ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 3, No. 2, pp. 295-305, March 2012

doi:10.4304/jltr.3.2.295-305

A Model for Training Teachers to Identify Common Reference Levels in Written Production Activities

José Cuadrado-Moreno
Department of Philology and Translation, Pablo de Olavide University, Seville, Spain
Email: josecuadrado@upo.es

María Reyes-Fernández
Department of Philology and Translation, Pablo de Olavide University, Seville, Spain
Email: mreyfer@upo.es

***Abstract*—This paper presents a teacher training programme, the purpose of which was to train a secondary education teacher of English to identify common reference levels in written production activities. The first phase of this programme was based on the familiarisation and standardisation phases in North *et al.* (2009) and allowed the participant to familiarise herself with the Common European Framework of Reference (Council of Europe, 2001) and to reliably identify common reference levels of English in both standardised and local performances. During the second phase, the participant assessed the written production competence of a sample of Year-9 pupils of English and assigned a common reference level of English to each pupil. After analysing her scores with generalizability theory, several high reliability coefficients were obtained. Such results demonstrate that this training programme, including its assessment procedure, can be adopted as a model for training secondary education teachers of English to identify common reference levels in written production activities.**

***Index Terms*—generalizability, reliability, Common European Framework of Reference, writing, training programme**

I. Introduction

At present, European language policy makers and planners are introducing the *Common European Framework of Reference* (CEFR) (Council of Europe, 2001) into the foreign-language curricula. One of the areas which is receiving a great deal of attention is assessment, since the CEFR has established six proficiency levels (called common reference levels, CRLs), upon which objectives, contents, methodology, learning activities and assessment should be built. Training courses are consequently being organised so that teachers can learn to assess pupils' productive performances according to the CRLs. However, the reliability of the scores provided by the participants in such training activities (and, therefore, the effectiveness of the training programme) has been relatively pushed into the background.

This paper presents a teacher training programme for a secondary-education teacher of English, the purpose of which was to enable the participant to identify the CRLs of local writing performances in a reliable way.

II. Literature Review

The *Common European Framework of Reference* is a document published by the Council of Europe in 2001. For Little (2006: 167),

The CEFR is a descriptive scheme that can be used to analyse L2 learners' needs, specify L2 learning goals, guide the development of L2 learning materials and activities, and provide orientation for the assessment of L2 learning outcomes.

The CEFR introduces six levels of language proficiency (A1, A2, B1, B2, C1 and C2), called *common reference levels* (CRLs), which are "appropriate to the organisation of language learning and the public recognition of achievement" (Council of Europe, 2001: 22-23). The CEFR describes explicitly the CRLs in terms of single holistic paragraphs (2001: 24, Table 1), major categories of language use (2011: 26-27, Table 2), communicative language activities (2001: chapter 4) and communicative language competences (2011: chapter 5). The CEFR has been deeply influenced by Brian North's doctoral dissertation about the development of a scale of language proficiency (North, 2000).

One of the problems of the CEFR has been the fact that many of the statements describing the common reference levels are based on the *perception* of the development of language proficiency by groups of teachers of English in Switzerland (North, 2000: 290), not on a rigorous description of actual performance or second language acquisition

theory (Fulcher, 2010: 114). North himself was aware of this problem with his rating scales in that, before the publication of the CEFR, North and Schneider (1998: 219-220) had already stated that

[T]here is no guarantee that the description of proficiency offered in a scale is accurate, valid or balanced […] Raters may actually be trained to think the same; inter-rater reliability correlations of over 0.8 are common […] But the fact that people may be able to use such instruments with surprising effectiveness doesn't necessarily mean that what the scales say is valid. Furthermore, with the vast majority of scales of language proficiency, it is far from clear on what basis it was decided to put certain statements at Level 3 and others at Level 4 anyway.

This situation has given rise to research projects looking for the criterial features which should allow examiners to distinguish among the various CRLs (Carlsen, 2010; Hendriks, 2008; Salamoura & Saville, 2010). Another problem which has been mentioned in connection to the CEFR tables has been the lack of clarity in its descriptors. The CEFR includes a series of guidelines to write descriptors describing levels in language attainment, one of which deals with the clarity of descriptors: "Descriptors should be transparent, not jargon-ridden" (Council of Europe, 2001: 206). However, in order to avoid the dangers of the overuse of jargon, the CEFR authors have often opted for vagueness. See, for example, the descriptor for B1 in the scale for grammatical accuracy (Council of Europe, 2001: 114):

B1	• Communicates with reasonable accuracy in familiar contexts. • Generally good control though with noticeable mother tongue influence. • Errors occur, but it is clear what he/she is trying to express.

The problem with descriptors such as this one is that, from our point of view, they are rather ambiguous: the assessor is given too much freedom to interpret phrases like "with reasonable accuracy" or "generally good control" and, therefore, to determine the examinee's CRL. This amount of freedom of interpretation may cause a negative index of interrater agreement (Tinsley & Weiss, 1975: 359).

The CEFR has been subsequently developed by a huge number of documents, among which two stand out for our purposes: *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Writing Tasks: Pilot Samples* (Council of Europe, s.d.) and North *et al.* (2009) (henceforth referred to as the *Manual*). Council of Europe (s.d.) is a collection of sample writing performances, to which CRLs have been assigned, while the *Manual* is a document to help the providers of examinations to develop, apply and report transparent, practical procedures in a cumulative process of continuing improvement in order to situate their examination(s) in relation to the Common European Framework (CEFR) (North *et al.*, 2009: 1).

