Lingue e Linguaggi
Lingue Linguaggi 6 (2011), 7-20
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i2239-0359v6p7
http://siba-ese.unisalento.it, © 2011 Università del Salento

FLUENCY AND PRONUNCIATION IN THE ASSESSMENT OF GRAMMATICAL ACCURACY IN SPOKEN PRODUCTION: An Empirical Study

THOMAS CHRISTIANSEN

1 Introduction

Due to the nature of spoken discourse, spontaneous and transient, debate continues about what precisely to assess in speaking tests (see Luoma 2004, North 1991, North and Schnieder 1998) and there is little research into how to assess it (see Hubbard, Gilbert and Pidcock 2006).

In this paper, we will look at how far different areas of performance may affect each other. The method we adopt involves the manipulation of data to observe how changes in pronunciation and fluency affect assessment of grammatical accuracy (henceforth GA), the wider issue being how well assessors manage to separate distinct elements of performance as required by the use of analytic scales (see Luoma 2004:62).

Of the different areas conventionally identified for assessment, pronunciation and fluency would seem to be the two where there is least overlap with GA. It would be possible, in theory, for a learner to display a low level of pronunciation or fluency but a high degree of GA or vice versa. In practice however, striking variation between different areas of performance is uncommon. Typically, a high level of punctuation or fluency will seem to correlate with a high level of GA, it being, however, impossible to determine in which direction the relationship runs: whether the one causes the other, or whether they are both caused by some third factor. Any correlation is therefore complex as is shown by the fact that learners seem to acquire pronunciation at a different rate from other skills[1].

Looking at two variables at once, pronunciation and fluency, certainly complicates research but is unavoidable as it is very difficult, both in practice and in theory, to separate the two, as indeed it is to separate fluency in a general sense from other assessment criteria[2].

Definition of fluency is itself problematic (see Fulcher 1987, Hasselgreen 2005). Lennon (1990) distinguishes between the term's broad and narrow sense. The former has been defined as "[…] the ability to express oneself intelligibly […] reasonably accurately and without too much hesitation." (Byrne 1986:9), the latter broken down into quantifiable factors such as: words per minute; repetitions; self-corrections; filled pauses; etc. (see Hughes 2002:113 for a complete list). Defining fluency precisely can therefore become highly technical; Fulcher (1993, 1996) provides 200-word descriptors for fluency at different levels. Furthermore, even simple objective criteria can prove troublesome to apply, as shown by Griffiths (1990:331-32), who notes the inability of speakers to reliably detect differences in speech rate.

[1] See the CEFR descriptors for phonological control where no distinction is made at the upper end of the scale between C1 and C2 levels.

[2] Indeed, revealingly, few assessment scales make recourse to fluency per se - see Luoma 2004.

In the normal practice of assessment, various factors contribute to make it harder for people to assess fluency: perception; the limits of the mind to register relevant information; and incorrect assumptions of what the norm is in terms of average native speaker performance[3].

With these observations in mind, the general hypothesis we will seek to test is the following:

> **Hypothesis 1.0**
> Fluency – *narrowly-defined* – and Pronunciation can affect the way a speaker's GA is perceived by listeners.

This hypothesis does not specify in what way other aspects of performance may be affected. To clarify this, we examine two separate hypotheses which are subordinate to Hypothesis 1.0.

> **Hypothesis 1.1**
> A high level of Fluency – *narrowly-defined* – and Pronunciation can lead assessors to **over**estimate the speaker's level of GA.

> **Hypothesis 1.2**
> A high level of Fluency – *narrowly-defined* – and Pronunciation can lead assessors to **under**estimate the speaker's level of GA.

Hypotheses 1.1 and 1.2 are not mutually exclusive, as they each could hold in different contexts. If both or either of Hypotheses 1.1 and 1.2 are found to be true, then it indicates that Hypothesis 1.0 is true. If only one is found true, Hypothesis 1.0 will be judged too general. If neither is found to be true, Hypothesis 1.0 will be definitely refuted.

2 Method

To test the hypotheses set out in Section1.0, a survey was conducted asking respondents to assess six short extracts of simulated speaking tests. One of the extracts at random was replaced by another in which the utterances were the same but the original voice (OV) had been replaced by a computer-generated (CG) one. The CG extracts were produced by making transcripts of the OV and using the commercially-available text-to-speech program *Text Aloud* (by NextUp.com).

Recordings of six OVs were used, two of whom were L1 Italian, two L1 German, one L1 Polish and one L1 British English. Three were male, three female. The same female US English CG voice ("Heather" by Accapela) was used for the CG versions of the five non-native speaker OVs. The native speaker OV was replaced by a female Latin American Spanish CG voice (AT&T Labs' *Natural Voices* "Rosa")[4]. Both the CG voices

[3] The central problem is: "Speaking is not naturally language-based, rather it is people focused. Language testing is not naturally people focused but by its nature will tend to be language focused. To give an example, many native speakers are extremely hesitant in their speech delivery, particularly when being asked to do something such as form an opinion 'on the hoof'. However, the listener will not tend to notice the pausing, um-ing and er-ing in their interest to hear the answer to their question or the opinion of the speaker." Hughes (2002: 77).

