Relating PhonePass\textsuperscript{TM} Overall Scores to the Council of Europe Framework Level Descriptors

John H.A.L de Jong\textsuperscript{1} & Jared Bernstein\textsuperscript{2}

\textsuperscript{1}Language Testing Services, Velp, The Netherlands  john.hal.dejong@wxs.nl
\textsuperscript{2}Ordinate Corporation Menlo Park, CA 94025  jared@ordinate.com

Abstract

This study is a preliminary report on an experiment relating PhonePass SET-10 scores to the scale of level descriptors in the Council of Europe Framework. This scale describes the content and level of second language proficiency from a functional communicative perspective. Speech samples from each of 121 non-native English speakers were (1) scored in SET-10, the automatic speaking test and (2) rated under the European Framework through the conciliation of three independent raters. Rater reliability in using the Council of Europe scale and the comparability of the human and automatic measures are reported.

1. Introduction

1.1 European Framework

The Council of Europe has put forward a descriptive system for the elaboration of language syllabuses, curriculum guidelines, examinations and textbooks. [1]. The Common European Framework (henceforth: CEF) has already obtained wide recognition throughout Europe in political and educational contexts and has been adopted officially as the preferred descriptive system in the European Union. The CEF describes what learners have to learn in order to use a language for communication and what knowledge and skills they have to develop. The CEF also defines levels of proficiency, which allow learners’ progress to be measured at each stage of learning [2], [3] and [4]. A large number of European countries have adopted the CEF level descriptors as a basis for their educational systems or are preparing to do so. The CEF also impacts on curricula and certificates for private language schools.

Ever since the late sixties the Council of Europe has focused on a modular approach to language learning. The Council recognized the necessity of dividing the process of language acquisition into smaller units in order to describe and acknowledge the increase in individuals' language proficiency. This resulted in the first major publication: the 'Threshold Level' authored by Van Ek [5]. A thorough revision of the English version was published in 1991 [6].

The CEF is based on the principles set out in the Threshold Level but incorporates it as one of a number of progressive stages in language acquisition, allowing to define descriptive scales for communicative competencies (e.g., spoken or written interaction). These are broken down into a range of language activities in relation to specific domains (e.g., addressing an audience, obtaining goods), and requiring particular strategies (e.g., planning, repairing), pragmatics (e.g., fluency) and linguistic control (e.g., vocabulary range, grammatical accuracy). In effect the scales divide the assumed underlying continuum of language development into a number of bands or levels. Three broad levels are defined: Basic, Independent, and Proficient User and labeled A, B and C respectively. Each of these three levels is divided primarily into two finer levels, A1 and A2, B1 and B2, C1 and C2, but can in principle be subdivided further given the continuous nature of the underlying variable. To illustrate the scales, descriptors have been developed and empirically validated for the six basic levels A1 through C2 [3]. In the work reported here a global scale for rating interactive oral communication was used (see appendix).

1.2 PhonePass Set-10

The PhonePass system delivers automated tests over the telephone. Set-10 is a 10-minute version that measures speaking and listening skills in interaction. PhonePass tests are scored automatically by a computer-based system using speech recognition technology. Each test item requires the candidate to understand a spoken utterance and speak in response to it. Spoken performances are scored according to the linguistic content of the responses (60%) and according to the manner of producing the responses (40%). The manner-of-production scores are calculated from a set of acoustic base measures on segments, words and phrases, which are scaled with IRT methods or parametrically combined to optimize fit to human listener judgments. PhonePass tests measure facility in spoken English, which includes the ease and immediacy in understanding and producing basic conversational English. PhonePass tests tap core skills, those that enable a person to understand spoken language about everyday topics and respond intelligently at a native conversational pace. Studies of the PhonePass test indicate an overall test score reliability of 0.94 [7]. Overall PhonePass scores are reported on a scale from 2 to 8.

2 Method

2.1 Subjects

Subjects were sampled from the total database containing several thousand subjects who have taken SET-10 through the years. The sample was selected to contain a minimum of 100 subjects and to optimally agree with the following constraints:
- about an equal representation of male and female subjects;
- a representation from a range of language backgrounds including at least Arabic, Chinese, Japanese, and Spanish;
- a representation of subjects who have known scores on a number of other measures of proficiency in English as a foreign language (OPI, TSE, TOEFL, TOEIC).

These constraints yielded a sample containing 121 subjects. The sample was somewhat unbalanced by gender with 47% males, 25% females and 28% gender unknown. In total 22 different language backgrounds were represented: languages from East Asia, the Middle East, South America and from both Eastern and Western European countries. The sample included 51 people that had OPI scores, 30 with TSE scores, 20 with TOEIC listening scores, and 20 with TOEFL listening scores. No constraints were put on age and educational background. Subjects
varied widely in age - from 18 to 72, with an average age of 33 and about 90% had some form of academic education.