In order to relate a language examination to the CEFR, the *Manual* recommends the following five-phase process:

1. *Familiarisation*, the purpose of which is "to ensure that participants in the linking process have a detailed knowledge of the CEFR, its levels and illustrative descriptors" (North *et al.*, 2009: 10).

2. *Specification*, that is, "a self-audit of the coverage of the examination (content and tasks types) profiled in relation to the categories presented in CEFR Chapter 4 […] and CEFR Chapter 5" (North *et al.*, 2009: 10).

3. *Training in standardisation/benchmarking*, the activities of which aim "(a) to help panellists to implement a common understanding of the CEFR levels; (b) to verify that such a common understanding is achieved, and (c) to maintain that standard over time" (North *et al.*, 2009: 37).

4. *Standard setting among the different CRLs*.

5. *Validation*, which is related to "the body of evidence put forward to convince the test users that the whole process and its outcomes are trustworthy" (North *et al.*, 2009: 90).

Given the relevance of the phases of familiarisation and training in standardisation/benchmarking for the participant's training process, these are presented more extensively. The *Manual* recommends carrying out the following activities during the familiarisation phase in approximately three hours (North *et al.*, 2009: 23):

- Brief presentation of the CEFR Familiarisation seminar by the coordinator (30 minutes)
- Introductory activity (d–e) and discussion (45 minutes)
- Qualitative activity (f–g) including group work (45 minutes)
- Preparation for rating (h–i) (45 minutes)
- Concluding (15 minutes)

Table 1 presents the (slightly revised) time management which North *et al.* (2009: 46) proposed for the training activities to standardise the participants' judgments concerning written performance samples.

TABLE 1.
TIME MANAGEMENT FOR ASSESSING WRITTEN PERFORMANCE SAMPLES (NORTH *ET AL.*, 2009: 46).

Stages
Stage 1: Familiarisation (60 minutes)
Stage 2: Practice with standardised samples - Step 1: Illustration with approximately three standardised performances (60 minutes) - Step 2: Controlled practice with approximately three standardised performances (60 minutes) - Step 3: Free stage with approximately three standardised performances (60 minutes)
Stage 3: Benchmarking local samples - Individual rating and group discussion of circa three performances (60 minutes) - Individual rating of circa five more performances (60 minutes)

The *Manual* has been the base of numerous training programmes to standardise judgments concerning written performances in terms of CRLs, such as "Rating procedures in the assessment of written learner productions of the A-levels" (Quality Agency Meißen, Sachsen, 2007), "Rating written learner productions of the A-levels" (Institute for Teacher Training, Saarbrücken, 2008) or "Reconocimiento de los niveles comunes de referencia en actividades de expresión e interacción escritas y orales" (Centro de Profesores, Sevilla, 2011). However, a datum which is missing in these training programmes –but present in other programmes, such as Wigglesworth (1993) or Lumley & McNamara (1995)– is, as far as we know, a statistic which estimates, with a high degree of dependability, the participants' proficiency in identifying the examinees' productive and interactive written competences in terms of CRLs.

The process of rating a written performance is related to what in educational and psychological testing is called *reliability*, which has been defined as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999: 180) (*Standards*, 1999 henceforth). In fact, the 1999 *Standards* devoted Standard 2.1 to the reliability of scores: "For each score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard error of measurement or test information functions should be reported" (*Standards*, 1999: 31). In a sense, this paper can be considered to be a report of the reliability and the standard error of measurement of the scores provided by the trainee during the training programme.

At present, one of the most powerful approaches to reliability estimation is *generalizability theory* (GT), which Shavelson and Webb (1991: 1) defined as "a statistical theory about the dependability of behavioural measurements". The first paper where generalizability theory was presented was Cronbach, Rajaratnam & Gleser (1963), who defined it as "a theory regarding the adequacy with which one can generalize from one observation to a universe of observations" (1963: 137). For Cronbach, Rajaratnam & Gleser (1963: 145), investigators had to specify a universe of conditions of observation, over which they would generalize. In GT, the term *conditions* is applied to "particular test forms or stimuli, observers, occasions or situations of observation" (Cronbach, Rajaratnam & Gleser, 1963: 145), that is, the values of the variables which the test evaluator considers to influence the testing procedure, while a *facet* is the "characteristic of a measurement procedure such as task, occasion, observer that is defined as a potential source of measurement error" (Shavelson & Webb, 2005b: 99), i.e., the variable influencing the testing procedure. GT distinguishes between *generalizability studies* (G studies), which are used "to obtain estimates of variance components associated to the universe of admissible observations" (Brennan, 1992: 3), and *decision studies* (D studies), the purpose of which is to provide data "to design efficient measurement procedures for operational use or to provide information for making substantive decisions about objects of measurement" (Brennan, 1992: 3). GT is based on a series of assumptions:

1. Conditions are randomly selected from the universe of conditions (Cronbach, Rajaratnam & Gleser, 1963:147).
2. Conditions are specified (Cronbach, Rajaratnam & Gleser, 1963: 145).
3. Conditions are experimentally independent (Cronbach, Rajaratnam & Gleser, 1963: 145).
4. The scores assigned to the conditions are numbers on an interval scale.