[4] We might have used the same CG voice for the native speaker voice, as this would have provided us with data to test the effect of the artificial CG voice on the assessor against the natural voice. However, we concluded that presenting assessors with a text displaying a native-speaker level of proficiency in GA but

used were powerful applications in themselves, rendering the intonation of the whole clause in a clear manner, as opposed to just the individual word in a robotic fashion. Audio-editing software was also used with the CG voices in an effort to make them sound as natural as possible.

At the transcription stage, we included all vocal sounds that could be rendered convincingly by the CG voice (e.g. ‘er’, ‘um’, ‘oh’, but not laughs). Fluency of the CG voice was thus increased in terms of speed but not in number of filled pauses (Lennon 1990) or *smallwords* (Hasselgreen 2005), except in the case of Speaker 6 CG (see below). Such vocalisations and fillers also have an effect on fluency, but to manipulate them would have been a delicate task beyond the scope of this present research. Table 1 shows the different timings between the original extracts and the CG versions:

	Speaker 1	**Speaker 2**	**Speaker 3**	**Speaker 4**	**Speaker 5**	**Speaker 6**
Original	114 secs	94 secs	101 secs	128 secs	117 secs	76 secs
CG	82 secs	56 secs	52 secs	99 secs	75 secs	105 secs
Difference CG to OV	- 32 secs	- 38 secs	- 49 secs	- 29 secs	- 42 secs	+ 29 secs

Table 1
Difference in time (seconds) between original extracts and those generated using CG voices

Only with Speaker 6 (the native speaker) was the CG version longer. Here the object had been to make the CG version less fluent. As well as using a slower speech rate, we added fillers such as ‘er’ or ‘um’, but in an as objective fashion as possible: at more or less regular intervals (every three content words).

The completed extracts were then posted on the internet[5] in the form of six randomly-accessed surveys. In each of these surveys, the respondent heard five OVs and one CG voice. In the first survey (labelled A), Speaker 1 was substituted by a CG voice, in survey B, Speaker 2, and so on. Within the surveys, the order of the extracts was then randomized.

Respondents to the survey were asked to assess each speaker only for GA, using descriptors adapted from the *Common European Framework of Reference* (CEFR 2001), as shown in Table 2:

▼ CEFR Level		Mark▼
C2	Student maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning).	8
C1	Student maintains a high degree of GA; errors are difficult to spot.	7
B2	Student shows good grammatical control; occasional 'slips' and minor flaws in sentence structure may still occur, but student often corrects these him- or herself.	6
	Student shows fairly good grammatical control. He/She does not make mistakes which lead to misunderstanding.	5
B1	Student can participate in discourse by communicating with reasonable accuracy; He/She shows generally good control though with noticeable mother tongue influence. Errors occur, but it is still clear what he/she is trying to say.	4
	Student can construct a variety of structures related to familiar topics. He/She uses a repertoire of commonly-used structures reasonably accurately.	3

continues

a noticeably lower level in pronunciation and fluency provided an interesting context in which to test hypothesises 1.0 – 1.2, in a reverse direction to that which were attempting to do with the other extracts.

[5] Setting up the internet site and writing the software to manage the randomised surveys, was the work of Alberto Milone, then a student at the Università del Salento (Lecce, Italy).

Table 2 continued

▼ CEFR Level		Mark▼
A2	Student can use a limited series of phrases and sentences to describe things and events in simple terms. Uses some but not all simple structures correctly, but still systematically makes basic mistakes – for example mixes up tenses; nevertheless, it is usually clear what he/she is trying to say.	2
A1	Student shows only limited control of a few simple grammatical structures and sentence patterns, and relies on a learnt repertoire.	1

Table 2
Descriptors for GA (adapted from CEFR, 2001: 114)

The descriptors cover the full range of the CEFR from A1 up to C2. Respondents were asked to score each performance according to the scale 1-8 but were not informed of the precise link between the mark and the corresponding CEFR level.

3 Survey sample

The period of the research ran from January to February 2008. In that time 40 respondents completed the surveys. Consequently, for each OV, there were 32 - 34 different scores to compare and for each CG between 6 and 7.

Before starting, respondents were asked to complete a short questionnaire. This provided background information for subsequent analysis. Table 3 summarises this:

Question 1 What is your native language? / What are your native languages?	**Question 4** How long have you been a teacher of English to Speakers of Other Languages?
English = 34; Italian = 8; English / Italian = 2; English / French; English / German; English / Mandarin Chinese; Albanian; Polish = 1 each (L1 English = 29: L1 Other = 11)	• Less than 1 year **= 0** • Between 1 and 3 years **= 3** • Between 3 and 10 years **= 13** • More than 10 years **= 24**
Question 2 Where do you currently live and teach?	**Question 5** Which different countries have you taught in?
Italy = 34; UK = 4; Australia; Albania = 1 each	**1 country = 13; 2 countries = 15; 3 countries = 8; 5 countries = 3; 7 countries = 1**
Question 3 What kind of teaching qualification(s) do you have? Please specify.	**Question 6** Have you ever received training as an oral examiner? Please give brief details:
Qualified = 38; Not Qualified = 2	**Trained OE = 20; Not trained OE = 20**

Table 3
Pre-survey questionnaire and results

As can be seen, typically respondents were native speaker English, living in Italy and had over 3 years teaching experience in a couple of countries, and had some teaching qualification. Fifty percent of the respondents had also been trained as oral examiners, mostly for Cambridge ESOL.