2.2 Instruments

2.2.1 PhonePass Set-10
PhonePass Set-10 consists of 5 parts. Part A (8 items) involves reading aloud 8 out of 12 printed sentences in the order requested. Part B (16 items) requires the examinee to repeat spoken sentences verbatim. Part C, (16 items) requires listening to words and producing 'opposites'. In Part D (16 items) the examinee listens to a question and responds with a short answer. Part E (2 items) involves listening to a question and producing long (30 seconds) answers. The overall PhonePass score is computed from responses on Parts A through D only. Part E is not scored automatically but can serve evaluation, benchmarking and identification purposes.

2.2.2 Council of Europe Oral Interaction Scale
In order to cover the full range of potential development as well as a wide range of aspects of conversational skill, a scale of oral interactional proficiency according to the European Framework was constructed. The scale was based on the Overall Spoken Interaction scale, but was rewritten to incorporate elements from the CEF Strategic, Pragmatic and Linguistic scales. The scale was written to contain the 6 basic levels used in the European Framework: A1, A2, B1, B2, C1 and C2 (see Appendix). As the lowest level (A1) still assumes some, albeit minimal, proficiency, a zero level was introduced to assign to subjects not reaching this minimal level or providing no evidence of proficiency at all (e.g. silence or avoiding behavior).

2.3 Procedures
Three raters were contracted from three different European countries: Netherlands, Switzerland and the UK. All three raters were well acquainted with the Council of Europe Framework and one rater is regularly involved as a professional rater in examinations of English as a foreign language. The raters were trained using two sets of examples. The first set consisted of video-recordings of an oral task performed by subjects at each of the six levels of the European Framework. Thus raters were made aware of the meaning of levels independent of the rating task to be performed in the experiment. The second set consisted of six digitized voice recordings of subjects performing one of the two open-ended tasks within PhonePass Set-10 (Part E). Subjects were selected to represent a range of levels, but not all of them different. Raters played these samples using the Windows media-player and were to rate them using the European Framework-based scale that was developed for the project. After rating them independently, raters discussed their results via e-mail and negotiated on discrepancies, thus acquiring a common feeling for the distinctive elements to be taken into account in assigning European levels to subjects.

The rating sample for each rater consisted of 121 subjects responding to the two open-ended questions of Part E. As subjects get random selections from the PhonePass item bank, 18 different items were involved. To control for intra-rater consistency, raters were presented with about 30% of the samples twice.

The rating samples were stored in a database. Raters called into the database via the telephone at their own convenience and rated as many samples per call as they wished. The whole procedure was fully automatic. Raters listened to the samples and punched in their ratings on the telephone keypad.

Data were analyzed using the multifaceted Rasch model as implemented in the computer program FACETS [8]. This program allows estimating rater severity, subject ability, and item difficulty in a single analysis. The model assumes a single underlying dimension, where values along this dimension are expressed on a logit scale. Interrater reliability was estimated using the computer program OVERTON [9] an ANOVA-procedure. Interrater reliability in this program is estimated as an intra-class correlation.

3 Results

3.1 Ratings according to the Council of Europe Framework
To evaluate the quality of the rater judgments Figure 1 presents the placement of the cut-offs between subsequent levels of the Council of Europe Framework as estimated from the ratings provided by the three independent raters.

Figure 1 shows that raters differed in overall severity and in their estimates of the length of the scale; where Rater 1 estimates the difference between the highest and the lowest cut-off as 7.5 on the logit scale, Rater 3 uses more than 9 logits. Rater 2, with 8 logits, takes a middle position. Raters are in quite good agreement on the placement of the cut-offs between levels A1 and A2 and between levels B2 and C1, and achieve moderate agreement on the cut-off between levels C1 and C2. Larger differences in their estimates for the cut-offs 0/A1, B1/B2 and B2/C1 are on no occasion larger than the difference between two levels, which in practice means that no subject will be placed further than one level apart by any two judges.

The full data matrix of 121 candidates, by 2 items, by three raters contains 726 data points. To assess intra-rater reliability raters were made to generate two independent scores for 111 random data points. Mean and standard deviation of these two independent ratings were similar up to the third decimal. The correlation between the two ratings was .95.

Summing over the two items that were rated for each candidate generated a total CEF score per candidate. The total variance in these total scores was analyzed to be attributable for 81% to candidates, for 7% to raters, with a residual of 13% (Due to rounding errors these values do not add up to exactly 100%). The interrater reliability was estimated at .949, which was considered sufficient in order to take the average over judgements as a stable estimate of candidates' positions on the CEF-scale.
Figure 2 presents the distribution of subjects over the Council of Europe levels using cut-offs averaged over the three raters. Figure 2 shows that a substantial number of subjects produced responses placing them at the zero level. Also, assuming a normal distribution for the remainder of the subjects, there appears to be an over-representation of subjects at the lower levels. Subjects’ average level was estimated a little below level A2 with a standard deviation of 1.4 levels. Reliability of these estimates was .95.