In GT, a person *p*'s score in item *i* (X_{pi}) is broken down as follows:

$$X_{pi} = \mu \text{ (grand mean)} + \mu_p - \mu \text{ (person component} = v_p) + \mu_i - \mu \text{ (item component} = v_i) + X_{pi} - \mu_p - \mu_i + \mu \text{ (residual component} = v_{pi})$$

Where:

$$\mu = E_p E_i X_{pi} \text{ (the mean over both the population of persons and the universe of items)}$$

$$\mu_p = E_i X_{pi} \text{ (an examinee's mean score over the universe of items)}$$

$$\mu_i = E_p X_{pi} \text{ (the population mean for item } i)$$

$$X_{pi} - \mu_p - \mu_i + \mu \text{ = residual component}$$

The basic assumptions underlying the model in equation 1 are:

1. The expected value of each of these components over the population of persons and the universe of items is set equal to zero ($\underset{p}{E}\nu_p = \underset{i}{E}\nu_i = \underset{pi}{E}\nu_{pi} = 0$).

2. Each of the n_p persons is administered each of the n_i items, and the responses X_{pi} are obtained (this is a description of a $p \times i$ G study, the same type as the study below).

The total test variance can also be broken down into different *variance components* (Cornfield and Tukey, 1956: 926; Cronbach, Rajaratnam & Gleser, 1963: 151; Gleser, Cronbach & Rajaratnam, 1965: 408), which can be estimated by first applying an analysis of variance to the set of data and, then, the following formulae:

$$\sigma^2(p) = \mathrm{E}(\mu_p - \mu)^2 = \frac{MS_p - MS_{pi}}{n_i}$$

$$\sigma^2(i) = \mathrm{E}(\mu_i - \mu)^2 = \frac{MS_i - MS_{pi}}{n_p}$$

$$\sigma^2(pi) = \mathrm{EE}(X_{pi} - \mu_p - \mu_i + \mu)^2 = MS_{pi}$$

Among the data that may be obtained with a D-study we can mention:

1. An index Φ of dependability for domain-referenced interpretations, corrected for chance agreement (Brennan, 1980: 205; 2001: 35; Kane & Brennan, 1980: 118). Values of Φ close to or greater than .80 are considered to be dependable (Glasswell & Brown, 2003: 4).

2. Absolute error variance (Brennan, 1980: 198). The *absolute error* Δ_{pI} is the error involved in using an examinee's observed mean score as an estimate of his or her universe score (Brennan, 2001: 31).

3. Error variance for estimating μ using $\overline{X}$: The error variance involved in using the mean over the sample of both persons and items ($\overline{X}$) as an estimate of the mean over both the population of persons and the universe of items (μ)

$$\sigma^2(\overline{X}) = \sigma^2(P) + \sigma^2(I) + \sigma^2(PI) = \frac{\sigma^2(p)}{n'_p} + \frac{\sigma^2(i)}{n_i} + \frac{\sigma^2(pi)}{n'_p n_i}$$

4. Signal-to-noise ratio: Brennan & Kane (1977: 616) developed a general index to measure the precision of the criterion-referenced test, based on the concept of signal-to-noise ratio (*S/T ratio*). In telecommunication engineering, the signal-to-noise ratio indicates "the amount by which a signal exceeds the noise in a specified bandwith" (Freeman, 2004: 146). In a telecommunication system, the material to be transmitted requires a minimum S/N ratio in order to satisfy customers or to make the receiving instrument function within certain specified criteria. When Brennan & Kane (1977: 616) applied the concept of S/N ratio to criterion-referenced tests, $\sigma^2(p)$ (which is a function of the magnitude of the deviation $\mu_p - \mu$) was associated with the signal and the absolute error variance $\sigma^2(\Delta)$ with the noise. Therefore, the amount by which a signal exceeds the noise in a criterion-referenced test situation can be estimated by means of the proportion $\sigma^2(p)/\sigma^2(\Delta)$, which indicates the degree to which the test is precise (Brennan, 1980: 204; 2001: 47; Brennan & Kane, 1977: 616).

5. Index of the context-specific precision or error-tolerance ratio (*E/T*) (Kane, 1996). Brennan (2001: 48) presents the following formula for the estimation of the error-tolerance ratio:

$$E/T(\lambda) = \sqrt{\frac{\sigma^2(\Delta)}{\sigma^2(p) + (\mu - \lambda)^2}}$$

Where:

$E/T(\lambda)$ = the error-tolerance ratio for cut score λ.

λ = a cut score, expressed a proportion of correct items.

From this error-tolerance ratio, a reliability-like coefficient is:

$$\Phi(\lambda) = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\Delta)}$$

GT has also been applied to the assessment of writing tests (Brown & Bailey, 1984; Gebril, 2010; Llabre, 1978; Sudweeks, Reeve & Bradshaw, 2004; Swartz *et al*., 1999).

Bearing in mind the literature review and the purpose of this study, the research was guided by the following general research questions:

1. Is the trainee familiarised with the CEFR?

2. Can the trainee identify the pupils' common reference levels?

3. Are the trainee's scores reliable?

Since the trainee intended to assess the writing competence of a sample of Year-9 pupils, the specific research questions in this study were:

1. What aspects influence the variation in her scores?
2. Is the contribution of the test to the dependability of the decision procedure very high?
3. What is the recommended number of rating criteria to be used when assessing the writing competence in a population of Year-9 pupils?
4. How great is the error involved in using a pupil's observed scores as an estimate of this pupil's universe score?
5. Is the trainee's writing test a precise instrument to measure the pupils' writing competence?
6. How precise are the trainee's scores?
7. Is the writing test a precise instrument to measure the pupils' writing competence?

