*Applied Linguistics 29/1: 24–49*
doi:10.1093/applin/amm017


# Assessed Levels of Second Language Speaking Proficiency: How Distinct?

[1]NORIKO IWASHITA, [2]ANNIE BROWN, [3]TIM McNAMARA and [3]SALLY O'HAGAN

[1]The University of Queensland, [2]Ministry of Higher Education and Scientific Research, Abu Dhabi, UAE, [3]The University of Melbourne

The study reported in this paper is an investigation of the nature of speaking proficiency in English as a second language in the context of a larger project to develop a rating scale for a new international test of English for Academic Purposes, TOEFL iBT (Brown *et al*. 2005). We report on a large-scale study of the relationship between detailed features of the spoken language produced by test-takers and holistic scores awarded by raters to these performances. Spoken test performances representing five different tasks and five different proficiency levels (200 performances in all) were analyzed using a range of measures of grammatical accuracy and complexity, vocabulary, pronunciation, and fluency. The results showed that features from each category helped distinguish overall levels of performance, with particular features of vocabulary and fluency having the strongest impact. Overall, the study contributes important insights into the nature of spoken proficiency as it develops and can be measured in rating scales for speaking, and has implications for methodological issues of the appropriateness of the use in language testing research contexts of measures developed in research on second language acquisition.

## INTRODUCTION

Proficiency in a second language is one of the most fundamental concepts in Applied Linguistics, and accordingly its character is the subject of ongoing and intense debate. Often this debate is about competing theories or models of second language proficiency and its development, as in the influential discussions by Canale and Swain (1980) and Bachman (1990). Less often, empirical data in the form of demonstrated proficiency at various levels of achievement in second language learning is used as the basis for such discussions. One important source of data for this kind of analysis is performance data from language tests. This is the source utilized in the present study, in which an investigation is made of the nature of proficiency at various levels of achievement in the context of the development of a scale for rating performance on speaking tasks in a test of English for Academic Purposes, TOEFL iBT. An in-depth analysis of learner speech is used to identify which features of language distinguish levels of proficiency in speaking in the context of this test.

The study can also be understood in another light, as addressing issues of test score validation, that is, providing evidence in support of or questioning the interpretations about learners' abilities made on the basis of scores awarded by judges using rating scales to rate spoken language performance. The validity issues involved have serious practical ramifications: Second and foreign language learners wishing to study overseas or to obtain a job which requires communicative skills in a particular language are often required to take a test to demonstrate their competence in the spoken language. Test scores awarded for test performance represent claims about the likely quality of performance of the learners in relevant real world contexts (Brindley 1986). However, as Douglas (1994) argues, a single summary score necessarily represents a more complex picture than can be practically reported, involving as it does an inevitable reduction of the complexity of test performance. For example, the same score does not necessarily represent the same quality of performance. If a learner has been awarded less than the highest achieved score, it does not necessarily mean that this person has performed any less well than the highest scoring person on every aspect. In general, then, the relationship of different aspects of performance to overall judgments of proficiency is an issue both for theories of the nature of language proficiency and for the interpretability of test scores.

## Language test data as evidence of the character of second language oral proficiency

In the Applied Linguistics literature, although the word 'proficient' is often used interchangeably with words such as 'good', 'fluent', 'knowledgeable', 'bilingual', 'competent', and so on, it is not always clear what speaking proficiency entails; the term may be used quite differently from researcher to researcher (Galloway 1987; McNamara 1996).

One very popular although much criticized notion of spoken proficiency in second language contexts is that described in the ACTFL Guidelines (1985 and 1999), where proficiency is presented in terms of communicative growth. Different levels of proficiency are described in a hierarchical sequence of performance ranges. The guidelines see four factors as constituting proficiency: function, content, context, and accuracy. The origin and the use of the scales means that they were written very much with classroom instruction and curriculum development in mind, and thus represent a policy statement about the nature of proficiency as much as the fruit of detailed empirical research on learner performance at each level, although considerable research has subsequently been carried out on this scale.

A number of such researchers have considered the relative weight of individual features of performance in determining overall judgments of proficiency based on the ACTFL Scale and its predecessors. For example, Adams (1980) investigated the relationship between the five factors which were identified in assessing the Foreign Service Institute (FSI) Oral Interview

Test of Speaking (i.e. accent, comprehension, vocabulary, fluency, and grammar) and the global speaking score (e.g. on a scale of 1–5) by analyzing analytic and overall score data drawn from test performances in various languages. The main factors distinguishing levels were found to be vocabulary and grammar, with accent and fluency failing to discriminate at several levels. Higgs and Clifford (1982) suggested that different factors contribute differently to overall language proficiency at the different levels defined in the FSI scale, and proposed the Relative Contribution Model (RCM) to describe rater perceptions of the relative role of each of five component factors making up global proficiency (i.e. vocabulary, grammar, pronunciation, fluency, and sociolinguistics). In their hypothesized model, vocabulary and grammar were considered to be the most important across all levels, but as the level increased, other factors such as pronunciation, fluency, and sociolinguistic factors would also become important. The hypothesized RCM was then presented to a panel of experienced teachers, whose opinions were elicited on the question of the relative contribution of factors at different levels. The results showed that teachers perceived vocabulary and pronunciation factors to be most important at lower levels with fluency and grammar factors contributing little; contributions from fluency and grammar increase as the level goes up. At higher levels, four factors (vocabulary, grammar, pronunciation, and fluency) show equal contributions, with the sociolinguistic factor contributing relatively less. Magnan (1988) examined the number of different types of grammatical errors in the transcripts of oral proficiency interviews conducted with 40 students studying French at college level, and then co-referenced this to oral proficiency interview (OPI) ratings. A significant relationship between percentage of grammatical errors and OPI rating was found, but the relationship was not always linear. Magnan explains that (1) the relationship of error to proficiency varies considerably depending on the category of error; (2) at higher levels, learners attempt more complex grammatical notions, and consequently make more errors.

Other researchers have also investigated the componential structure of proficiency at varying levels using other test instruments. De Jong and van Ginkel (1992) used speaking test data from 25 secondary school level students of French to investigate the relative contribution of different aspects of oral proficiency to the global proficiency score. The results revealed that the pronunciation category contributed most to global proficiency at the lower level, but as the level went up fluency became more important. The contribution of accuracy and comprehensibility did not vary across the levels. McNamara (1990), validating the Speaking sub-test of the Occupational English Test (OET), a specific purpose test for health professionals, investigated the relationship between the global score (Overall Communicative Effectiveness) and five analytic scales (Resources of Grammar and Expression, Intelligibility, Appropriateness, Comprehension, and Fluency). An analysis using Rasch Item Response Modelling identified Resources of

Grammar and Expression as the strongest determinant of the score for Overall Communicative Effectiveness; it was also the most 'difficult', that is the most harshly rated criterion (comprehension was scored most leniently).

Taken as a whole, the studies cited above appear to show that across levels grammatical accuracy is the principal determining factor for raters assigning a global score, with some variations in contribution of other factors depending on level. It should be noted that the data used for analysis in those studies were mostly quantitative analyses of score data, using subjective ratings of test performances or feedback; in addition, opinions from experienced teachers have been used. In other words, actual performance data from language proficiency interview transcripts has not formed the basis of the evidence, with the notable exception of the study by Magnan (1988), who analyzed performance data from interview transcripts. Let us look more closely at the potential of this latter methodology for capturing the nature of spoken proficiency at differing levels, as this was the methodology used in the present study.