Of the 18 items used in this study, 7 items were estimated at the B1 level, 9 at the B2 level and 2 at the C1 level, suggesting as could be expected that SET-10 is most appropriate at moderate and higher levels of ability.

3.2 Relation between Council of Europe Levels and PhonePass Set-10 scores

Figure 3 shows a mapping of students’ estimated levels on the CEF scale against their Overall PhonePass Set-10 scores. As can be seen, the relation is not strictly linear resulting in a moderate correlation of 0.84. However, this is to be expected as North [3] has reported that the CEF levels are not equidistant on the ability scale. A polynomial trend-line has been drawn as best fit between the two measures.

Figure 4 shows the best prediction of the CEF level cut-offs from the PhonePass Set-10 score reporting scale. Figure 4 suggests that Table 1 may be used as a transformation table to predict CEF levels from PhonePass Set10- scores.

Table 1: Score Transformation Table

<table>
<thead>
<tr>
<th>PhonePass Set 10 Score</th>
<th>Prediction of CEF Level in Oral Interaction Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0 - 3.9</td>
<td>&quot;no evidence&quot;</td>
</tr>
<tr>
<td>4.0 - 4.9</td>
<td>A1</td>
</tr>
<tr>
<td>5.0 - 5.5</td>
<td>A2</td>
</tr>
<tr>
<td>5.6 - 6.1</td>
<td>B1</td>
</tr>
<tr>
<td>6.2 - 6.7</td>
<td>B2</td>
</tr>
<tr>
<td>6.8 - 7.2</td>
<td>C1</td>
</tr>
<tr>
<td>7.3 - 8.0</td>
<td>C2</td>
</tr>
</tbody>
</table>

4 Conclusion

The procedure used to score subjects responding to the two open-ended PhonePass Set-10 questions on the Council of Europe scale produced reliable estimates of the subjects’ position on this scale. These estimates showed reasonable correspondence with the scores for these subjects in responding generated automatically from their responses to the other set items within the PhonePass Set-10 test of Spoken English. PhonePass Set10 scores can therefore be used to predict CEF levels on the CEF scale of Oral Interaction Skills with reasonable accuracy. Further research is called for to investigate whether PhonePass Set-10 scores form equally reliable predictors for CEF levels over the complete scale and whether components in the PhonePass Set-10 overall score predict the CEF levels differentially.

5 References


Appendix

Holistic Scale for Assessing Speaking onto CEF Levels as used by the raters

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conveys finer shades of meaning precisely and naturally.</td>
<td>Shows fluent, spontaneous expression in clear, well-structured speech.</td>
<td>Relates comprehensibly main points he/she wants to make on familiar matters.</td>
<td>Relates basic information on, e.g. work, background, family, free time etc.</td>
<td>Makes simple statements on personal details and very familiar topics.</td>
<td>Candidate performs below level defined as A1</td>
</tr>
<tr>
<td></td>
<td>Can express him/herself spontaneously at length with a natural colloquial flow. Consistent grammatical and phonological control of a wide range of complex language, including appropriate use of connectors and other cohesive devices.</td>
<td>Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Clear, natural pronunciation. Can vary intonation and stress for emphasis. High degree of accuracy; errors are rare. Controlled use of connectors and cohesive devices.</td>
<td>Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Pronunciation is intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur. Reasonably accurate use of main repertoire associated with more predictable situations. Can link discrete, simple elements into a connected sequence.</td>
<td>Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Pronunciation is generally clear enough to be understood despite a noticeable foreign accent. Uses some simple structures correctly, but still systematically makes basic mistakes. Can link groups of words with simple connectors like &quot;and&quot;, &quot;but&quot; and &quot;because&quot;.</td>
<td>Can manage very short, isolated, mainly pre-packaged utterances. Much pausing to search for expressions, to articulate less familiar words. Pronunciation is very foreign.</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>C2</td>
<td>5</td>
<td>B1</td>
<td>B2</td>
<td>A2</td>
<td>A1</td>
</tr>
<tr>
<td>5</td>
<td>C1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>&lt;A1</td>
</tr>
<tr>
<td>4</td>
<td>B2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>dirigible</td>
</tr>
<tr>
<td>3</td>
<td>B1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>&lt;A1</td>
<td>Candidate's response cannot be graded:</td>
</tr>
<tr>
<td>2</td>
<td>A2</td>
<td>1</td>
<td>0</td>
<td>&lt;A1</td>
<td>Candidate's response cannot be graded:</td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>A1</td>
<td>0</td>
<td>&lt;A1</td>
<td>Candidate's response cannot be graded:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Insufficient evidence to decide on score category
- Silence
- Just a few loose words, no meaningful idea expressed
- Single utterance no longer than 6 words
- Just repeat of question
- Irrelevant response
- Avoiding behaviour: 'I don't know'
- Unintelligible response