III. Methodology

In order to answer these research questions, a training programme was designed so that the trainee could familiarise herself with the CEFR and acquire experience in standardisation/scoring local samples of written performances in English. The programme used the model established in North *et al.* (2009) as a guideline, though several modifications were introduced:

1. Familiarisation (4 hours).
2. Practice with standardised samples
   a. Step 1: Explanation of rating criteria (2 hours).
   b. Step 2: Illustration (3 hours).
   c. Step 3: Controlled practice (1 hour and 30 minutes).
   d. Step 4: Free practice (1 hour and 30 minutes).

Since the trainee is a secondary education teacher, it was decided that she would have to recognize the common reference levels in a sample of local Year-9 pupils' performances. There was the problem that the Andalusian curriculum for secondary education (Andaluc á, 2007) does not include a general instructional objective for foreign language learning in terms of the common reference levels. The authors then decided to set the common reference level A2 as the goal to attain in English at the end of secondary education (Year 10).

The familiarisation phase was composed of the following activities:

1. Presentation of the training program (30 minutes).
2. Presentation of the CEFR (30 minutes).
3. Self-assessment of the trainee's linguistic competence in terms of the CEFR common reference levels.
4. Learning activity 1: Reordering the descriptors in the CEFR table "Overall written production" (Council of Europe, 2001: 61).
5. Learning activity 2: Reordering the descriptors in the CEFR table "Overall written interaction" (Council of Europe, 2001: 83).
6. Learning activity 3: Reordering the descriptors in CEFR "Table 2. Common Reference Levels: self-assessment grid" (Council of Europe, 2001: 26-27).

As *familiarisation with the CEFR* was defined as the ability to reconstruct CEFR tables, the null hypothesis h_0 to be tested in each of the familiarisation learning activities 4-6 was that there was no difference between the descriptor distribution in the corresponding CEFR table and the distribution made by the trainee. Since, as far as we know, there exist no data concerning the degree of familiarisation with the CEFR among Spanish secondary education teachers in terms of CEFR tables, the (non-parametric) sign test (traditional method) was used in order to test the above null hypotheses ($\alpha = 0.05$) (Cochran, 1937; Triola, 2006: 678-686; Wilcoxon, 1945).

During the phase of practice with standardised samples, the trainee was given some materials and asked to read the samples of written performances (ranging from A1 to A2 common reference levels) and to rate each criterion of written performance with 0, 1 or 2 (0 = A1-, 1 = A1, 2 = A2). The practice phase was composed of four steps:

1. Step 1 (*explanation*): The trainee was explained the concepts of grammatical range, lexical range, cohesion and descriptive writing, and was given the following materials:

a. For grammatical range: the clause structures presented in Salamoura & Saville (2010: 116) using Quirk, Greenbaum, Leech & Svartvik (1985)'s terminology (see Appendix A).

b. For lexical range: the word lists from Trim (1998: 157-176) (WLA1) and University of Cambridge Local Examinations Syndicate (2009c) (WLA2). Lexical range was divided into lexical range for A1 level and lexical range for A2 level on the basis of these vocabulary lists, and these terms were defined as follows:

$$\textit{Lexical range}_{\text{A1}} = \text{WLA1}$$
$$\textit{Lexical range}_{\text{A2}} = \text{WLA2} - \text{WLA1}$$

c. For coherence: the descriptors of the coherence criterion for levels A1 and A2 from Table C4 in North *et al.* (2009: 187) (see Appendix B).

d. For descriptive writing: the descriptors of the description criterion for levels A1 and A2 from Table C4 in North *et al.* (2009: 187) (see Appendix B).

2. Step 2 (*illustration*): Presentation of two CEFR illustrative standardised performances (A1 and A2) (Council of Europe, s.d.). The trainee was asked to rate each aspect of the written language use in each sample and to justify her ratings. Afterwards, feedback was provided.

3. Step 3 (*controlled practice*): Presentation of three CEFR (A1, A2 and B1) standardised samples from CEFTrain (s.d.). The trainee was asked to use Appendices A and B, WLA1 and WLA2 in order to rate four aspects of written language use (grammatical range, lexical range, cohesion and description) for three subjects by assigning a value (0 = A1–; 1 = A1; 2 = A2 or higher) to each aspect of the subject's written language use according to her estimation of the writer's aspect. Data were then collected and feedback provided.

4. Step 4 (*free activity*): Presentation of three local CEFR (A1-A2) standardised performances. The trainee was asked to rate, with the same material as during the controlled-practice phase, four aspects of written language use (grammatical range, lexical range, cohesion and description) for three subjects and to assign a value (0 =A1–; 1 = A1; 2 = A2 or higher) to each aspect of the subject's written language use. Later, data were collected and feedback was provided.

In order to assess the degree of consistency of the trainee's scores, the trainee repeated steps 3 and 4 fifteen days later (test-retest reliability) (Linn and Gronlund, 2000: 110). The collected data were then analyzed in order to test the null hypotheses h_0 that there was no difference between the first and the second rating. Since, as far as we know, there exist no data concerning the degree of familiarisation with the CEFR among Spanish secondary education teachers, the (non-parametric) sign test (traditional method) was used in order to test these hypotheses ($\alpha = 0.05$).

