

# Research Note: PTE General – Using the *Common European Framework of Reference for Languages* to Rate Test Taker Responses

Kirsten Ackermann  
Pearson, London, UK  
Kirsten.ackermann@pearson.com

July 2011

## 1. Introduction

As part of our on-going commitment to maintaining the highest standards in English proficiency testing, Pearson started to revise its suite of General English tests (formerly known as London Tests of English) in 2007. The revised version of the Pearson Test of English General (PTE General) was introduced in the November 2010 session.

Part of this revision process was the introduction of new marking criteria for all constructed responses of the speaking and the writing component of PTE General at all six levels. In order to enhance the alignment to the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001), it was decided to use the CEFR descriptors as a basis to develop the new marking criteria as well as relate the marking scale to the framework.

This paper summarizes the process of using the CEFR in the creation of new marking criteria for constructed responses that require expert marking. It explains the rationale, discusses the advantages and disadvantages of this approach and summarizes the development and validation processes.

### 1.1 Pearson Test of English General

The aim of PTE General is to assess English language proficiency in the four skills (reading, writing, listening, and speaking) for general, practical purposes. The test is offered at six levels (A1 to Level 5) which are designed to correspond to the six Common Reference Levels of the CEFR (A1-C2).

PTE General is intended for teenagers and young adults of all nationalities whose first language is not English. It assesses each language level as practical communicative ability. In practice this means that test takers are assessed according to (1) types of language situation they can deal with, (2) texts, both spoken and written, that they can understand and produce and (3) language functions they can perform and recognize.

The writing component of PTE General consists of three item types: *dictation*, *correspondence* and *write text*. As mentioned above this study is only concerned with the marking of constructed responses and thus only looks at item types *correspondence* and *write text*. *Correspondence* assesses test takers' ability to write a piece of correspondence, e.g., postcard, email, letter in response to a written stimulus, whereas *write text* assesses test takers' ability to write texts from own experience, imagination or knowledge for creative or educational purposes. Table 1 summarizes the objectives of the item types.

**Table 1:** Writing component

Item type	Objective
Correspondence	To assess the ability to write a piece of correspondence.
Write text	To assess the ability to write a text from own experience, knowledge or imagination.

The speaking component consists of a structured face-to-face interaction between a single test taker and an interlocutor. The length of the speaking test varies from five minutes at the lowest levels to eight minutes at the highest. During the speaking test, the interlocutor explains the tasks, asks questions and generally interacts with the test taker. Assessment is done either by a separate examiner who is present or offline from a recording.

The speaking component comprises three sections (A1 and Level 1) or four sections (Levels 2 to 5) designed to elicit a varied sample of language performance. At levels A1 and 1 test takers are required to deliver a monologue about matters of personal interest, comment on one picture and engage in a role play. At levels 2–5 there is one additional section which requires test takers to engage in a discussion about a given topic. Test takers at levels 3–5 receive two pictures to compare and contrast. Table 2 illustrates the speaking component.

**Table 2:** Speaking component

Item type	Objective
Sustained monologue	To assess the ability to talk about matters of personal information, personal interest and experience and opinion.
Discussion <i>Levels 2-5 only</i>	To assess the ability to engage in a discussion on a concrete or abstract issue.
Describe picture	To assess the ability to speak continuously on a topic in response to a visual stimulus.
Role play	To assess the ability to perform language functions by role-playing in a given situation.

### 1.3 Rationale for aligning marking criteria to the CEFR

The Common European Framework of Reference for Languages was published by the Council of Europe in 2001 to provide a common basis for describing aspects of language learning, teaching and assessment. According to the Council of Europe the CEFR offers “a practical tool for setting clear standards to be attained at successive stages of learning and for evaluating outcomes in an internationally comparable manner”<sup>1</sup>. The purpose of referencing the marking criteria to the CEFR is to contribute to the alignment of PTE General to a widely accepted framework whose levels of language proficiency are understood and to a large extent described.