4 Results

In Table 4, the marks of the 40 respondents are summarised, giving the mean scores for each speaker, and the standard deviations (SD) of the same:

	Mean	SD		Mean	SD
Speaker 1 OV	5.32	1.39	Speaker 1 CG	3.00	2.76
Speaker 2 OV	5.36	1.54	Speaker 2 CG	4.00	3.15
Speaker 3 OV	3.03	1.20	Speaker 3 CG	2.63	2.62
Speaker 4 OV	5.76	1.48	Speaker 4 CG	4.50	2.48
Speaker 5 OV	5.03	1.85	Speaker 5 CG	5.33	2.25
Speaker 6 OV	7.58	1.03	Speaker 6 CG	4.00	1.83
Mean value	**5.35**	**1.41**	**Mean value**	**3.91**	**2.51**

Table 4
Summary of marks for OV and for CG extracts

There is a marked difference between the average scores given for the OVs (5.35) and for the CG voices (3.91). This supports Hypothesis 1.2 as the voices which presumably display better pronunciation and fluency receive lower marks for GA. However, with Speaker 5, the GA of the CG voice is marked higher than that of the OV and thus supports Hypothesis 1.1. Similarly, for Speaker 6 OV, the GA of the CG voice (non-native pronunciation and fluency) receives lower marks than the OV (native speaker)

It is significant that there is a higher standard deviation for CG voices than for the OVs, reflecting less agreement among respondents. This may indicate that it is harder to assess the GA of the CG voice, one explanation for which being that the CG versions are perhaps less intelligible.

4.1 Intelligibility

To investigate the issue of how intelligible the CG voices were, we set up a reference group of six expert informants (all experienced ESOL professionals, L1 English but based in Italy) and asked them to each assess the twelve extracts using a custom-written set of descriptors (Table 5). Members of the reference group had not done the survey and were informed that half the extracts used CG voices.

	Mark▼
It is possible to recognise **all** the words uttered by the speaker and **every one** is **unambiguous** and **clear**.	6
It is possible to recognise **nearly all** the words uttered by the speaker but **one or two** are **ambiguous** or **unclear**.	5
It is possible to recognise **most** of the words uttered by the speaker but **a few** are **ambiguous** or **unclear**.	4
It is possible to recognise **many** of the words uttered by the speaker but **many** are **ambiguous** or **unclear**.	3
It is possible to recognise **some** of the words uttered by the speaker but **most** are **ambiguous** or **unclear**.	2
It is possible to recognise **only a few** of the words uttered by the speaker and **all but a few** are **ambiguous** or **unclear**.	1

Table 5
Intelligibility descriptors

The results of this restricted survey are given in Table 6, and below that in a graph, Figure 1, where they are compared with the marks given in the general survey for GA:

	Mean	SD		Mean	SD
Speaker 1 OV	5.17	0.41	Speaker 1 CG	4.33	1.21
Speaker 2 OV	5.50	0.84	Speaker 2 CG	4.50	1.38
Speaker 3 OV	4.17	0.75	Speaker 3 CG	4.33	1.63
Speaker 4 OV	5.67	0.52	Speaker 4 CG	4.00	1.67
Speaker 5 OV	4.83	0.75	Speaker 5 CG	4.33	1.37
Speaker 6 OV	5.83	0.41	Speaker 6 CG	2.50	0.84
Mean value	**5.20**	**0.61**	**Mean value**	**4.00**	**1.22**

Table 6
Summary of marks for intelligibility of OV and of CG extracts

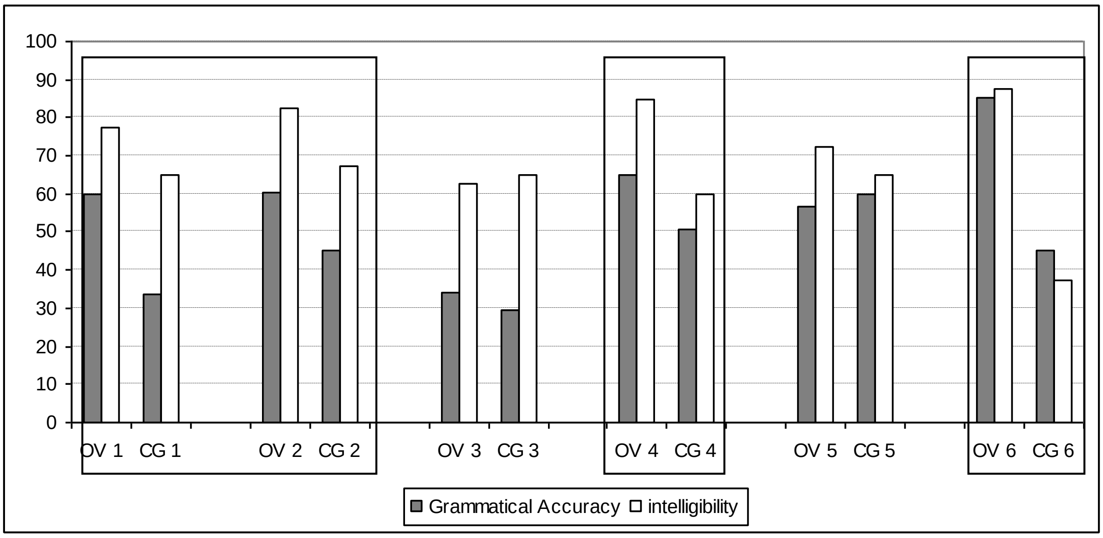

Figure 1
Marks awarded for GA compared with marks for pronunciation

With the exception of Speaker 3, OVs were judged more intelligible than their CG counterparts. Again, as seen in Table 4, there was more agreement among respondents about the marks for the OVs than there was for CG voices. From Figure 3, it can be seen that, for Speakers 1, 2, 4 and 6, there is an apparent correspondence between intelligibility and the pattern of marks for GA for OV and CG voices.