After the training programme was completed, the trainee rated 191 samples of written performances (of a population of 191 Year-9 pupils) from two secondary education schools in Montequinto (Dos Hermanas, Spain). These pupils took a writing test (Appendix C) and the trainee was asked to use the same rating procedure as the one used during the controlled-practice and free-activity phases.

TABLE 2.
ESTIMATED VARIANCE COMPONENTS FOR G STUDY IN THE UNIVARIATE ANALYSIS WITH A $P \times I$ DESIGN FOR THE TEST OF PRODUCTIVE AND INTERACTIVE COMPETENCE ($N_P = 191$, $N_I = 4$) (TWO-WAY ANOVA WITHOUT REPLICATION, $\Lambda = 0.05$)

Components	*Pupils (p)*	*Criteria (i)*	*Interaction (pi)*	*Total*
Df	190	3	570	763
SS	316.973	62.108	99.141	478.223
MS	1.668	20.702	0.173	
Estimated G study variance components	0.374	0.107	0.174	
Total variance (%)	57.03	19.54	31.62	
Estimated standard error	0.043	0.069	0.009	

TABLE 3.
RESULTS FOR RANDOM COMPONENTS D STUDY $P \times I$ DESIGN

$\widehat{\Phi}$	0.842
$\hat{\sigma}^2(\Delta)$	0.070
S/N ratio	5.310
$\hat{\sigma}^2(\bar{X})$	0.029

TABLE 4.
ESTIMATED $E/T(\Lambda)$ AND $\Phi(\Lambda)$ FOR D STUDY IN THE UNIVARIATE ANALYSIS WITH A $P \times I$ DESIGN FOR THE TEST OF WRITING PRODUCTIVE AND INTERACTIVE COMPETENCES ($N_P = 191$)

λ	$E/T(\lambda)$	$\Phi(\lambda)$
0.1	0.293	0.921
0.2	0.319	0.908
0.3	0.347	0.893
0.4	0.376	0.876
0.5	0.405	0.859
0.6	0.429	0.844
0.7	0.446	0.834
0.8	0.450	0.831
0.9	0.442	0.836
1	0.423	0.848

IV. ANALYSIS

A. Analysis of Trainee Data

The sign test procedure ($\alpha = 0.05$) provided the following results with the data about the familiarisation phase (Table 5), when testing the null hypotheses h_0 that there was no difference between the descriptor distribution in the corresponding CEFR table and the distribution made by the trainee during familiarisation learning activities 1-3 (R ós-Lorenzo, 2009):

TABLE 5.
DATA OBTAINED WITH THE FAMILIARISATION LEARNING ACTIVITIES

Activity	*Test statistic*	*Critical value on the left*	*Critical value on the right*
Learning activity 1	0	*	*
Learning activity 2	0	*	*
Learning activity 3	4	1	8

The asterisk indicates that, with that test statistic, it is impossible to obtain a value in the critical region and, therefore, the corresponding null hypothesis h_0 must be admitted. On the basis of the data collected in each familiarisation learning activity 1-3, it was concluded that there are not enough data to reject the null hypotheses h_0 that there was no difference between the descriptor distribution in the CEFR tables and the distributions made by the trainee. Put in a simpler way, it was inferred that the trainee was familiarised with the CEFR.

The data obtained during both the controlled practices and the free activities were also analysed with the sign test procedure. Table 6 presents the test statistics obtained and the corresponding critical values.

TABLE 6.
DATA OBTAINED DURING BOTH THE CONTROLLED PRACTICES AND THE FREE ACTIVITIES

Activity	*Test statistic*	*Critical value on the left*	*Critical value on the right*
Controlled practice 1	0	*	*
Controlled practice 2	0	*	*
Controlled practice 3	0	*	*
Free activity 1	0	*	*
Free activity 2	0	*	*
Free activity 3	0	*	*

In every learning activity, the trainee identified the writer's CRL and assigned the same scores to the writer's performances in both sessions. On the basis of the data collected during the controlled practice and the free activities, it was concluded that there were no difference between the scores in both sessions. In other words, it was inferred that (i) the trainee could identify the writer's CRL and (ii) her scores were highly reliable.

B. Analysis of Pupil Data

a. Analysis of the Data Obtained in the G Study

On the basis of the relative sizes of the estimated G-study variance components associated with pupils, rating criteria and the interaction between pupils and rating criteria, it can be inferred from Table 2 that:

1. The largest estimated variance component was the one associated with the individual differences among pupils' writing competence, which accounts for 57.03% of the total variance.

2. The second largest estimated variance component was the one associated with the interaction pupils-rating criteria, which accounts for 31.62% of the total variability. This variability reflects the interaction and influence of factors such as the pupils' educational histories and other residual characteristics like noise, light, pupils' health conditions, etc., which, since they have not been measured, are conflated in this variance component. For example, a particular rating criterion may have been easier for those pupils who have received additional feedback on that criterion during their instruction.

3. The smallest estimated variance component was the one associated with the rating criteria, which accounts for 19.54% of the total variance. This percentage of the total variability is associated with the difficulty of the rating criteria.