## Studies of features of spoken language in oral assessment

In fact, an increasing volume of research in language testing has analyzed various features of the language produced by candidates in oral assessment. Shohamy (1994) argues that insights from such analysis provide a significant contribution to defining the construct of speaking in oral tests in general. Van Lier (1989) also stresses the importance of analysis of speech, especially the need to look at oral tests using data from the test performance (i.e. what test-takers actually said) and to analyze the test as a speech event, in order to address issues of validity. McNamara *et al*. (2002) provide a survey of studies of oral test discourse, which indicates that, while the number of studies investigating test-taker discourse has been growing, to date few studies have examined the relationship between the substance of the test-taker performance and the scores awarded.

One important exception to this is the work of Douglas and Selinker (1992, 1993), who argue that raters, despite working from the same scoring rubrics, may well arrive at similar ratings for quite different reasons. In other words, speakers may produce qualitatively quite different performances and yet receive similar ratings. Building on this insight, Douglas (1994) compared test scores with transcripts of semi-direct speaking test performance from six Czech graduate students. Various aspects of test-taker performance (local and global errors, risky versus conservative response strategies, style and precision of vocabulary, fluency, content, and rhetorical organization) were analyzed, and the actual language produced by subjects who received similar scores on the test was compared. The results revealed that very little relationship was found between the scores on the test and the language actually produced by the subjects. Douglas speculated that one of the reasons for the discrepancy could be that raters were influenced by aspects of the discourse that were not

included in the rating scales. It is generally accepted that language in use is a multi-componential phenomenon, and thus raters' interpretations of test-taker performance may vary according to which facets are being attended to and how these interact. Douglas suggested that think-aloud studies of rating processes be undertaken in order to understand more thoroughly the bases upon which the raters are making their judgments, a strategy which was adopted in the larger study (Brown *et al.* 2005) from which the data in this paper are drawn.

A further important exception is a study by Fulcher (1996), who is more optimistic about the relationship between characteristics of candidate speech and the wording of rating scales. He analyzed the transcripts of 21 ELTS interviews in terms of the rating category 'fluency'. Using Grounded Theory Methodology (Strauss and Corbin 1994), eight different aspects of fluency in the interview transcripts were considered in detail. All twenty-one transcripts were coded into eight explanatory categories, and further cross-referenced with the ELTS band scales using discriminant analysis. The results showed that all eight explanatory categories taken together discriminated well between students. The relationship between actual band scores and predicted band scores was further examined by comparing which bands would have been awarded purely on the basis of the explanatory categories. In only one case out of twenty-one was the hypothesized band different from the actual band score.

To summarize, then, an important resource for investigating the character of language proficiency as it develops is performance data from language tests. From the point of view of the validation of rating scales for oral assessment, too, investigations have been carried out in order to discover what performance features distinguish proficiency levels, and how each feature contributes to overall speaking proficiency scores. To date, the detailed examination of test performance data that have been carried out have been relatively limited in scope, and in addition their conclusions have raised serious questions of test validity. The development of TOEFL iBT provided a context in which a much more ambitious and thorough study of this issue could be carried out. In the study to be reported in this paper, we build on the Douglas (1994) study and investigate the quality of learner performance on a number of aspects of speaking proficiency that expert EAP specialist raters had identified as important in related earlier studies (Brown *et al.* 2002; Brown *et al.* 2005).

## RESEARCH QUESTIONS

The present study addresses the following two research questions.

(1) In what ways does performance on EAP speaking tasks differ by level?

(2) What are the distinguishing features of test performance at each of five assessed levels of proficiency?

## THE STUDY

### Data

The data used for the study were initially collected as part of the piloting of prototype tasks for TOEFL iBT (Lee 2005). Performances on five pilot oral test tasks had been double-rated by trained Educational Testing Service (ETS) test development staff using a draft global scale with five levels (levels 1–5); a G-study of the data estimated reliabilities of 0.88 for double-rated speech samples with five tasks (the data for this study) (further details can be found in Lee 2005: 9). In fact, reliabilities were likely to have been higher than this, as the G-study was conducted on unadjudicated scores, whereas the current study used adjudicated scores, whereby if raters disagreed by more than one point the sample was scored a third time. Approximately 2–5 per cent of samples required adjudication (M. Enright, personal communication, 17 October 2006). For the purposes of this project, for each task ten samples at each of the five levels were initially selected from a larger pool of pilot test data, a total of 50 performances per task; 250 in total. Of these, two performances at each level were discarded because of problems of audibility. In order to ensure the same number of performances at each level and on each task we were left with eight performances at each level on each task; 200 in total. The ESL learners who took the trial test varied in terms of age, L1, and length of residence in an English-speaking country and prior time spent studying English, but all were studying English to prepare for tertiary study in the USA at the time of data collection.

### Tasks

The five test tasks used in the present study were of two types, independent and integrated, based on whether performance involved prior comprehension of extended stimulus materials. In the independent tasks, participants were asked to express their opinion on a certain topic that was presented with no accompanying material to read or hear. In the integrated tasks, participants first listened to or read information presented in the prompt, and then were asked to explain, describe, or recount the information. The amount of preparation and speaking time varied for each task, but longer preparation and speaking times were given for the integrated tasks than for the independent ones. A summary of the five tasks is given in Table 1.

## METHODS OF ANALYSIS

The seven features analysed in the present study are grouped according to three conceptual categories that EAP specialist raters had identified as important in assessing performances on the tasks (Brown *et al*. 2002; Brown *et al*. 2005). In these preceding studies, a bottom-up approach was used to derive the categories: raters participated in a think-aloud procedure as they listened to and re-listened to the performances, and the researchers sifted

*Table 1: The tasks*

| Task | Type | Targeted functions and discourse features | Preparation time (secs) | Speaking time (secs) |
|---|---|---|---|---|
| 1 | Independent | Opinion; Impersonal focus; Factual/ conceptual information | 30 | 60 |
| 2 | Independent | Value/significance; Impersonal focus; Factual/conceptual information | 30 | 60 |
| 3 | Integrated; Monologic lecture | Explain/describe/recount; Example/event; cause/effect | 60 | 90 |
| 4 | Integrated; Dialogic lecture | Explain/describe/recount; Process/procedure; Purpose/results | 60 | 90 |
| 5 | Integrated; Reading | Explain/describe/recount; Process/procedure; Purpose/results | 90 | 90 |

through the comments to derive the categories of performance to which the raters seemed oriented. The resulting broad conceptual categories were *Linguistic Resources*, *Phonology*, and *Fluency*. (While these appear to be conceptually overlapping categories, it seemed as if they corresponded to distinct areas of rater orientation.) A larger number of specific features were identified in these studies for each of three conceptual categories, and from them, a number of features were selected for analysis in this study: for *Linguistic Resources*, the features *grammatical accuracy*, *grammatical complexity*, and *vocabulary* were analysed; for *Phonology*, the features chosen were *pronunciation*, *intonation*, and *rhythm*; and *Fluency* was regarded as a single feature, analysed in multiple ways, as we shall see. For each of these seven features, a number of methods of analysis were identified, drawing on relevant literature on discourse analysis and interlanguage analysis in SLA, and on consultations with linguistics experts. A phonetician was especially informative with respect to the analysis of the three Phonology features. Details of each of the analyses are given below; fuller information is provided in Brown *et al*. (2005).

All 200 speech samples were transcribed using transcription guidelines described in a study by Ortega *et al*. (in progress). The segmented and coded speech samples were entered into a database for use with the CLAN (Computerized Language Analysis) program developed as part of the *CHILDES* project (MacWhinney 1999) for the analyses for grammatical accuracy, complexity, and fluency. The CLAN program allows a large number of automatic analyses to be performed on data, including frequency counts,

word searches, co-occurrence analyses, and calculation of type/token ratios. For vocabulary analysis, a web-based program *VocabProfile* (Cobb 2002), and for phonological analysis, the software *Xwaves* (Rommark 1995) were used respectively.