<sup>1</sup> Council of Europe. Education and Languages: Education Policy. Retrieved 20 January 2011 from [http://www.coe.int/T/DG4/Linguistic/CADRE\\_EN.asp](http://www.coe.int/T/DG4/Linguistic/CADRE_EN.asp)

The CEFR contains common reference levels and illustrative scales of descriptors (hereafter CEFR descriptors) which do not only describe qualitative aspects of language, e.g., degree of accuracy or fluency, but also what learners need to do with the language, e.g. ability to give a monologue or write creatively. They, thus, provide a basis for assessing qualitative aspects of test takers’ language use as well as their ability to use different language functions, which is also the aim of the speaking and writing components of PTE General using the CEFR as it serves the communicative approach to language testing.

## 2. Development of the Marking Criteria

The development of new marking criteria started in January 2009 with the decision to use the CEFR as the starting point. There were five main stages of the development process: (1) gathering expert judgments on the usability of the CEFR descriptors for the creation of marking criteria, (2) identifying gaps and inconsistency in the CEFR, (3) producing additional guidelines to flesh out the descriptors, (4) field testing the new marking criteria and (5) using quantitative and qualitative methods for construct validation. Outcomes of this development include new marking criteria supported by test-specific commentaries and a marking scale based on the CEFR.

During the initial stage a panel consisting of eight experienced external examiners gave recommendations on which CEFR descriptors could serve as marking criteria. The panellists were asked (1) to reference the CEFR to the item types of the writing and speaking components of PTE General; (2) to select potential CEFR descriptors for both test components; and (3) to try out different marking procedures for the speaking component in order to evaluate the usability of the considered CEFR descriptors.

As the end of this initial stage the following consensus was reached: It was agreed that each task of the speaking and writing test should be assessed on the performance of the specific language function tested using one individual trait, e.g., *overall written interaction* in the *correspondence* task or *thematic development* in the *describe picture* task. Qualitative traits, e.g., *accuracy*, *range* were to be assessed overall. By looking at what a test taker can do with the language and how well s/he can use the language a fair assessment of test taker performances is ensured. Tables 3 and 4 give an overview of the CEFR descriptors to be adapted as marking criteria for the speaking and writing component of PTE General.

**Table 3:** CEFR descriptors speaking component

Section in PTE General	Item type	Individual trait
10	Monologue	Sustaining Monologue
11 (L2-L5 only)	Discussion	Turn Taking
12	Picture description	Thematic Development
13	Role play	Sociolinguistic Appropriateness
<i>Overall qualitative traits</i>	Fluency                      Interaction	Phonological Control    Range    Accuracy

**Table 4:** CEFR descriptors writing component

Section in PTE General	Item type	Individual traits	Qualitative Traits
8	Correspondence	Overall Written Interaction	Range Accuracy Coherence Orthographic Control
9	Write text	Overall Written Production	Range Accuracy Coherence Orthographic Control

This first stage as well as the subsequent trialling of the marking criteria, however, revealed the limitations of adapting the CEFR descriptors. The following challenges in working with the framework as pointed out by Alderson, Figueras, Kuijper, Nold, Takala, & Tardieu (2006, p.9) shaped further discussions: (1) inconsistencies, (2) terminology issues, (3) lack of definition, and (4) gaps within the framework.

The following steps were taken in order to address these challenges. If CEFR descriptors were unavailable for a specific level, marking criteria were based on descriptors of alternative scales, e.g. ELTDU *Stages of Attainment Scale 1976*. Furthermore, commentaries were devised for each marking criterion at each level specifying how the CEFR descriptors are to be interpreted in the test context in order to operationalize the marking criteria. Test specifications, item writer guidelines and feedback from examiners who were trialling the new marking criteria informed the development of the commentaries. These commentaries serve to resolve ambiguities in terminology, enhance consistency and provide definitions in the test-specific context, i.e. to explain, clarify and if appropriate exemplify the individual and qualitative traits in relation to the six levels of PTE General. Table 5 provides an excerpt from commentaries for the speaking and writing tests Level 2 (designed to be aligned with CEFR level B1).