The total variability of pupils' test scores was thus influenced by (ordered from the most to the least prominent) (i) the overall writing competence, (ii) the pupils' degree of familiarity with the rating criteria and other environmental characteristics, and (iii) the difficulty of each rating criterion.

b. Analysis of the Data Obtained in the D Study

We can see in Table 3 that the Φ index is over .80. Thus, the rating procedure contributes a great deal to the dependability of the decisions or estimations based on this rating procedure. The absolute error variance $\hat{\sigma}^2(\Delta)$ is extremely low (0.07), which indicates that the error involved in using an examinee's observed mean score as an estimate of his or her universe score is also extremely low. Fig. 1 shows the influence of different numbers of rating criteria on the estimations of the Φ index and the absolute error variance. The Φ index shows a value over .80 just when the number of rating criteria is equal to or higher than 4, and, as for the absolute error variance, there exists a very small difference among its values when the number of rating criteria is equal to or higher than 4. Therefore, for the rating procedure with this trainee, it is concluded that four is an optimum number of rating criteria.

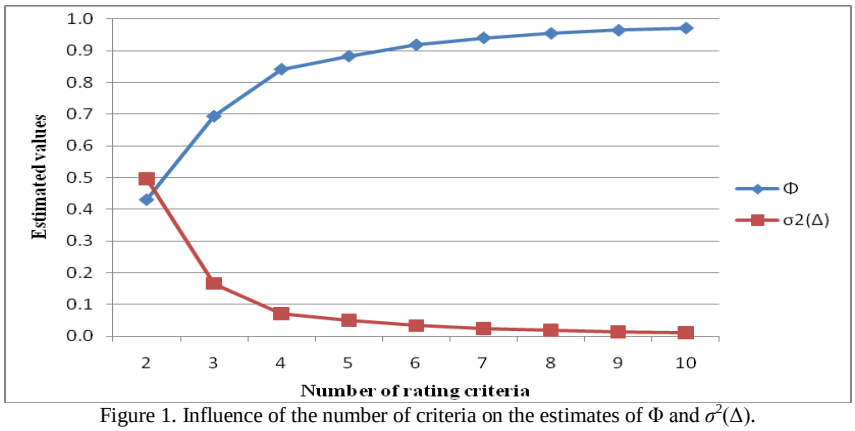

Figure 1. Influence of the number of criteria on the estimates of Φ and $\sigma^2(\Delta)$.

The signal-to-noise ratio is 5.31, which means that the magnitude of the estimated variance components for persons is more than five times the magnitude of the estimated criterion-referenced absolute error. Thus, the written test is precise to measure the pupils' writing competence.

With respect to the index of tolerance for error for this particular population and test administration, Fig. 2 shows the estimated values of *E/T*(λ) and Φ (λ) for different cut scores (see Table 4). As predicted by Brennan (2001: 48), when λ is equal to the mean $\overline{X}$ over both the populations of pupils and the rating criteria, at $\overline{X}$ (in our study $\overline{X}$ = 0.787) the estimate of *E/T*(λ) achieves its maximum value and the estimate of Φ(λ) its minimum value. All in all, all the estimates of *E/T*(λ) are small, which is an indication that the measurements have substantial precision for the intended use, while every estimate of Φ(λ) is over 0.8, which is a signal that the test contributes a great deal to the dependability of the decision procedure.

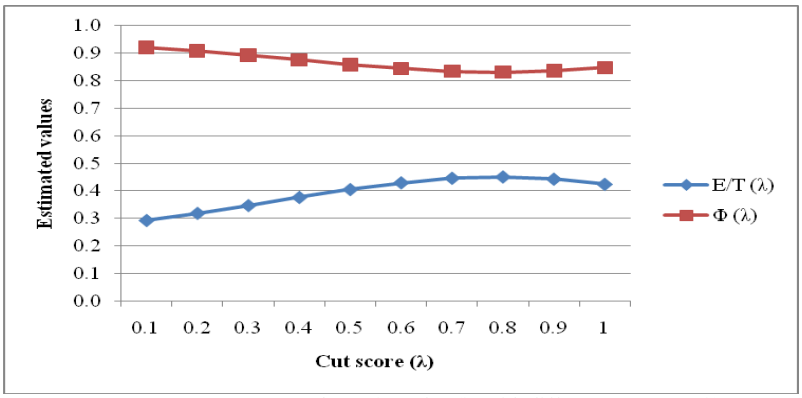

Figure 2. Estimated values of E/T (λ) and Φ (λ) with different cut scores (λ).

V. Conclusions

1. The trainee was familiarised with the CEFR.
2. The trainee could identify the pupils' common reference level and her scores were highly reliable.
3. The total variability of pupils' test scores was influenced by (ordered from the most to the least prominent) (i) the pupil's overall writing competence, (ii) factors such as differences in the pupils' educational histories and other residual characteristics, which have not been directly measured, and (iii) the degree of difficulty of each rating criterion.
4. Given the high estimated Φ coefficient obtained by the trainee, the contribution of the test to the dependability of the decision procedure is very high.
5. The recommended number of rating criteria to be used when assessing the writing competence in this population with this trainee and this population is four.
6. The error involved in using an examinee's observed mean score as an estimate of his or her universe score is extremely low.
7. The writing test is a precise instrument to measure the pupils' writing competence.

The first two conclusions constitute the answers to those general research questions which guided the present study and show that this model can be used to familiarize secondary education teachers with the CEFR, to teach them how to identify pupils' common reference level in written production activities and to obtain reliable scores when rating. Those results and these conclusions support the adoption of this training programme (including its assessment procedure) as a

model for training secondary education teachers of English to identify common reference levels in written production activities.

