) *Testing English for university study.* London: Modern English Publications/ British Council, 45–110

Weir, C. (1988b) Construct validity. In A. Hughes, D. Porter and C. Weir (eds) *ELTS Validation Project: Proceedings of a conference held to consider the ELTS Validation Project Report. English Language Testing Service Research Report 1 (ii).* London: British Council /University of Cambridge Local Examinations Syndicate, 15–25.