**Table 5:** Sample of commentaries for marking speaking and writing at Level 2

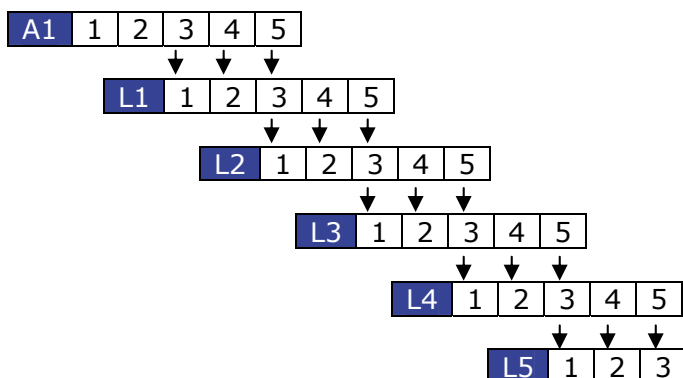
Item trait	CEFR descriptor (B1)	Commentary (L2)
Sociolinguistic Appropriateness	Can perform and respond to a wide range of language functions, using their most common exponents in a neutral register. Is aware of the salient politeness conventions and acts appropriately. Is aware of, and looks out for signs of, the most significant differences between the customs, usages, attitudes, values and beliefs prevalent in the community concerned and those of his or her own.	Test takers may be required to perform the following functions and respond to them: requesting, offering, suggesting, thanking, rejecting, apologising or congratulating. While test takers' language will generally be limited to a neutral register, some awareness of appropriateness (e.g., in terms of degrees of formality) is expected.
Fluency	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Test taker's delivery should not become strenuous for the assessor. If pauses dominate the conversation, the test taker should be penalized.
Overall Written production	Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements into a linear sequence.	Test takers can give their opinions, reasons, or accounts of experiences, describing feelings, thoughts and reactions in simple connected text. Test takers should have a repertoire of linking devices including <i>because, so, firstly, secondly, thirdly, afterwards, lastly</i> etc.
Range	Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interests, work, travel and current events.	This includes describing experiences and events, dreams, hopes & ambitions and briefly giving reasons and explanations for opinions and plans.

A five-point marking scale<sup>2</sup> was developed to measure the extent to which test takers fulfil each marking criterion as shown in Table 6.

**Table 6:** Marking scale

Score	Comment
1	The response does not meet the level requirements. The test taker’s English language proficiency is below the level.
2	The response meets some of the level requirements. The test taker’s English language proficiency is just at the level.
3	The response meets all level requirements. The test taker’s English language proficiency is at the level.
4	The response exceeds some of the level requirements/meets some of the requirements of the level above. The test taker’s English language proficiency is almost at the next level.
5	The response fully exceeds the level requirements/meet the requirements of the level above. The test taker’s English language proficiency is at the next level.

As the levels are a continuum the relations between the scores at adjacent levels are illustrated in Figure 1 below. For example, if a test taker achieves a score of 5 at level L1, s/he fulfils all the requirements for level L2. However, as a score of 5 represents the upper limit of a test’s measurement capacity, it cannot be reliably inferred that the test taker is at level L2; s/he might be even higher. Similarly, if a test taker scores 1 at level L1 this indicates that his or her level is no higher than level A1 (but could be even lower).



**Figure 1:** Overlapping marking scale for PTE General writing and speaking

The marking criteria, commentaries and marking scale were then field-tested between December 2009 and March 2010 as part of the validation process. The feedback provided by examiners helped further refine the commentaries. The quantitative and qualitative results from the validation process are summarized in section 3.

<sup>2</sup> Level 5 has a three-point scale as it is the highest level and cannot be referenced against a level above.

### 3. Validation Process

In order to validate the performance of the marking criteria and marking scale, data collected during marking oral and written responses<sup>3</sup> were analyzed quantitatively and qualitatively. Five oral examiners and five written examiners took part in this validation study. Examiners had on average 16.1 years (SD=11.96) of experience in teaching English as a foreign language, extensive experience as examiners and moderate to good knowledge of the CEFR. The scores examiners awarded were analyzed quantitatively. Examiners were also asked to provide think alouds while marking individual written and oral test taker responses.