In contrast to studies such as Wigglesworth (1993) or Lumley & McNamara (1995), this programme has provided an index that estimates the high degree of dependability of the participant's ratings. The present study has also made use of a higher number of confirmatory GT indices ($\hat{\Phi}$, $\hat{\sigma}^2(\Delta)$, *S/N* ratio and $\hat{\sigma}^2(\bar{X})$), compared to those used in other GT studies on performance assessment (Kim, 2009). Finally, another significant point which should be highlighted is that highly satisfactory results can be attained in similar training programmes if the number of rating criteria is reduced to 4, which may contribute to shorten the length of the training programme and the effort demanded of the participants involved. The study, however, possesses the significant limitation of having been focused on a single trainee. Future research should confirm if the application of this model to training programmes with a higher number of diverse participants may also attain such fruitful results.

APPENDIX A GRAMMATICAL RANGE

List of clause structures used to assess grammatical range

	Clause structure	Example
A1		
A2	SV: S V SV: S V ; V = intr. reciprocal SVA: S V A ; V = prepositional SVO: S V O SVO: S V O ; V = phrasal V SVO: S V O $V_{particle}$; V = phrasal V SVOA: S V O_d O_i ; O_i = PP SVO: S V O ; O = *-ing* clause SVO: S V O ; O = *to*-infinitive SVO: S V O ; O = *that*-clause	He went They met They apologized to him He loved her She looked up the number She looked the number up She added the flowers to the bouquet His hair needs combing I wanted to play They thought that he was late
B1	SVOO: S V O_i O_d ; O_d = NP SV: S V ; V = intr. phrasal V SVO: S_{NP} V_{tr} O_d ; O_d = *wh*-infinitive clause SVOA: S_{NP} V_{tr} O_d A; A = *-ing* clause SVOO: S_{NP} V O_d $O_{i\text{ - }PP}$; O_i = PP; Prep = *to* SVOC: S V O (*to V*) C_O SVA: S $V_{prepositional}$ A ; A = V NP ; VP = *-ing* form SVO: S V O A; V = phrasal V SVO: S V O $V_{particle}$ A; V = phrasal V SVOO: S V O_i O_d ; O_i = PP, Prep = *to*; O_d = *that*-clause SVO: S V O ; O = finite *wh*-clause SVA: S V A ; A = PP, Complement = finite *wh*-clause	She asked him his name She gave up He explained how to do it I caught him stealing He gave a big kiss to his mother I found him (to be) a good doctor He wanted the children found They failed in attempting the climb I separated out the three boys from the crowd I separated the three boys out from the crowd They admitted to the authorities that they had entered illegally He asked whether he should come He thought about whether he wanted to go

APPENDIX B DESCRIPTORS OF COHERENCE AND DESCRIPTIVE TEXTS FOR A1 AND A2 LEVELS

	Cohesion	*Descriptive texts*
A1	Can link words or groups of words with very basic linear connectors like *and* and *then*.	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do, etc.
A2	Can link groups of words with simple connectors like *and, but* and *because*.	Can write very short, basic descriptions of events, past activities and personal experiences Can write short simple imaginary biographies and simple poems about people.

APPENDIX C WRITING TEST

Read this notice from the magazine in your school.

Are you a good writer?
Would like to write for the school magazine?
Write about your best friend.
Please, send us your article!

Write about your best friend.
You have 20 minutes to write your answer in the answer sheet.
Write 60-80 words.

REFERENCES

[1] American Educational Research Association; American Psychological Association; National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association.
[2] Andaluc á. (2007). Orden de 10 de agosto de 2007, por la que se desarrolla el curr ículo correspondiente a la Educaci ón Secundaria Obligatoria en Andaluc á [Regulation concerning the curriculum for compulsory secondary education in Andalusia (10 August 2007)]. *Bolet ín Oficial de la Junta de Andaluc á* 171, 23-65.
[3] Bartning, I., M. Martin & I. Vedder. (eds.). Communicative proficiency and linguistic development: Intersections between SLA and language testing research. No place: The European Second Language Association.
[4] Brennan, R. L. (1980). Applications of generalizability theory. In R. A. Berk (ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, Maryland: The John Hopkins University Press, 186-232.
[5] Brennan, R. L. (1992). Elements of generalizability theory. Iowa City, Iowa: American College Testing.
[6] Brennan, R. L. (2001). Generalizability theory. New York: Springer.
[7] Brennan, R. L. & M. Kan. (1977). Signal/noise ratios for domain-referenced tests. *Psychometrika* 42.4, 609-625.
[8] Brown, J. D. & K. M. Bailey. (1984). A categorical instrument for scoring second language writing skills. *Language Learning* 34, 21-42.
[9] Carlsen, C. (2010). Discourse connectives across CEFR-levels: A corpus based study. In Bartning *et al.* (eds.), 191-210.
[10] CEFTrain Project (no date). CEFTrain Project. Helsinki: Helsinki University. http://www.ceftrain.net/ (accessed 20/1/2011).
[11] Cochran, W. G. (1937). The efficiencies of the binomial series tests of significance of a mean and of a correlation coefficient. *Journal of the Royal Statistical Society* 100, 69-73.
[12] Cornfield, J. & J. W. Tukey. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics* 27.4, 907-949.
[13] Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.
[14] Council of Europe. (no date). Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment: Writing tasks: Pilot samples. Strasbourg: Council of Europe. http://www.coe.int/T/DG4/Portfolio/documents/exampleswriting.pdf (accessed 20/1/2011).
[15] Cronbach, L. J., N. Rajaratnam & G. C. Gleser. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology* 16.2, 137-163.
[16] Freeman, R. L. (2004). Telecommunication system engineering. Hoboken, New Jersey: John Wiley and Sons.
[17] Fulcher, G. (2010). Practical language testing. London: Hodder Education.
[18] Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing* 15.2, 100-117.
[19] Glasswell, K. & G. T. L. Brown. (2003). Accuracy in the scoring of writing: Study in large-scale scoring of Asttle writing assessments. Auckland: University of Auckland & Ministry of Education.
[20] Gleser, G. C., L. Cronbach & N. Rajaratnam. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika* 30.4, 395-418.
[21] Hendriks, H. (2008). Presenting the English profile programme: In search of criterial features. *Research Note* 33, 7-10.
[22] Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education* 9.4, 355-379.
[23] Kane, M. T. & R. L. Brennan. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement* 4.1, 105-126.
[24] Kim, Y. H. (2009). A G-theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics* 30.3, 435-440.
[25] Linn, R. L. & N. E. Gronlund. (2000). Measurement and assessment in teaching. Upper Saddle River, New Jersey: Prentice Hall.
[26] Little, D. (2006). The common European framework of reference for languages: Content, purpose, origin, reception and impact. *Language Learning* 39, 167-190.
[27] Llabre, M. M. (1978). An application of generalizability theory to the assessment of writing ability. Ph.D dissertation, University of Florida.
[28] Lumley, T. & T. F. McNamara. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing* 12.1, 54-71.
[29] North, B. (2000). The development of a common framework scale of language proficiency. New York: Peter Lang.
[30] North, B., N. Figueras, S. Takala, P. Van Avermaet & N. Verhelst. (2009). Relating language examinations to the common European framework of reference for languages. Learning, teaching, assessment (CEFR). A manual. Strasbourg: Council of Europe.
[31] North, B. & G. Schneider. (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15.2, 217-262.