## Linguistic resources

In the analyses of *Linguistic Resources*, we focused on three features: *grammatical accuracy*, *grammatical complexity*, and *vocabulary*. It will be noted that the features chosen represent sentence-level phenomena, rather than more complex features of discourse, particularly academic discourse, such as overall discourse organization, discourse coherence and cohesion, or the accuracy and complexity of the content. In fact, within this category the EAP specialists in the rater cognition study (Brown *et al*. 2002) also identified *textual* features such as use of connectives, cohesion, and discourse markers. Details of the results for these features, and an analysis of how the accuracy and complexity of the content varied across levels, are reported in Brown *et al*. (2005); we report on a reduced set of features in this paper in the interests of length, but wish to stress that the features we have chosen to discuss here are not being proposed as an adequate operationalization of the construct of academic speaking proficiency.

### Grammatical accuracy

Empirical studies in both language testing and SLA have reported measures of grammatical accuracy of learner speech either in terms of *global accuracy* (i.e. identifying any and all types of error) (e.g. Foster and Skehan 1996; Skehan and Foster 1999) or in terms of *specific types of error* (e.g. Robinson 1995; Wigglesworth 1997; Ortega 1999). The global accuracy approach has the advantage of being potentially the most comprehensive in that all errors are considered. However, it is also the hardest in which to establish consistency of coding. In an earlier study (Iwashita *et al*. 2001), it was found that coders tended not to agree on what they considered to be errors or on whether they should be classified as grammatical or lexical. In the studies of global accuracy reported in Foster and Skehan (1996) and Skehan and Foster (1999), no inter-coder reliability measures were reported. Given these uncertainties, a decision was made to measure grammatical accuracy through both methods: accuracy of use of specific grammatical features, and global accuracy. The specific features chosen were verb tense, third person singular, plural markers, prepositions, and article use. Global accuracy was examined by calculating error free T-units as a percentage of the total number of *T*-units. A *T*-unit is defined as an independent clause and all its dependent clauses (Hunt 1970). Error free *T*-units are *T*-units free from any grammatical errors including both the specific errors defined above as well as other grammatical errors (e.g. word-order, omission of pronouns).

A decision was also made to identify errors using the 'target-like-use' (TLU) analysis developed by Pica (1983) rather than a 'supplied in obligatory context' (SOU) analysis. The difference here is that the TLU analysis includes learner errors produced in both non-obligatory contexts and obligatory contexts. In counting errors in (and correct use of) the features listed above, the transcribed speech was first pruned by excluding features of repair. This meant that learners were considered to have shown evidence of correct use of the target-like feature when it was demonstrated in their repaired utterance.

## Grammatical complexity

Grammatical complexity refers to characteristics of utterances at the level of clause relations, that is, the use of conjunctions and, in particular, the presence of subordination. The following four measures were reported in the present study.

(1) the number of clauses per *T*-unit (the *T*-unit complexity ratio);

(2) the ratio of dependent clauses to the total number of clauses (the dependent clause ratio);

(3) the number of verb phrases per *T*-unit (the verb–phrase ratio);

(4) the mean length of utterance (MLU).

The first three of these measures were identified in a review of second language writing studies by Wolfe-Quintero *et al.* (1998) as the measures which best capture grammatical complexity, and have also been used in studies involving the analysis of learner speech in both pedagogic and testing contexts (e.g. Skehan and Foster 1999; Iwashita *et al.* 2001). Segmentation of the learner utterances into *T*-units, clauses, and verb phrases was carried out following the guidelines developed in Ortega *et al.* (in progress). Mean length of utterance (MLU) was measured by calculating the number of morphemes per utterance, with an utterance including both *T*-units and fragments.

## Vocabulary

Vocabulary knowledge was examined using the web program *VocabProfile* (Cobb 2002), which measures the proportions of low and high frequency vocabulary used. Both type and token measures were calculated. The token measure was used as it was assumed that for weaker participants not all of the time allowed would be taken up with speech, and even if it was, it was likely to be slower and thus yield fewer tokens. The type measure was

chosen as a measure of the range of vocabulary used; it was hypothesized that more proficient speakers would use a wider range of types.

## Phonology

The phonological analysis was undertaken using the speech data labelling application software *Xwaves* (Rommark 1995). *Xwaves* was used because it allowed a number of features to be transcribed against a recording, tagging them to particular time-points. The frequency of each type of tag could then be calculated. Part of the *Xwaves* program is a labelling module, which can be set up as one likes. Thus, comments or labels in the form of ASCII text can be entered onto the screen and aligned with particular points in the speech file. This way, it was possible to add labels for words, segments and any other features of interest. Because of their time-consuming nature, analyses were carried out on a portion of the data only: 30 seconds from each performance on one independent task (Task 2) and one integrated task (Task 3).

### Pronunciation

The analysis of pronunciation features was conducted at word level and sub-word level. In the word level analysis, the coders first categorized words as meaningful or not meaningful. The pronunciation of meaningful words was then classified as 'target-like', 'marginally non-target-like', or 'clearly non-target-like'. In the sub-word level analysis, syllables were again assessed as to whether they were 'target-like', 'marginally non-target-like', or 'clearly non-target-like'.

### Intonation

The assessment of intonation was conducted in terms of the number of completed intonation units. The analysis considered whether learners produced completed units, cut-off or incomplete units, or isolated words. Performances were first categorized as displaying *many* or *few* intonation units; the *many* category was then further broken down into performances showing English-like intonation (E), nearly English-like (Nr), and non-English-like (N). Criteria for allocation into these sub-categories included: Do they follow general patterns such as rising pitch to indicate continuation, and falling phrase-final pitch to end a thematic section? Do they place pitch accent on focused words and phrases in the sentence? Do they pronounce their English using the intonation patterns of another language (i.e. learner's L1)?

### Rhythm

Most varieties of English have a rhythm based on word stress, so that stressed syllables come at a regular rate (i.e. they are *stress timed*). In contrast, many other languages, and even some varieties of English (e.g. Indian, Singaporean), are *syllable-timed*: generally, each syllable comes at a regular speed. Syllable-timed speech is known to be particularly problematic for

speakers of stress-timed languages and vice versa. The categories used for the analysis of rhythm were: *stress-timed*, *syllable-timed*, *variable* (denotes speakers who wavered between the two), and *unclear* (when judges could not really tell: this tended to happen when the speech samples were not long enough).

## Fluency

The following features were identified as suitable measures of fluency: *filled pauses* (*um*s and *er*s), *unfilled pauses*, *repair*, *total pausing time* (as a percentage of total speaking time), *speech rate*, and *mean length of run*. The number of *unfilled pauses* was calculated by counting the number of pauses of 1 second or more that occurred in the speech (Mehnert 1998). In order to enable comparisons, instances of *filled pauses*, *unfilled pauses*, and *repair* were counted per 60 seconds of speech, because the actual speaking time of individual learners varied (as a function of the amount of pause time and filled pauses). *Repair* refers to repetition of exact words, syllables or phrases; replacement; reformulations (grammatical correction of structural features); false starts; and partial repetition of a word or utterance (Freed 2000). *Total pausing time* was calculated by adding up all the unfilled pauses. *Speech rate* was calculated by dividing the total number of syllables produced in a given speech sample by the total time expressed in seconds (Ortega 1999). First, the transcribed speech was pruned by excluding features of repair; then the resulting total number of syllables was divided by the total speech time excluding pauses of three or more seconds. *Mean length of run* was calculated in terms of the mean number of syllables produced in utterances (Towell *et al.* 1996).