The standardization process for both groups, oral and written examiners, followed the same methodology. Examiners underwent a CEFR re-familiarization exercise followed by a detailed introduction of the new marking criteria, commentaries as well as the marking scale. Examiners then marked test taker responses as a group. During this marking session examiners were given the opportunity to ask questions and discuss marking issues. Once examiners felt confident that they were applying the marking criteria appropriately, examiners started marking independently.

#### 3.1 Quantitative analysis

Multi-faceted Rasch analysis of the data was conducted providing estimates of test taker ability, rater severity and item trait difficulty by mapping these facets on a common log-linear logit scale. The findings of the analyses for levels L2 speaking and L3 writing (*Correspondence*) should serve here as an example of this validation process.

##### 3.1.1 Results from the quantitative analysis of the speaking test

The five oral examiners assessed 21 L2 test taker responses. Figure 2 presents graphically the measures for test taker ability, rater severity and item trait difficulty mapped on a common scale. Examinees are ordered with the most able test taker at the top and the least able at the bottom. In terms of examiners, the most severe rater is the one at the top of the figure. Likewise, the most difficult trait is at the top and the least difficult at the bottom.

---

<sup>3</sup> Examiners marked test taker responses that were collected as part of the PTE General Concordance study, which has two objectives: (1) Equating PTE General test scores with other high-stakes English language tests and (2) Calibrating test scores to be put on a common Pearson Scale of English. Both objectives provide direct evidence of the comparability of test results.

Measr	-Examiners	+examinees	-Traits				Scale
3							(5)
		28823558	28920401	28923683			4
2		28317361					
		27869288					
		28922413					---
1	E14						
		28423521			AC		
	E10	28929273					
	E15	28239189					
	E13	28760860			PC SM TD TT		3
0					SA		*
		28423521					
		28154932			FL IN		
		28963706					
	E7	28960098			RA		
-1		28873415	28916068				
		28277175					
		28798668					---
		28239339					
-2							
		27970963					
-3							2
		28874333					
-4							(1)
Measr	-Examiners	+examinees	-Traits				Scale

Figure 2: Facets<sup>4</sup> analysis: Speaking test level L2<sup>5</sup>

<sup>4</sup> Facets refer here to examiner, examinees and marking criteria (=traits).

<sup>5</sup> AC=Accuracy, PC=Phonological Control, SM=Sustained Monologue, TD=Thematic Development, TT=Turn Taking, SA=Sociolinguistic Appropriateness, FL=Fluency, IN=Interaction, RA=Range

Figure 2 indicates that there is a spread of test takers' language proficiency across the level with six examinees showing a higher ability and two a lower ability than required to be considered at the level. The range of examinee measures is about 6.33 logits<sup>6</sup> [2.56 -(-3.77)] with five statistically distinguishable levels of performance in this sample of examinees with one examiner. This shows that the five-point marking scale is sufficient to distinguish between different proficiency levels and that examiners use the scale appropriately.

In terms of rater severity, examiner E14 is most severe in her marking and E7 more lenient than the other examiners. The examiner measurement report shows a reliability of .94 indicating that the analysis is reliably separating examiners into different levels of severity. However, in this case, a low reliability is desirable, since ideally examiners would be equally severe. The logit spread of the examiners is about 1.7 logits, which corresponds to one statistically distinguishable level of performance for the examinees or almost one category (= one score point) on the rating scale. Therefore, failing to adjust for examiner severity could result in a higher performer with a severe examiner and a lower performer with a lenient examiner, whose performances are actually two score points different, being reported in the same category.

In order to analyze inter-rater reliability, the Intra-class Correlation coefficient (ICC) was calculated using SPSS. ICC measures the ratio of between-groups variance to total variance. The coefficient ranges from 0 to 1, and will be close to 1.0, when there is little variation among the scores given to each response by the examiners indicating high inter-rater reliability.