That said, except for Speaker CG 6, all extracts scored an average of over 4, so none were so unintelligible so as to affect assessment of GA. Speaker CG 6 has a low average of 2.5 (and a low standard deviation indicating greater agreement among respondents) which suggests that intelligibility, or lack of, does affect the score it received for GA. Speaker 5 is also interesting, as here the CG voice receives higher marks for GA in the general survey, even though the OV is judged by the reference group to be more intelligible. Such an inconsistency is possibly due to the fact that intelligibility is a somewhat subjective criterion which, like general fluency, is not an aspect of performance in itself but rather an amalgam of various factors.

4.2 Pronunciation and Fluency

To examine effect of pronunciation and fluency on the marks given, the reference group assessed all 12 recordings for pronunciation (using the CEFR descriptors for phonological control, 2001:117) and fluency (CEFR, 2001:114). The scales for pronunciation go from 1 (A1 level) to 5 (C1 / C2), those for fluency from 1 (A1) to 9 (C2). In order to compare these scores with those for GA (1-8 marks), the figure 90 serves as a common base: the marks for pronunciation multiplied by 18, those for fluency by 10, and those for GA by 11.25.

Looking first at pronunciation, here again the standard deviation for OVs was considerably lower than that for CG voices (0.61 as opposed to 1.35). In Figure 2, the marks given in the general survey for GA are compared with those of the reference group for pronunciation.

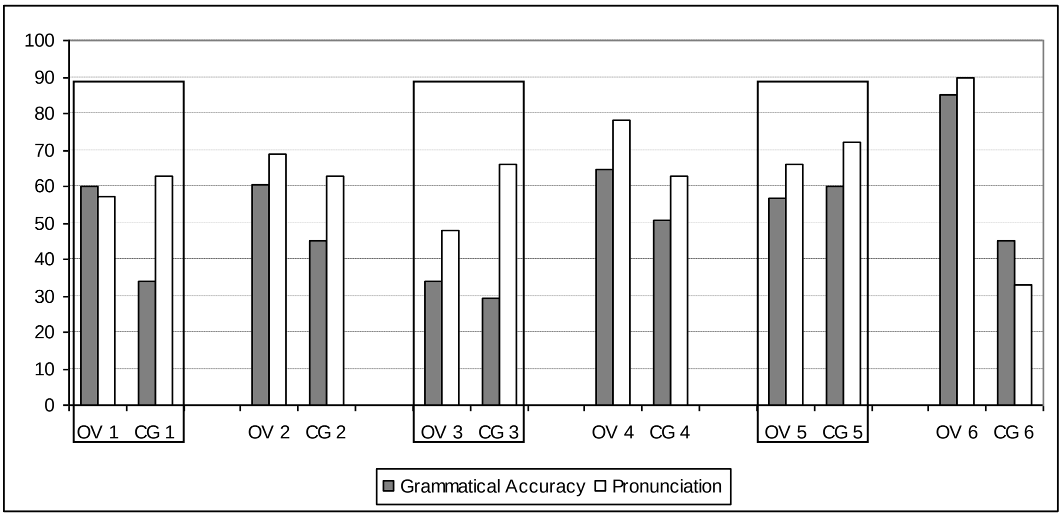

Figure 2
Marks awarded for GA compared with marks for pronunciation

Only with Speakers 1, 3 and 5 did the reference group prefer the pronunciation of the CG. Given that the same voice used for 5 of the CG versions, the differences in marks must be attributable to intonation and prosody, not individual sounds.

Overall, there is some erratic correspondence between the marks for GA and pronunciation for Speakers 4 and 5. In each case, the pattern of marks for the OV and CG is very similar both for GA and pronunciation. Speaker 6 is also similar but there is a much greater difference between the scores for pronunciation for OV and CG than there is for GA, but intelligibility may be a factor here (Section 4.1).

The patterns for Speakers 1-3 are apparently more random. With Speaker 1 there is little difference between the scores for pronunciation but a large one between those for GA. By contrast for Speaker 3, there is more difference between the marks for pronunciation than there is for GA. Speakers 2 and 5 show a similar level of GA and pronunciation but the pattern of marks for each category is markedly different.

None of this seems explainable in terms of the OV accent; both groups contained a mixture of accents: Speakers 1-3, where no possible correlation is found, two L1 Italians and one L1 German; Speakers 4-6, where a possible correlation is found, one L1 German, one L1 Polish, and one L1 English.

Turning to fluency, Figure 3 replicates Figure 2 inserting the data for Fluency in place of those for pronunciation.