## Statistical analyses

For most features, results are reported in terms of inferential statistics. The exceptions are intonation and rhythm given the nature of the analysis in each case. For the inferential statistical analysis, Analysis of Variance (ANOVA) with two factors (i.e. level and task) ($2 \times 2$ design) was used for most of the data. However, for those variables that did not need to be converted into frequency data per amount of time, that is for three of the complexity measures, given that they were ratio data. Analysis of Covariance (ANCOVA) was performed instead in order to eliminate the potential effect of the amount of speech. For both ANOVA and ANCOVA analyses, some data were skewed and variances were not homogenous. In these cases, transformation of the data (e.g. a log or square root transformation) was considered. However, after consultation with a statistician, it was decided not to use transformed data for two reasons: (1) transformed variables are generally hard to interpret; (2) both ANOVA and ANCOVA statistics are robust to violations of their assumptions especially when the sample size is large.

## Inter-coder reliability

A portion of the data (i.e. approximately 10 per cent of the data) was coded twice by a second coder to calculate inter-coder reliability. Inter-coder reliability was calculated using the Spearman–Brown prophesy formula (Table 2). Achieved levels were high in almost all cases, with the exception

*Table 2: Summary of inter-coder reliability*

| Category | Coding units | *N* | *Spearman–Brown* |
|---|---|---|---|
| *Linguistic Resources* | | | |
| *Grammatical accuracy* | | | |
| Specific types of errors | 1. Correct | | |
| | article | 19 | 0.96 |
| | plural | 19 | 0.98 |
| | preposition | 19 | 0.98 |
| | tense | 19 | 0.99 |
| | third person singular | 19 | 0.91 |
| | 2. Error | | |
| | article | 19 | 0.93 |
| | plural | 19 | 0.86 |
| | preposition | 19 | 0.88 |
| | tense | 19 | 0.71 |
| | third person singular | 19 | 0.78 |
| Global accuracy | Error free T-unit | 19 | 0.98 |
| | Error T-unit | 19 | 0.99 |
| *Grammatical complexity* | | | |
| T-unit complexity ratio | T-unit | 20 | 0.91 |
| | Clause | 20 | 0.94 |
| Dependent clause ratio | Dependent clause | 20 | 0.99 |
| Verb phrase ratio | Verb phrase | 20 | 0.98 |
| MLU[a] | Morphemes | 20 | 1.00 |
| *Fluency* | | | |
| Filled pauses (um and ers) | Filled pauses | 20 | 1.00 |
| Repair | Repair | 20 | 0.98 |
| MLR[b] | Syllables | 20 | 1.00 |
| | Utterances | | 0.98 |

[a]MLU = mean length of utterance.
[b]MLR = mean length of runs.

*Table 3: Grammatical accuracy (1)*

| Level | Articles | | | Tense marking | | | 3rd person singular verbs | | |
|---|---|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | *N* | *M* | *SD* | *N* | *M* | *SD* |
| 1 | 40 | 69.49 | 25.71 | 30 | 47.89 | 37.43 | 30 | 69.96 | 34.54 |
| 2 | 40 | 73.63 | 22.32 | 36 | 65.19 | 35.06 | 31 | 64.40 | 41.14 |
| 3 | 38 | 70.19 | 17.30 | 40 | 64.12 | 32.28 | 28 | 70.70 | 36.09 |
| 4 | 40 | 75.37 | 17.40 | 39 | 75.20 | 28.31 | 34 | 78.80 | 32.08 |
| 5 | 40 | 84.43 | 17.16 | 40 | 86.58 | 14.87 | 26 | 91.11 | 19.60 |

of two of the grammatical accuracy features (tense and 3rd person) where the level of agreement was marginally below 0.8.

The phonological analysis was carried out by two trained phoneticians, and the question of reliability was addressed somewhat differently from that used in the other analyses. The following procedure was followed to establish adequately high inter-coder agreement. In the earlier stages, while the two coders were refining the feature categories, they went through data from ten learners together. Then, in a test-run for the final categories, they first went through three learners together, then transcribed five learners independently, and finally compared their results for the data from the five learners. In comparing the counts of non-target features, they differed in the following way: over the five learners, the average disagreement between the feature-counts differed by at most 1 token for both 'clear' and 'marginal' features.

## RESULTS

### Linguistic resources

#### Grammatical accuracy

The descriptive statistics of all six measures revealed that for all measures Levels 4 and 5 are distinct from other levels, but the pattern is less clear at the lower levels (see Tables 3 and 4 for descriptive statistics). ANOVA analyses ($2 \times 2$) were performed separately for each variable, and for each one, highly significant differences were observed (*Articles*, $F\ (4,\ 188) = 3.41$, $p = 0.001$, $eta = 0.07$; *Tense marking* $F\ (4,\ 175) = 7.45$, $p = 0.001$, $eta = 0.15$; *3rd person singular* $F\ (4,\ 139) = 3.01$, $p = 0.02$, $eta = 0.08$; *Plural* $F\ (4,\ 173) = 9.58$, $p = 0.001$, $eta = 0.17$; *Preposition* $F\ (4,\ 188) = 7.42$, $p = 0.001$, $eta = 0.14$; *Global accuracy* $F\ (4,\ 173) = 13.51$, $p = 0.001$, $eta = 0.22$). However, the effect sizes (*eta*) were all marginal, ranging from 0.07 (*3rd person singular*) to 0.22 (*Global accuracy*), which reflects the wide standard deviations for each measure at each level.

*Table 4: Grammatical accuracy (2)*

| Level | Plural | | | Prepositions | | | Global accuracy | | |
|---|---|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | *N* | *M* | *SD* | *N* | *M* | *SD* |
| 1 | 37 | 66.88 | 33.63 | 39 | 73.28 | 25.49 | 38 | 16.50 | 24.31 |
| 2 | 40 | 58.99 | 34.22 | 39 | 83.07 | 16.58 | 40 | 21.29 | 27.85 |
| 3 | 40 | 73.65 | 27.73 | 40 | 85.26 | 11.73 | 40 | 20.96 | 19.01 |
| 4 | 39 | 81.12 | 16.89 | 40 | 88.16 | 10.26 | 40 | 30.98 | 22.98 |
| 5 | 40 | 94.07 | 9.10 | 40 | 90.71 | 11.99 | 40 | 50.93 | 21.83 |

*Table 5: Grammatical complexity*

| Level | T-unit complexity | | | DC ratio | | | VP ratio | | | MLU | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | *N* | *M* | *SD* | *N* | *M* | *SD* | *N* | *M* | *SD* |
| 1 | 33 | 2.17 | 1.12 | 33 | 0.42 | 0.21 | 33 | 0.40 | 0.55 | 40 | 15.19 | 6.37 |
| 2 | 39 | 2.03 | 0.87 | 39 | 0.39 | 0.23 | 39 | 0.44 | 0.36 | 40 | 17.47 | 5.33 |
| 3 | 40 | 1.93 | 0.49 | 40 | 0.41 | 0.14 | 40 | 0.39 | 0.25 | 40 | 17.53 | 6.20 |
| 4 | 40 | 1.92 | 0.59 | 40 | 0.41 | 0.14 | 40 | 0.44 | 0.38 | 40 | 18.31 | 7.90 |
| 5 | 40 | 2.04 | 0.65 | 40 | 0.41 | 0.16 | 40 | 0.54 | 0.47 | 40 | 19.77 | 6.34 |

## Grammatical complexity

The four *grammatical complexity* measures yielded mixed results (see Table 5 for descriptive statistics). The expected gradient of increasing complexity per level was found for only one of the measures (*MLU*). No such pattern was observed for the *T-unit complexity ratio* ($F$ (4, 179) = 0.95, $p = 0.22$) or the *Dependent clause ratio* ($F$ (4, 181) = 1.4, $p = 0.24$). When the *number* of utterances produced in each performance was taken into consideration in an analysis of covariance (ANCOVA), significant differences were found across levels for *Verb phrase complexity* ($F$ (4, 182) = 3.50, $p = 0.01$, *eta* = 0.07) and *MLU* ($F$ (4, 187) = 2.82, $p = 0.02$, *eta* = 0.19), but both effect sizes were marginal.