**Table 7:** Intraclass Correlation Coefficients PTE General speaking test

Item trait	Intraclass Correlation Coefficient
Sustained Monologue	0.77
Turn Taking	0.84
Thematic Development	0.90
Sociolinguistic Appropriateness	0.84
Fluency	0.88
Interaction	0.83
Range	0.87
Accuracy	0.87
Phonological Control	0.89

Table 7 shows strong agreement amongst examiners when assessing all item traits. Inter-rater reliability is generally very high and was highest for *thematic development* and *phonological control* and lowest for *sustained monologue*. This can partially be explained by the findings of the verbal protocols analysis and post-marking questionnaires. Examiners stated that it was more difficult to assess *sustained monologue* as some test takers provided too small a language sample to make a reliable assessment. On the other hand, examiners pointed out that *phonological control* was a clear marking criterion that the majority of examiners found the least difficult to assess.

With regard to test taker performance on the item traits, Figure 2 indicates that it is more difficult for test takers to obtain high scores for Accuracy than it is for Range. This indicates that either test takers are less able to use grammar and vocabulary accurately or that examiners are more severe when assessing *accuracy*. The verbal protocol analysis supports the latter assumption as examiners concentrate on listing inaccuracies when assessing test taker performances.

<sup>6</sup> Logit is the unit of measure used by Rasch analysis.

Figure 3 shows the probability of occurrences for each category (= one score point). The pattern of the curves indicates that each category is in turn the most likely category at successive points along the latent variable (=ability). In other words, the higher the test taker's ability the higher the awarded score will be indicating the reliability of the marking scale.

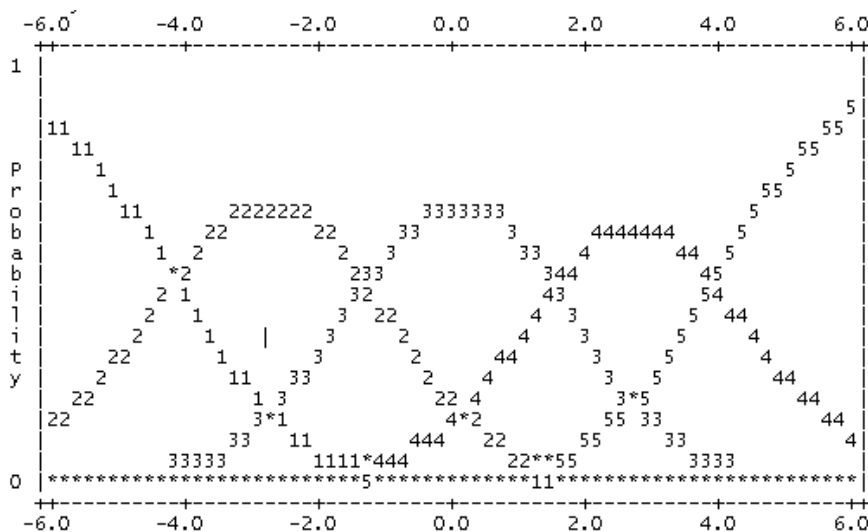


Figure 3: Probability curves level L1 speaking test

### 3.1.2 Results from the quantitative analysis of the writing test

Five examiners assessed 24 written test taker responses at level L3. Figure 4 shows that the marking criteria and the marking scale sufficiently discriminate between test taker performances on the writing test. Four examiners are clustered around 1 indicating that they show a greater degree of severity than expected. The reliability index of .59 shows that the analysis is less reliably separating examiners into different levels of severity, i.e. the examiners show a similar degree of severity. As mentioned above, a low reliability is desirable, since ideally examiners would be equally severe. The logit spread of the examiners is about 0.6 logits, i.e. examinees are very likely to receive the same score from all five examiners.

This finding of high inter-rater reliability is also reflected in the high ICC as shown in Table 8.

Table 8: Intraclass Correlation Coefficients PTE General writing test

Item trait	Intraclass Correlation Coefficient
Overall Written Interaction	0.84
Range	0.87
Accuracy	0.87
Coherence & Cohesion	0.82
Orthographic Control	0.84

Table 8 shows strong agreement amongst examiners when assessing all item traits of the *correspondence* task. Inter-rater reliability is generally very high and highest when assessing *range* and *accuracy*. Based on the post-marking questionnaire, higher agreement was expected when assessing *orthographic control* as most examiners felt that the marking criterion for this trait was straightforward and easy to apply.