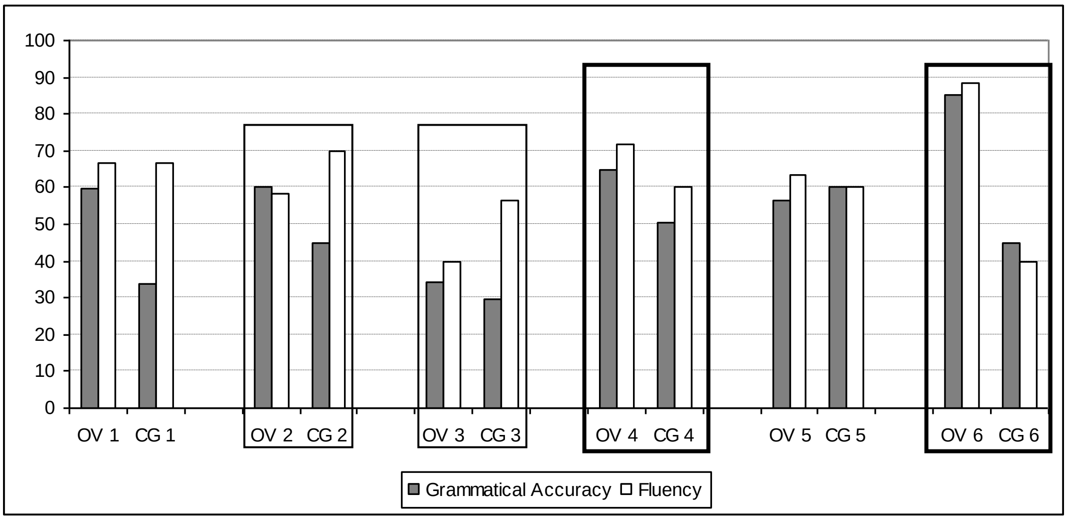

Figure 3
Marks awarded for GA compared with marks for fluency

Generally, there is even less correspondence between the scores for GA and fluency than there is between GA and pronunciation, although the patterns between GA and fluency for Speakers 4 and 6 are similar. Only with Speakers 2 and 3 is the CG voice judged more fluent than the OV. The standard deviation is again considerably larger for the CG version (2.3 vs. 0.96 for the OV).

The search for a direct correlation between level of either pronunciation or fluency and marks awarded for GA has thus proved inconclusive. In individual cases, links can be posited but no general pattern emerges that would enable useful generalisations. In particular, it is difficult to establish whether a high level of fluency and pronunciation consistently results in assessors' overestimating or underestimating the speaker's GA. Again, only Hypothesis 1.0 is confirmed.

4.3 Marks according to category of respondent

It has not been possible to explain why assessors give different marks for GA to OV and CG voices in terms of pronunciation or fluency. Perhaps, the answer lies in some aspect of voice quality that is more subjective than either pronunciation or fluency as defined in the descriptors used for the reference group, that is to say the general impression that the speaker conveys to the assessor. To investigate this point it is useful to look not just at the marks given but also at those giving the marks: the respondents. This will first require dividing the respondents into categories and then deciding which of these categories are significant as regards marking behaviour.

Figure 4 uses data from Table 3 to show how category of respondent affects marks given (excluding teaching qualifications as too few respondents defined themselves as unqualified):

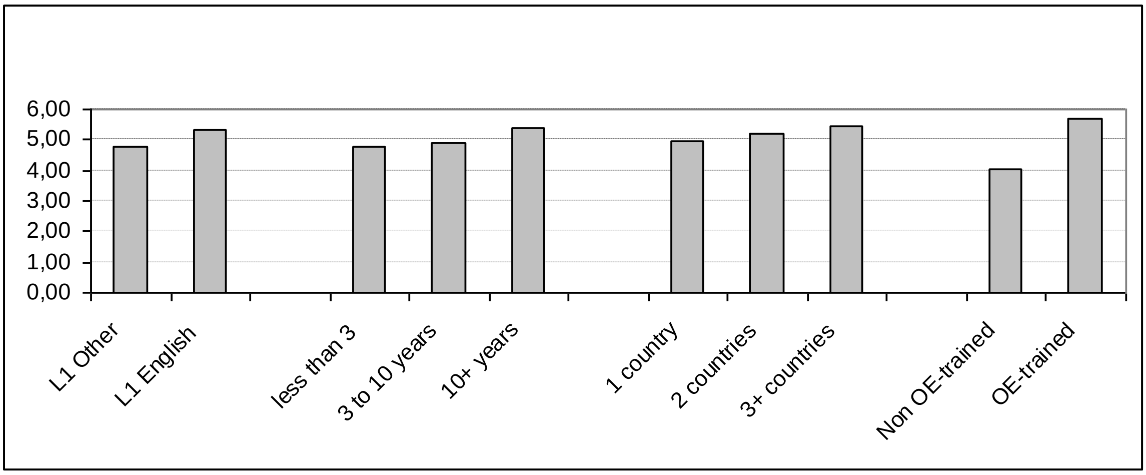

Figure 4
Average marks according to category of respondent

Figure 4 identifies three trends. Firstly, L1 English speakers tend to mark higher than L1 Other. Secondly, teachers with more experience (either in terms of years or of number of countries in which they have worked) mark progressively higher. Finally, respondents who have received oral examiner training (OE) mark higher than those who have not (Non-OE).