## Vocabulary

Table 6 shows that increases in level were associated with an increase in the number of words produced (tokens) and a wider range of words (type). ANOVA analyses showed significant differences for token and type (Token $F$ (4, 190) = 62.32, $p = 0.001$, *eta* = 0.57; Type $F$ (4, 190) = 47.88, $p = 0.001$, *eta* = 0.50), with medium effect sizes.

*Table 6: Vocabulary*

| Level | | Token | | Type | |
|---|---|---|---|---|---|
| | *N* | *M* | *SD* | *M* | *SD* |
| 1 | 40 | 55.68 | 18.86 | 38.02 | 12.94 |
| 2 | 40 | 69.92 | 18.76 | 42.73 | 9.78 |
| 3 | 40 | 86.87 | 20.08 | 49.05 | 12.97 |
| 4 | 40 | 100.08 | 22.62 | 56.39 | 14.04 |
| 5 | 40 | 118.09 | 22.64 | 66.04 | 14.55 |

*Note*: Word-token and type data are frequency data (per 60 seconds).

*Table 7: Pronunciation*

| | Level 1 *N* = 14 | | Level 2 *N* = 16 | | Level 3 *N* = 16 | | Level 4 *N* = 17 | | Level 5 *N* = 1 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| *Word level: per 10 words* | | | | | | | | | | |
| Meaningful words, target like | 8.68 | 1.05 | 8.44 | 1.01 | 8.56 | 0.69 | 8.98 | 0.58 | 9.06 | 0.58 |
| Meaningful words, but marginally non-target like | 0.13 | 0.21 | 0.14 | 0.14 | 0.16 | 0.20 | 0.14 | 0.18 | 0.09 | 0.11 |
| Meaningful words, but clearly non-target like | 0.12 | 0.23 | 0.23 | 0.38 | 0.13 | 0.16 | 0.16 | 0.19 | 0.03 | 0.07 |
| Non-meaningful words | 1.06 | 0.92 | 1.19 | 0.92 | 1.16 | 0.66 | 0.72 | 0.48 | 0.82 | 0.61 |
| *Sub-word level: per 10 syllables in meaningful words* | | | | | | | | | | |
| On target syllables | 8.25 | 0.93 | 8.03 | 0.85 | 8.64 | 0.68 | 9.08 | 0.53 | 9.46 | 0.30 |
| Marginally non-target-like syllables | 0.31 | 0.29 | 0.35 | 0.24 | 0.29 | 0.29 | 0.22 | 0.21 | 0.12 | 0.10 |
| Clearly non-target-like syllables | 1.27 | 0.95 | 1.39 | 0.73 | 1.01 | 0.58 | 0.65 | 0.54 | 0.42 | 0.30 |

*Note*: All data reported in this Table are frequency data (per 10 words and syllables).

## Phonology

### Pronunciation

Pronunciation was assessed at the word level and at the sub-word level. The results are presented in Table 7.

In the *word level analysis*, the proportion of meaningful words classified as showing 'target-like' pronunciation increased across levels, with the

*Table 8: Intonation*

| Level | Native like ← | | | Non-native like → | |
|---|---|---|---|---|---|
| | E | Nr | N | F | Total |
| 1 | 1 | 0 | 7 | 6 | 14 |
| 2 | 0 | 0 | 10 | 6 | 16 |
| 3 | 2 | 2 | 5 | 7 | 16 |
| 4 | 3 | 4 | 9 | 0 | 16 |
| 5 | 7 | 5 | 4 | 0 | 16 |
| Total | 13 | 11 | 35 | 19 | 78 |

exception of Levels 1 and 2. The proportion of words that were classified as 'marginally non-target like' or 'clearly non-target like' in pronunciation was not sensitive to level, nor was the number of words classified as 'non-meaningful'. An ANOVA analysis was performed to compare the frequency of meaningful words across levels, but no significant difference was observed ($F$ (4, 69) = 1.75, $p$ = 0.15, *eta* = 0.09). No analysis was performed on the other categories as frequencies were very low.

In the *sub-word level analysis*, more noticeable differences across levels were found than in the word-level analyses. Syllables were again assessed as to whether they were 'target-like', 'marginally non-target-like', or 'clearly non-target-like'. In general, the number of 'non-target-like' syllables (especially 'marginally non-target-like' syllables) was sensitive to level. The results of the ANOVA analysis of the frequency of 'target-like' syllables showed a highly significant difference across levels ($F$ (4, 69) = 11.49, $p$ = 0.001); the effect size was small (*eta* = 0.40). Again, statistical analyses of the other categories were not seen as meaningful because of the low observed frequencies associated with them.

## Intonation

The expected pattern emerged, that is that the intonation units of higher-level learners were more frequently categorized as 'Many and English-like', compared with those of lower level learners. More than half of the lower level learners fell into the two lowest performance categories ('Many and not English-like' and 'Few'). Few lower level learners achieved 'English-like' intonation, and many were assessed as performing in the 'not English-like' category. Half of all learners below Level 3 fell into the two lowest categories (see Table 8).

*Table 9: Rhythm*

| Level | Native like ← | | | Non-native like → | |
|---|---|---|---|---|---|
| | St | V | U | Sy | Total |
| 1 | 3 | 0 | 6 | 5 | 14 |
| 2 | 3 | 2 | 4 | 7 | 16 |
| 3 | 6 | 1 | 4 | 5 | 16 |
| 4 | 6 | 4 | 4 | 2 | 16 |
| 5 | 10 | 4 | 2 | 0 | 16 |
| Total | 28 | 11 | 20 | 19 | 78 |

*Table 10: Fluency measures*

| Level | | Filled pauses | | Unfilled pauses | | Total pause time | | Repair | | Speech rate | | Mean length of run | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| 1 | 39 | 4.62 | 4.88 | 5.98 | 5.44 | 24.78 | 19.71 | 4.35 | 2.83 | 1.32 | 0.44 | 26.81 | 19.32 |
| 2 | 40 | 6.07 | 4.63 | 5.13 | 2.83 | 18.68 | 11.70 | 4.45 | 3.07 | 1.66 | 0.44 | 22.85 | 12.25 |
| 3 | 40 | 6.06 | 4.58 | 3.93 | 2.43 | 11.22 | 7.79 | 4.39 | 2.63 | 2.02 | 0.45 | 20.84 | 11.22 |
| 4 | 39 | 6.61 | 4.91 | 3.04 | 2.84 | 7.79 | 8.05 | 4.49 | 2.10 | 2.36 | 0.46 | 21.30 | 11.85 |
| 5 | 40 | 6.64 | 5.64 | 1.49 | 1.83 | 3.81 | 4.68 | 3.46 | 2.44 | 2.83 | 0.50 | 22.80 | 10.16 |

*Note*: The numbers of filled pauses, unfilled pauses and repairs reported are frequency data (per 60 seconds).

## Rhythm

Rhythm proved sensitive to differences in level: few lower level speakers were assessed as managing stress-timed speech, which seemed to be a characteristic of higher-level performances. The prosodic timing of more than half of the Level 4 and 5 learners was coded as 'stress timed'. In contrast, the coders' judgments of prosodic timing on many of the Level 1 and 2 learners fell into the categories 'unclear' and 'syllable timed' (Table 9). For rhythm, there was no difference in the number of Level 3 and Level 4 learners assessed as 'stress timed', but more Level 3 learners were assessed as 'unclear' and 'syllable-timed' than Level 4 learners.