Measr	-Examiners	+examinees	-Traits	Scale
3	+	+	+	(5)
2	+	+ 28883129	+	---
		28836808		
	E14	28208873	28940681	
	E15 E9	28517718		
	E8			
1	+	+ 28164174	28768868	+
		28902808	AC	
	E12	28184521	28579793	
		28873680	29129904	3
*	0	*	* OWI	*
		28238817		
		28546557	28963775	CC OC RA
		28239189		
		28904975		
		28260398		
-1	+	+ 28152481	28885868 29203417	+
				---
-2	+	+ 28787850	+	+
		28564172		
-3	+	+	+	2
		28239339		
-4	+	+	+	(1)

Figure 4: Facets analysis: Writing test level L3<sup>7</sup>

<sup>7</sup> AC=Accuracy, OWI=Overall Written Interaction, CC=Coherence and Cohesion, OC=Orthographic Control, RA=Range

Figure 4 indicates again that it is slightly more difficult for test takers to obtain high scores for *accuracy* than it is, for example, for *range*.

The examinee measurement report shows that the analysis is separating test takers into different levels of language proficiency with a reliability of .92. The range of examinee measures is about 5.77 logits with four statistically distinguishable levels of performance in this sample of examinees with one examiner. This can be explained by the fact that the examiners hardly ever used a score of four as it becomes apparent in Figure 5.

Figure 5 shows that test takers scores increase with increasing ability. The five scale categories on the PTE General marking scale are appropriately ordered and the scale is functioning properly as a five-point scale. However, it also becomes apparent that examiners were less likely to award a score of four. This is most likely due to the small cohort of test takers (n=24) as the statistical program cannot estimate the probability of obtaining a score of four reliably given this number of cases. This assumption is confirmed by the analysis of the verbal protocol, which reveals that examiners are using the scale to a large extent correctly. Nonetheless, future examiner training must ensure that examiners feel confident in using all five scale points.

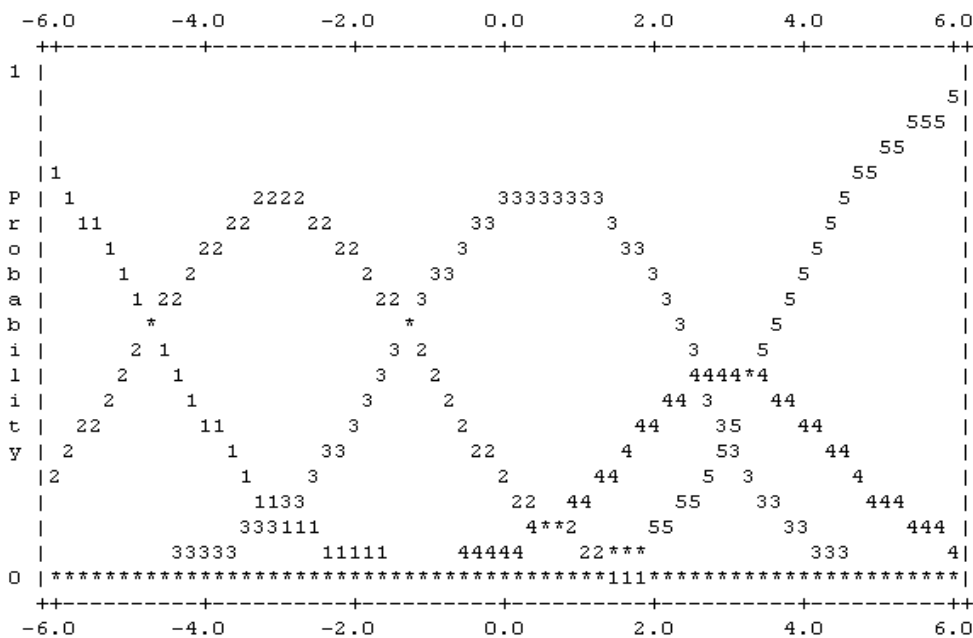


Figure 5: Probability curves L3 writing test

### 3.1.3 Summary

Multi-faceted Rasch analysis showed that examiners are able to use the new marking criteria and marking scale appropriately, although with varying degree of severity. Overall, the inter-rater reliability was high in both speaking and writing. The qualitative analysis below provides additional insight into the application of the marking criteria.