[32] Quirk R., S. Greenbaum, G. Leech & J. Svartvik. (1985). A comprehensive grammar of the English language. Harlow, Essex: Longman.
[33] Ríos-Lorenzo, J. L. (2009). Programa de familiarización con el marco común europeo de referencia para las lenguas y las actividades lingüísticas comunicativas de expresión e interacción escrita para profesores de español como lengua extranjera [Programme to familiarise teachers of Spanish as a second language with the common European framework of reference for languages and linguistic communicative activities of written production and interaction]. M.A. dissertation, Pablo de Olavide University.
[34] Salamoura, A. & N. Saville. (2010). Exemplifying the CEFR: Criterial features of written learner English from the English profile programme. In Bartning *et al.* (eds.), 101-132.
[35] Shavelson, R. J. & N. M. Webb. (2005a). Generalizability theory: A primer. Newbury Park, California: Sage.
[36] Shavelson, R. J. & N. M. Webb. (2005b). Generalizability theory. In K. Kempf-Leonard (ed.), *Enciclopedia of social measurement* (vol. 2). Amsterdam: Elsevier, 99-105.
[37] Swartz, C. W. *et al.* (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement* 59.3, 492-506.
[38] Tinsley, H. E. A. & D. J. Weiss. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology* 22.4, 358-376.
[39] Trim, J. L. M. (2010). Breakthrough. Manuscript. http://www.englishprofile.org/images/stories/ep/breakthrough.doc (accessed 20/1/2011).
[40] Triola, M. (2006). Elementary statistics. Boston: Pearson.
[41] University of Cambridge Local Examinations Syndicate. (2007). Cambridge young learners English: Movers reading and writing: Sample paper. Cambridge: University of Cambridge Local Examinations Syndicate. http://www.candidates.cambridgeesol.org/cs/digitalAssets/105810_yle_movers_sample_test.zip (accessed 20/1/2011).
[42] University of Cambridge Local Examinations Syndicate. (2009a). Cambridge young learners English tests: Starters, movers, fliers: Handbook for teachers. Cambridge: University of Cambridge Local Examinations Syndicate. https://www.teachers.cambridgeesol.org/ts/digitalAssets/104521_yle_hb.pdf (accessed 20/1/2011).
[43] University of Cambridge Local Examinations Syndicate. (2009b). Key English test for schools: Handbook for teachers. Cambridge: University of Cambridge Local Examinations Syndicate.
[44] University of Cambridge Local Examinations Syndicate. (2009c). Vocabulary list: Key English test (KET), Key English test for schools (KETfS). Cambridge: University of Cambridge Local Examinations Syndicate. https://www.teachers.cambridgeesol.org/ts/digitalAssets/113295_ket_vocablist09.pdf (accessed 20/1/2011).
[45] Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10.3, 305-319.
[46] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 80-83.

José Cuadrado-Moreno received a B.A. in English Philology from Granada University in 1987, a B.A. in Linguistics from Cádiz University in 2001 and his Ph.D. in English Philology from Granada University in 1996. He is currently an assistant lecturer at Pablo de Olavide University in Seville (Spain) and his research interests include language teaching and assessment.

María Reyes-Fernández received her B.A. in English Philology from Extremadura University in 1988 and her Ph.D. from Seville University, Spain, in 2001. She is currently an assistant lecturer at Pablo de Olavide University in Seville (Spain) and her research interests include applied linguistics, language teaching, assessment and translation.