## Fluency

The results for the following three measures showed a clear relationship with proficiency level: *speech rate*, *number of unfilled pauses*, and *total pause time* (see Table 10 for descriptive statistics). Higher-level learners spoke faster with less pausing, and fewer unfilled pauses. ANOVA analyses showed significant differences across levels for each of these measures: *speech rate* ($F$ (4, 189) = 71.32, $p = 0.001$, *eta* = 0.60), *unfilled pauses* ($F$ (4, 190) = 12.19, $p = 0.001$, *eta* = 0.20), and *total pause time* ($F$ (4, 190) = 20.62, $p = 0.001$, *eta* = 0.30) with medium or small effect sizes. No significant differences were observed for *filled pauses* ($F$ (4, 190) = 0.75, $p = 0.56$, *eta* = 0.02), *repair* ($F$ (4, 190) = 0.99, $p = 0.41$, *eta* = 0.02), and *mean length of run* ($F$ (4, 190) = 1.6, $p = 0.18$, *eta* = 0.03).

## Summary

Overall, a number of measures provided evidence that features of the test-taker discourse under analysis varied according to proficiency level. Significant differences across levels in the expected direction were found for at least some of the measures of each of the following features: *Grammatical accuracy* (all measures), *Grammatical complexity* (verb phrase complexity and mean length of utterance), *Vocabulary* (both token and type), *Pronunciation* (target-like syllables), and *Fluency* (speech rate, unfilled pauses, and total pause time). However, the distinctions between adjacent levels were not always absolutely clear-cut in that the differences between levels where they did exist were not as great as might have been expected, with the exception of some of the pronunciation and fluency measures, and the number of word tokens. Also, large standard deviations were found for most measures, indicating broad variation among learners assessed at any one level, and overlap between levels, which are reflected in the mostly modest effect sizes (see a summary of the findings of the statistical analyses in Table 11). In other words, for most measures, while the differences across level were real, that is, not attributable to chance, their impact on the overall level assigned to the test taker was not particularly strong. This is in itself not terribly surprising, given that the level scores reflect judgments of performance in terms of a composite of potentially many features. However certain measures, while in themselves showing small or medium effect sizes, had a greater relative impact on overall scores. These were measures of *Grammatical accuracy* (i.e. global accuracy), *Vocabulary* (i.e. word type and token), *Pronunciation* (target-like syllables), and *Fluency* (i.e. unfilled pauses, total pause time, and speech rate). Of these, the strongest were *Vocabulary* (token) and *Fluency* (speech rate). (Note that the effect sizes of the differences in the *Grammatical complexity* measures were all marginal.)

*Table 11: Summary of statistical analyses*

| | | Difference | Effect size |
|---|---|---|---|
| *Linguistic resources* | | | |
| *Grammatical accuracy* | Article | ✓ | 0.07 |
| | Tense marking | ✓ | 0.15 |
| | 3rd person singular | ✓ | 0.08 |
| | Plural | ✓ | 0.17 |
| | Preposition | ✓ | 0.14 |
| | Global accuracy | ✓ | 0.22 |
| *Grammatical complexity* | T-unit complexity | | |
| | DC ratio | | |
| | VP ratio | ✓ | 0.07 |
| | MLU | ✓ | 0.19 |
| *Vocabulary* | Token | ✓ | 0.57 |
| | Type | ✓ | 0.5 |
| *Phonology* | | | |
| *Pronunciation* | Meaningful words | | |
| | Target-like syllables | ✓ | 0.4 |
| *Fluency* | No. of filled pauses | | |
| | No. of unfilled pauses | ✓ | 0.2 |
| | Total pause time | ✓ | 0.3 |
| | Repair | | |
| | Speech rate | ✓ | 0.6 |
| | MLR | | |

*Notes*: ✓ = statistical difference: Effect size (eta); M = marginal (< 0.2); S = small (> 0.2 to < 0.5); MED = medium (> 0.5 to < 0.8); L = large (> 0.8).

## DISCUSSION

In the present study, we investigated various features of learner discourse by conducting in-depth analyses of test-taker performances in order to see in what ways task performances differ by level and what features distinguish levels more clearly than others. Our findings can be considered in terms of the nature of proficiency, and the validity of the scale being developed in the large study (Brown *et al*. 2005).

First, we found that a set of features that seemed to have an impact on the overall assigned score. These were *Vocabulary* (i.e. word type and token), *Fluency* (i.e. unfilled pauses, total pause time, and speech rate), *Grammatical accuracy* (i.e. global accuracy), and *Pronunciation* (target-like syllables); of these, *Vocabulary* and *Fluency* seemed particularly important. Our results

reveal that features drawn from a wide range of categories were making independent contributions to the overall impression of the candidate. Features from each of the three main conceptual categories investigated (i.e. Linguistic Resources, Phonology, and Fluency) were shown to have the greatest influence on overall scores; no category was omitted. It is also notable that more macro-level categories—speech rate, the main vocabulary measures, a global pronunciation measure, and the global grammatical accuracy measure—appear to have most influence on scores, which is what we might expect. Our results showed that even if one aspect of language (for example, grammatical accuracy) is not as good as other aspects, the rating of the overall proficiency of that speaker is not necessarily determined solely by their performance on that one aspect of language. A combination of aspects determines the assessment of the overall proficiency of the learner. This appears to conflict with the findings of earlier studies that perceptions of grammatical accuracy are the main drivers of oral proficiency scores, as discussed earlier (e.g. Higgs and Clifford 1982; McNamara 1990), although it should be noted that the category 'Resources of Grammar and Expression' which McNamara found to be so important includes two of the categories found to be important here (i.e. grammatical accuracy and vocabulary). The in-depth analysis of a number of features of performance in speaking presented here provides further insight into the characteristics of oral proficiency, and contributes to further understandings of the development of oral proficiency, for example by demonstrating that various features of oral proficiency do not develop in a linear fashion.

Second, however, level differences at adjacent levels were not always distinguished by measures of the features under investigation. This is especially evident in the performances of lower level learners and on some features (i.e. *Grammatical accuracy* and *Grammatical complexity*). As we have seen, for many features of learner speech, Level 5 and Level 4 learners demonstrated clearly better performances, but the performances of Level 1 learners were not always the worst. In other words, for some features, the slope of increasing levels of proficiency at the lower end was not always in the expected direction (i.e. a performance at Level 3 was not always better than one at Level 1). Also we found differences in production features (speech rate, total pause time, intonation, rhythm) and vocabulary to be more distinctive than grammatical accuracy and complexity features at these levels.

The unexpected direction of the slope of the increasing levels of proficiency for certain features among the lower level learners can be understood in terms of the way raters handled problematic features in the speech of these learners. Although far more problematic features were observed in the speech of lower level learners than among higher level ones, these problematic areas of speech were not homogeneous among learners. For example, some performances were problematic in just one area of

grammatical accuracy (e.g. tense marking), and others showed problems with a number of different areas of grammatical accuracy (e.g. articles, plural markings). Also, at the macro level, some learners' pronunciation was very weak, but their vocabulary knowledge was very good. It is possible that the poor quality of the pronunciation turned out to be a decisive factor in determining the learner's proficiency in the initial level assignment conducted by ETS raters. Some support for this idea is found in the comments made by expert EAP teachers in the larger study (see Brown *et al*. 2005), where a high percentage of comments in the 'Pronunciation' category made reference to intelligibility *and* they were often negative. If a listener cannot make out the words, then they are not in a position to judge the ideas, syntax, and so on; pronunciation may therefore be acting as a sort of first level hurdle.