For statistical purposes, there is a need to calculate the probable significance of each category by estimating how likely finding similar results as shown on Figure 4 would be if this research was repeated with different respondents. A contingency table can be used to compare observed with expected frequencies, calculating from the difference between them the so-called χ^2 (chi square) statistic.

Putting the data from L1 English and L1 Other from Figure 4 into a contingency table, the following results are obtained (Table 7):

Marks	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total O**
L1 ENG O	4	17	12	27	27	25	33	29	174
L1 ENG E	*7.25*	*18.9*	*11.6*	*28.3*	*23.2*	*27.6*	*29.0*	*28.3*	
L1 Other O	6	9	4	12	5	13	7	10	66
L1 Other E	*2.75*	*7.15*	*4.4*	*10.7*	*8.8*	*10.4*	*11*	*10.7*	
Total O	10	26	16	39	32	38	40	39	240
χ^2 = 11.4	**Probability (Null Hypothesis) = 0.122 (12.2%)**				**Probability of significance = 87.8%**				

Table 7
Contingency table L1 English speaker versus L1 Other

The values found in our survey are marked as O (observed); those expected if the null hypothesis were true as E. Figures relate to the frequency of different marks for GA (for example L1 English award the mark 1 four times, mark 2 seventeen times etc.). As can be seen, after calculating the χ^2 statistic, it transpires that the null hypothesis has a 12.2% likelihood of being correct, so it is only 87.8% probable that there is significance in the difference between marks awarded by L1 English speakers and L1 Other speakers. This is below the figure cited conventionally (95%) as sufficient for significance, and thus there is not enough statistical evidence to maintain that the categories of L1 English and L1 Other are significant in our data.

Similar calculations, using the data for the other categories outlined on Figure 4, show that experience in years had only between a 54.8% and 77.5% probability of being significant (depending on whether the comparison was between -3 and 3+ years, or -10 and 10+ years), experience in terms of countries taught only 56.7%. The one category that did prove statistically significant was whether a respondent was OE or Non-OE (95.7%). Indeed, in Figure 4, this is the category where a difference is most marked. Intuitively as well, this makes sense as whether a respondent is OE or Non-OE would be the best indication of whether he or she had experience of using descriptors of the kind used in our survey.

The figure for L1 English versus L1 Other, while below the conventional benchmark for significance, is still high relative to the other categories, and, in the case of Speaker 6, the difference between L1 English and L1 Other does intuitively seem significant (Figure 5).

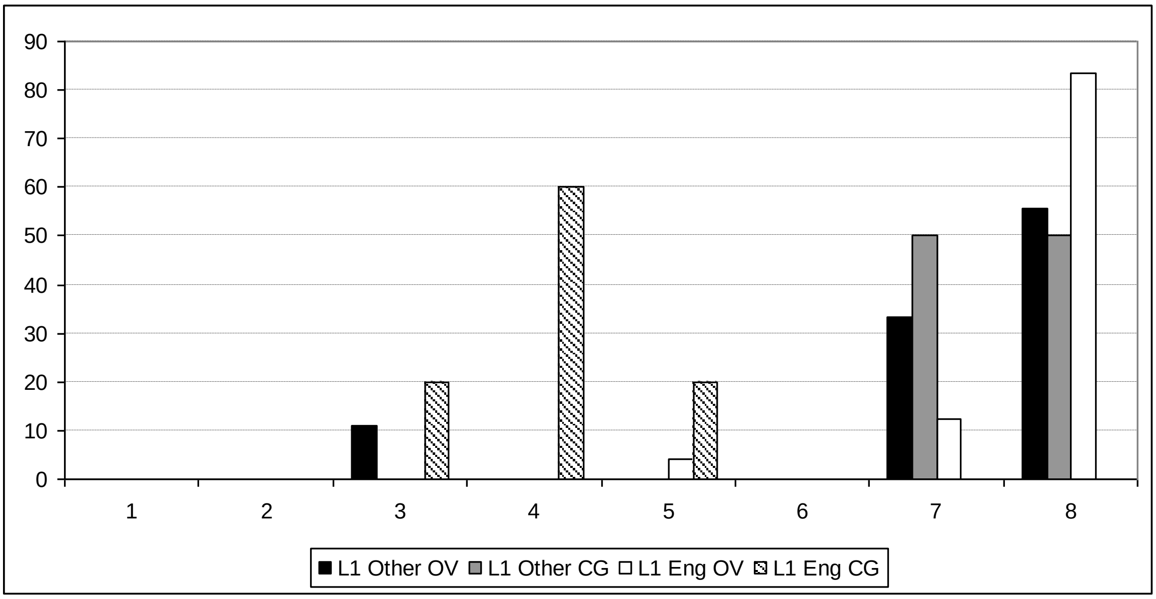

Figure 5
Percentage of marks for GA (1-8) given to Speaker 6 (OV and CG) by L1 English respondents and by L1 Other respondents.