An interesting possibility for understanding the fuzziness of distinctions at adjacent levels is raised in a study by Lee and Schallert (1997) of the relationship between L2 proficiency and L2 reading comprehension. They found that there was no difference in correlation coefficients of the relationship between L1 reading and L2 reading from adjacent L2 proficiency levels when the performances were assigned to ten L2 proficiency-level groups, but when the data were grouped into five proficiency levels, a clear difference appeared for the critical comparison between adjacency levels. It is perhaps possible in the present study that if the initial assigned proficiency level had been in terms of three proficiency levels instead of five for instance, level distinctions might have been more clear-cut.

In general, however, the complexity of the configuration of components in any overall judgment of proficiency, and the fuzziness of distinctions between levels because of the large standard deviations for features at any one level, appear to support the insight of Douglas and Selinker (1992, 1993) and Douglas (1994) that speakers may produce qualitatively quite different performances and yet receive similar ratings. To what extent this invalidates ratings, however, is a different question. Clearly, raters are making an 'on balance' judgment, weighing several factors in making a rating category decision. As Lumley (2005) has shown in relation to the assessment of writing, raters are faced with the task of fitting their responses to the great complexity of the texts they are rating into the necessarily simple category descriptions available within rating scales. We need more detailed study of how raters go about this demanding task in relation to the assessment of speaking, and in particular the process through which they balance the multiple features they are attending to. The good news is that the speculation by Douglas that raters were influenced by aspects of the discourse which were not included in the rating scales does not appear necessarily to be supported by this study, as a range of features studied appeared to be making a contribution to the overall judgment, and these features were drawn from each of the principal categories studied.

*Table 12: Numbers of T-units and clauses, and grammatical complexity measures (five test takers)*

| Test-taker ID | Level | *T*-unit | Clause | *T*-unit complexity |
|---|---|---|---|---|
| 1–130037 | 1 | 2 | 3 | 1.5 |
| 2–320028 | 2 | 4 | 6 | 1.5 |
| 3–130067 | 3 | 8 | 15 | 1.87 |
| 4–320003 | 4 | 9 | 17 | 1.89 |
| 5–320071 | 5 | 12 | 19 | 1.58 |

## Methodological issues

One very obvious methodological limitation in the present study is that the speaking proficiency investigated represents a narrow interpretation of the construct, as it consists of monologic task performances. Ability to sustain conversation and communicate with speakers of the target language was not investigated. The relationship of the speaking proficiency investigated here to this latter ability is unclear, and is the subject of ongoing debate within language testing research (e.g. O'Loughlin 2001). Further studies investigating interactional data are required to validate the findings of the current study.

In terms of methods of analysis of speech data, our findings have a number of implications. First, there is an issue with the ratio measures used (the *T*-unit complexity ratio, the dependent clause ratio, and the verb–phrase ratio), even though, as stated above, these had been recommended on the basis of previous studies as among the most useful measures of complexity (Wolfe-Quintero *et al*. 1998). If we look at a sample of actual performances at different levels, we note interesting apparent differences (see examples in the Appendix, available online). First is the sheer volume of clauses and *T*-units produced at the higher levels, which contrasts with the lower levels. This difference is, however, cancelled out when ratios are used. Table 12 shows examples of *T*-units and clauses produced by five learners representing all five levels. As the level goes up, the number of *T*-units and clauses also increases, but the ratio does not increase accordingly.

Secondly, there is some evidence of increasing complexity per level, although this is not a strong or uniform effect. Shorter and simpler sentences with little subordination were more frequently observed at lower levels, whereas complex sentences with several instances of subordination were in general a feature at higher levels, although there was not a strong difference between Levels 3, 4, and 5. Table 13 summarizes the degree of subordination in terms of the number of clauses per *T*-unit. While in the speech produced by Level 1–3 learners, only one level of subordination (i.e. two clauses per

*Table 13: Degree of subordination (five test takers)*

| Test-taker ID | Level | Degree of subordination | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| 1–130037 | 1 | 1 | 1 | 0 | 0 |
| 2–320028 | 2 | 4 | 1 | 0 | 0 |
| 3–130067 | 3 | 1 | 6 | 0 | 0 |
| 4–320003 | 4 | 4 | 6 | 2 | 1 |
| 5–320071 | 5 | 7 | 2 | 3 | 0 |

*Note*: The degree of subordination is identified by the number of clauses per *T*-unit or fragment (e.g. 2 = two clauses per *T*-unit or fragment).

*T*-unit) was observed, higher-level learners' speech contains a higher degree of subordination (i.e. three or four clauses per *T*-unit).

It is possible that these measures are useful only with longer stretches of text, as they have previously been used mainly in the analysis of written discourse, and it was on the basis of those studies that Wolfe-Quintero *et al*. (1998) recommended their use. Vermeer (2000) has pointed out similar problems in the use of ratio data in the assessment of vocabulary.

Finally, it is possible that there is a discrepancy between, on the one hand, the global proficiency scale used for the initial level assignment and the features of performance commented on by expert EAP specialists, and what is known from research in second language acquisition on the nature of second language development on the other. As Brindley (1998) points out, global rating scales describing features of 'real life' performance in specific contexts of language use based on systematic observation and documentation of language performance, and even (as here) incorporating expert assessors' comments, are not based on a theory of second language learning. In other words, there is no linguistic basis for positing the proposed hierarchies in those global rating scales (Brindley 1998: 118). In the present study, we examined how features of performance vary according to the level initially assigned by a global rating scale, and as the measures we employed in the analysis were adapted from SLA studies, our study raises again the potential significance of this gap.

## CONCLUSION

The present study investigated a number of features of learner performance on EAP speaking tasks and compared them with proficiency levels assigned by raters. We found that performance levels could be distinguished in terms of a range of performance features oriented to by raters, and as much if not

more by production features such as fluency and pronunciation and by vocabulary knowledge, as by grammatical features.

The results have strong implications for scale development, test preparation and teaching/learning in general. The detailed information gained on various aspects of learner performance at different levels provides a potential sketch of what may distinguish performance at various levels of achievement in language learning. Moreover, while the literature on the analysis of test discourse and interlanguage development has tended to focus on a limited range of features, the present study has extended the scope of this research by providing a cross-sectional investigation of a far more extensive range of features (e.g. including phonological features) at the same time.

The finding that a range of features including production features and vocabulary knowledge appear to distinguish different levels of learner proficiency indicates that an exaggerated emphasis on grammatical accuracy, reflected in the attitudes and behavior of many learners and teachers, is misplaced. In developing a speaking scale and in training assessors, it is recommended that the focus should include not only grammatical accuracy but also other features, in particular, production features and vocabulary knowledge. As well, in test preparation and learning/teaching more generally, teachers should encourage students to focus their attention on each of these features.

Overall, this study has advanced our understanding of the relative contribution to overall oral proficiency of particular features of learner production, and has given an insight into the range of factors balanced by raters in their determination of overall proficiency levels. The study needs to be extended to a broader range of learner populations and with learners engaged in a wider range of speaking tasks, particularly interactive ones; and a wider range of discourse features needs to be considered. The proficiency range of the participants under investigation in the present study is relatively narrow (mostly advanced learners who have studied English for a considerable amount of time) and the content of the tasks are relatively specific (i.e. EAP). Future investigations of a range of discourse features for learners of different proficiency levels using a wider variety of speaking tasks are required.

## ACKNOWLEDGEMENTS

An earlier version of this paper was presented at the AAAL conference in Arlington, Virginia, March 22–3, 2003. The research project reported in this paper was funded by Educational Testing Service (ETS) for the work of the TOEFL Speaking Team. We gratefully acknowledge the assistance of the people who participated in this study. We are also grateful to the following graduate students who, as research assistants, played an integral role in carrying out the discourse analyses: Michael Fitzgerald Jr., Felicity Gray, Daphne Huang, Clare Hutton, Debbie Loakes, Erich Round, Yvette Slaughter, Mary Stevens, and Susan Yi Xie. Thanks are also due

to Natalie Stephens for the immense amount of transcription that she undertook for this project, to Margaret Donald of the University of Melbourne Statistical Consulting Centre for statistical advice, and to Debbie Wallace for dealing with the numerous administrative requirements of the project. We are also thankful for Lourdes Ortega for her valuable comments on the interpretation of the data. Finally, we are grateful to Mary Enright of ETS for her assistance, advice, and encouragement in designing and completing this research.