As can be seen on Table 3, there are 29 L1 English, and 11 L1 Other. Twenty-four of the L1 English assessed Speaker 6 OV and five Speaker 6 CG; the figures for L1 Other are 9 and 2 respectively. The values shown on Figure 5 are percentages of the total for that category for that speaker: 11.11% of the eight L1 Other respondents (i.e. one person) gave Speaker OV 6 only 3 for GA, while the other eight (88.89%) gave either 7 or 8. Where there is a major difference between L1 English and L1 Other is with Speaker 6 CG. Twenty per cent of the L1 English give it a 3 for GA, 60% a 4 and 20% a 5. The two L1 Other give a 7 and an 8. As regards intelligibility[6], they do not find the non-native speaker accent of the CG voice as off-putting as their L1 English counterparts. Alternatively, it is possible that L1 English speakers, recognising Speaker 6 CG as a non-native speaker, marked GA low based on the expectation that it would be. The CG voice's relative lack of intelligibility may also have discouraged them from assessing GA as carefully as they might.

[6] This raises an interesting question about the concept of intelligibility and native and non-native speakers' perceptions of the same. Unfortunately, in our reference group (see Sections 4.1 and 4.2), we did not think of including any non-native English speakers, so are unable to investigate this point here.

Beyond this specific case, statistics show that it is only with the categories of OE and Non-OE that significant differences are found in results. In Table 8, the marks given by these two groups for the various speakers are summarised. Below that, the average marks are plotted on a graph (Figure 6).

	Mean OE	Mean Non-OE	SD OE	SD Non-OE
Speaker 1 OV	6.06	4.67	1.39	1.03
Speaker 2 OV	5.72	4.93	1.36	1.67
Speaker 3 OV	3.60	2.53	1.06	1.12
Speaker 4 OV	5.71	5.82	1.36	1.63
Speaker 5 OV	5.53	4.53	1.77	1.84
Speaker 6 OV	7.76	7.38	0.75	1.26
Mean value OV	**5.73**	**4.98**	**1.28**	**1.42**
Speaker 1 CG	2.50	4.00	2.38	4.24
Speaker 2 CG	4.50	4.40	4.95	2.97
Speaker 3 CG	4,00	2.33	2.83	2.31
Speaker 4 CG	5.33	4.33	2.08	3.21
Speaker 5 CG	7.33	4.00	0.58	2.00
Speaker 6 CG	3.67	6.00	0.58	1.83
Mean value CG	**4.59**	**4.18**	**2.23**	**2.76**

Table 8
Summary of marks awarded by OE-trained and Non-OE-trained respondents for OV and for CG voices

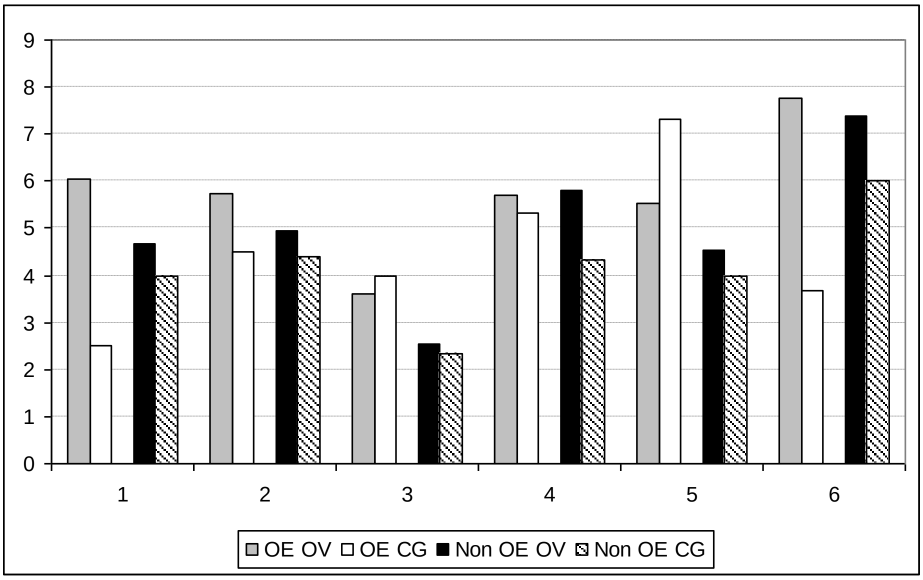

Figure 6
Comparison of average marks awarded for each speaker by OE-trained respondents and Non-OE-trained ones

Like Figure 4, Table 8 shows that OEs give higher marks. The lower standard deviations show that, as a group, they also agree more. In general, in both categories, there is more variation for CG voices. For OE, there is much lower than average standard deviation (0.58) for Speakers 5 CG and 6 CG even compared to OV, indicating more than usual agreement. With Speaker 6, there is a marked difference in scores given by OE and Non-OE but less so than between L1 English and L1 Other. The L1 difference would thus seem more significant in this specific case, as it also transpires that all 20 of the OE were L1 English while only 9 of the Non-OEs were.

Generally, the pattern of marks for each pair of recordings, OV and CG, is markedly different for OE and Non-OE. While Non-OE consistently award lower marks to CG (more pronounced with Speakers 4 and 6), for OE, the pattern is less easily characterised. They give higher marks only to OV Speakers 1, 2, 4 and 6.

It would seem overall that Hypothesis 1.2 holds more for Non-OE, although the extension of this trend to Speaker 6 actually conforms to Hypothesis 1.1, but here, as already shown, intelligibility is relevant.

With OE, it would seem that both Hypotheses 1.1 and 1.2 hold with a slight preference for the former (four out of the six speakers, including the problematic Speaker 6). In the next conclusion, we will discuss the implications of this and then suggest areas for further research.

5.0 Conclusions

From the results of the survey, the general Hypothesis 1.0 was confirmed with the more specific Hypothesis 1.2 holding more for Non-OEs than OEs, and Hypothesis 1.1 only in a few cases, such as Speaker 6, and for no particular category of respondent.

That Hypothesis 1.2 is confirmed in more cases than Hypothesis 1.1 might be because the higher level of pronunciation and fluency of the CG versions actually makes deficiencies in GA clearer. It is possible that Non-OEs have greater difficulty identifying errors in the less fluent discourse of a speaker where there is also less phonological control. As we saw in Section 4.1, intelligibility as such does not seem to be the problem (although as we have noted, perhaps it is relevant that the Non-OE group included all the L1 Others, and the reference group measuring intelligibility contained only L1 English), but rather one of overall effect of the CG voice on the listener.

Apart from specific figures and hypotheses, this study shows that it is indeed difficult to assess one aspect of performance in complete disregard of other areas, as analytical mark scales demand. However, results show that the tendency to confuse supposedly separate aspects of performance does not affect all assessors in the same way or to the same degree. This suggests that such a thing is not inevitable and that, by seeking to understand, through studies such as this one, the causes of the phenomenon with assessors from different backgrounds, remedies may be found that may be incorporated into oral examiner training. For example, that L1 Other respondents seem more able to deal with the non-native accent of Speaker 6 CG than L1 English do invites one to examine why exactly this is so in the hope of identifying factors that may be useful to L1 English assessors.

The data produced by this small-scale experiment raise a multitude of questions, each of which deserves fuller attention that we have been able to devote here: something which shows the potential of the type of interventionist approach adopted. For example, it would be interesting to conduct similar studies in languages other than English, especially ones where the conventional perceptions of correct and incorrect pronunciation hold more firmly than in English, which, having become a global language, no longer has one standard (see Crystal 2003).

The research method devised could be used to investigate numerous other variables and aspects of speaking tests. Regarding corroboration, the experiment could be replicated using different extracts and different respondents or developed using longer extracts or indeed substituting the CG voices with another human voice, ideally a trained actor, or voice-changing software could be considered. Both of these alternatives may serve to validate or otherwise the practice of using CG voices.

The issues raised when comparing genuine to artificial voices permits us to view from a completely new perspective the complexity of the processes behind assessment: in particular, what factors make a voice sound natural[7] and how far this may be allowed to contribute to assessment. The fact that the CG voices proved problematic for some respondents to assess[8] nonetheless reconfirms that assessment of one area of performance can be affected by factors that, strictly speaking, relate to quite different areas, which is the premise behind Hypothesis 1. Finally, problems encountered with any method can be in themselves illuminating and may be seen as an opportunity for further research; investigation is invited into what it is in the elusive processes of assessment which is adversely affected and how, as it is often only when there is a malfunction somewhere in the system that one can observe the hidden mechanisms underlying a complex process.

[7] No doubt this is an issue that occupies the developers of speech-to-text programs, and it would be interesting to know how they have approached the problem.

[8] While many respondents no doubt did not notice anything odd about the CG voices, it is clear from unelicited personal communication that more than a few respondents did.

References

Byrne, D. 1986, *Teaching Oral English*, Longman, London.

Council of Europe 2001, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge University Press, Cambridge.

Crystal D. 2003, *English as a Global Language*, Cambridge, Cambridge University Press.

Fulcher, G. 1987, *Tests of oral performance: The need for data-based criteria*, in "ELT Journal, 41 (4)", pp. 287-291.

Fulcher, G. 1993, *The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language*, unpublished Ph.D. thesis, Lancaster University, Lancaster, UK.

Fulcher, G. 1996, *Does thick description lead to smart tests? A data-based approach to rating scale construction*, in "Language Testing 13, 2", pp. 208 - 238.

Griffiths, R. 1990, *Speech Rate and NNS comprehension: A preliminary study in time-benefit analysis*, in "Language Learning, 40(3)", pp. 311-336.

Hasselgreen, A. 2005, *Testing the Spoken English of Young Norwegians: a study of test validity and the role of 'smallwords' in contributing to pupils' fluency,* Cambridge University Press, Cambridge.

Hubbard, C, Gibert, S and Pidcock J. 2006, *Assessment processes in Speaking tests: a pilot verbal protocol study*, in "Research Notes", University of Cambridge ESOL Examinations, issue 24, May 2006, pp. 14-19.

Hughes, R. 2002, *Teaching and Researching Speaking*, Longman, Harlow.

Lennon, P. 1990, *Investigating fluency in EFL: A qualitative approach*, in "*Language Learning*", *40* (3), pp. 387-417.

Luoma, S. 2004, *Assessing speaking*, Cambridge University Press, Cambridge

North, B. 1991, Standardisation of Continuous Assessment Grades, in Alderson, JC and North, B (eds.) *Language Testing in the 1990s*, Macmillan, London/British Council, pp.167-177.

North, B and Schneider, G. 1998, *Scaling descriptors for Language Proficiency Scales*, in "Language Testing 15/2", pp. 217-262.