## REFERENCES

**ACTFL Proficiency Guidelines.** 1985. Hasting-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

**ACTFL Proficiency Guidelines.** 1999. Hasting-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

**Adams, M. L.** 1980. 'Five co-occurring factors in speaking proficiency' in J. Firth (ed.): *Measuring Spoken Proficiency*. Washington, DC: Georgetown University Press, pp. 1–6.

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

**Brindley, G.** 1986. *The Assessment of Second Language Proficiency: Issues and Approaches.* Adelaide: National Curriculum Resource Centre.

**Brindley, G.** 1998. 'Describing language development? Rating scales and SLA' in L. Bachmann and A. Cohen (eds): *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, pp. 112–40.

**Brown, A., N. Iwashita, T. McNamara,** and **S. O'Hagan.** 2002. 'Getting the balance right: Criteria in integrated speaking tasks.' Paper presented at the 24th Language Testing Research Colloquium, Hong Kong, December 12–15.

**Brown, A., N. Iwashita,** and **T. McNamara.** 2005. *An Examination of Rater Orientations and Test Taker Performance on English for Academic Purposes Speaking Tasks*. (Monograph Series MS-29). Princeton, NJ: Educational Testing Service.

**Canale, M.** and **M. Swain.** 1980. 'Theoretical bases of communicative approaches to second language teaching and testing,' *Applied Linguistics* 1/1: 1–47.

**Cobb, T.** 2002. *The Web Vocabulary Profiler*. http://www.er.uqam.ca/nobel/r21270/textools/web_vp.html

**De Jong, J. H. A. L.** and **L. W. van Ginkel.** 1992. 'Dimensions in oral foreign language proficiency' in J. H. A. L. De Jong (ed.): *The Construct of Language Proficiency*. Philadelphia: John Benjamin, pp. 112–40.

**Douglas, D.** 1994. 'Quantity and quality in speaking test performance,' *Language Testing* 11/2: 125–44.

**Douglas, D.** and **L. Selinker.** 1992. 'Analysing oral proficiency test performance in general and specific purpose contexts,' *System* 20: 317–28.

**Douglas, D.** and **L. Selinker.** 1993. 'Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants' in D. Douglas and C. Chapelle (eds): *A New Decade of Language Testing Research.* Alexandria, VA: TESOL Publications, pp. 235–56.

**Foster, P.** and **P. Skehan.** 1996. 'The influence of planning and task type on second language performance,' *Studies in Second Language Acquisition* 18: 299–323.

**Freed, B.** 2000. 'Is fluency, like beauty, in the eyes (and ears) of the beholder?' in H. Riggenbach (ed.): *Perspectives on Fluency*. Ann Arbor: The University of Michigan Press, pp. 243–65.

**Fulcher, G.** 1996. 'Does thick description lead to smart tests? A data-based approach to rating scale construction,' *Language Testing* 13/2: 208–38.

**Galloway, V.** 1987. 'From defining to developing proficiency. A new look at the decision' in H. Byrnes (ed.): *Defining and Developing Proficiency: Guidelines, Implementations, and Concepts*. Lincolnwood, IL: National Textbook Company, pp. 25–73.

**Higgs, T.** and **R. Clifford.** 1982. 'The push towards communication' in T. V. Higgs (ed): *Curriculum, Competence, and the Foreign Language Teacher*. Lincolnwood, IL: National Textbook Company, pp. 57–79.

**Hunt, K. W.** 1970. 'Syntactic maturity in school children and adults,' *Monographs of the Society for Research in Child Development* 35/1: (1, Serial No. 134).

**Iwashita, N., T. McNamara,** and **C. Elder.** 2001. 'Can we predict task difficulty in an oral

proficiency test? Exploring the potential of an information processing approach to task design,' *Language Learning* 21/3: 401–36.

**Lee, J.-W.** and **D. Schallert.** 1997. 'The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL Context,' *TESOL Quarterly* 31/4: 713–39.

**Lee, Y.-W.** 2005. *Dependability of Scores for a New ESL Speaking Test: Evaluating Prototype Tasks*. (Monograph Series MS-28). Princeton, NJ: Educational Testing Service.

**Lumley, T.** 2005. *Assessing Second Language Writing: The Rater's Perspective*. Frankfurt: Peter Lang.

**McNamara, T.** 1990. 'Item response theory and the validation of an ESP test for health professionals,' *Language Testing* 7/1: 52–75.

**McNamara, T.** 1996. *Measuring Second Language Performance*. London and New York: Longman.

**McNamara, T., K. Hill** and **L. May.** 2002. 'Discourse and assessment,' *Annual Review of Applied Linguistics* 22: 221–42.

**MacWhinney, B.** 1999. *The CHILDES Project: Tools for Analyzing Talk* 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.

**Magnan, S.** 1988. 'Grammar and the ACTFL oral proficiency interview: Discussion and data,' *The Modern Language Journal* 72/3: 266–76.

**Mehnert, U.** 1998. 'The effects of different lengths of time for planning on second language performance,' *Studies in Second Language Acquisition* 20/1: 83–106.

**O'Loughlin, K.** (2001) *The Equivalence of Direct and Semi-Direct Speaking Tests*. Cambridge: Cambridge University Press.

**Ortega, L.** 1999. 'Planning and focus on form in L2 oral performance,' *Studies in Second Language Acquisition* 21/1: 109–48.

**Ortega, L., N. Iwashita, S. Rabie,** and **J. Norris.** in progress. *A Multi-lingual Comparison of Syntactic Complexity Measures and their Relationship to Foreign Language Proficiency*. Honolulu, Hawai'i: University of Hawai'i, National Foreign Language Resource Center.

**Pica, T.** 1983. 'Methods of morpheme quantification: Their effect on the interpretation of second language data,' *Studies in Second Language Acquisition* 6/1: 69–79.

**Robinson, P.** 1995. 'Task complexity and second language narrative discourse,' *Language Learning* 45: 141–75.

**Rommark, K.** 1995. *Xwaves*. Los Angeles: The University of California, Los Angeles. http://www-ssc.igpp.ucla.edu/~bryan/xwaves

**Shohamy, E.** 1994. 'The validity of direct versus semi-direct oral tests,' *Language Testing* 16: 99–123.

**Skehan, P.** and **P. Foster.** 1999. 'The influence of task structure and processing conditions on narrative retellings,' *Language Learning* 49: 93–120.

**Strauss, A.** and **J. Corbin.** 1994. 'Grounded theory methodology: An overview' in N. Denzin and Y. S. Lincoln (eds): *Handbook of Qualitative Research*. London: Sage.

**Towell, R., H. Hawkins** and **H. Bazergui.** 1996. 'The development of fluency in advanced learners of French,' *Applied Linguistics* 17/1: 84–119.

**Van Lier, L.** 1989. 'Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation,' *TESOL Quarterly* 23/3: 489–508.

**Vermeer, A.** 2000. 'Coming to grips with lexical richness in spontaneous speech data,' *Language Testing* 17/1: 65–83.

**Wigglesworth, G.** 1997. 'An investigation of planning time and proficiency level on oral test discourse,' *Language Testing* 14/1: 85–106.

**Wolfe-Quintero, K., S. Inagaki,** and **H.-Y. Kim.** 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity* Technical report #17. Honolulu: Second Language Teaching & Curriculum Center, The University of Hawai'i.