### 3.2 Qualitative analysis

In order to qualitatively validate the new marking criteria, the examiners’ cognitive strategies whilst marking and their marking decisions were analyzed using verbal protocols. Verbal protocols are a well-established method of qualitative data collection used in various fields of research, e.g. in education and assessment. Ericsson and Simon (1993) regard verbal protocols as a source of evidence about human cognition and according to Kuusela and Paul (2000) it is the most widely used thought process-tracing method today.

This paper will only report the findings from the verbal protocol analyses that correspond to the same set of test taker responses analyzed quantitatively. All examiners were given the same oral and written responses at Level 2 and 3 respectively. Each examiner was asked to produce one verbal protocols. Examiners assessing the written response were asked to provide non-mediated concurrent protocols, i.e. they would mark the *correspondence* task while concurrently thinking aloud.

Although the concurrent method is widely believed to be more effective in revealing thought processing than the retrospective method (Greatorex & Sütő, 2008), oral examiners were asked to provide non-mediated retrospective verbal protocols to accommodate the difficulty of having to listen to a recording and speak at the same time. In other words, oral examiners would listen to one section of a test taker's speaking test, e.g. *discussion*, pause the recording and give a think aloud providing information about their marking of this sections retrospectively, before continuing with the next section.

All verbal protocols were recorded, transcribed and analyzed qualitatively using MAXQDA 2007/2010. In the development of a valid coding scheme Green's approach (1998) was adopted. Initially four protocols were examined with respect to the type of information the respondents mentioned. The established codes were as specific as possible but also generalizable across all protocols. The developed coding scheme allowed the analyses of (1) individual marking sequences, (2) interpretation of the marking criteria, (3) appropriate use the rating scale and (4) any difficulties in rating. The codes were assigned to each identified segment of each verbal protocol.

### 3.2.1 Results for the qualitative analysis of the speaking test

The analysis of the verbal protocols confirmed the quantitative findings that examiners use the new marking criteria largely correctly and apply the marking scale appropriately, but also revealed issues regarding the marking process. Examiners refer regularly to the marking criteria when assessing each item trait. E15, for example, cites from the marking criteria:

So, for turn taking, er I think I'm going to give this one a three, because even though.... it [marking criterion] says 'can initiate, maintain and close simple face to face conversation', I definitely think that this candidate maintains the conversation, erm, and does close it at the end, maybe rather abruptly by just saying 'yes that's a good point', but I still think that that is what happens, erm, erm, yes, I definitely don't think it's below that or at the lower end of the band, I think it's just right at level, so, I'm going to give it a three. (SL2E15)

In addition, E15 shows a very stable marking strategy, by which she first consults the marking criterion at the level, assigns a preliminary score, confirms the score or re-scores by consulting the marking criteria of the level above and/or below:

Okay. Right, so for this one, thematic development, erm, can reasonably, fluently relate straightforward narrative, yes, I think so, erm, now I'm looking at the B2 [L3] 'develop a clear description or narrative extending', definitely I don't think she expands really, or supports this sort of argument, erm, A2 [L1], 'can tell a story or describe something in the simplest of points', she does more than that, erm, I think I'll just give this one a three, again, because she kind of, she gives it, yes, a straightforward narrative, she does, and a straightforward description, hmm, but I don't think it's much more than that really, so, I'm going to give that a three. (SL2E15)

E14 uses the same pattern but less consistently. E14 is the only examiner who also refers to the commentaries to assist her marking:

Erm, she didn't keep talking for the minimum length of time that she should have done, which would have been about forty to fifty seconds according to the, er, language descriptors, erm, so, she didn't really talk at length, as a monologue, erm, about the topic and therefore I felt that she could only get a two as a borderline rather than a three. Okay. (SL2E14)

Examiners occasionally cite the test taker when justifying the awarded mark:

I was surprised by some of the fixed phrases that she actually had picked up, things like, erm, 'at first sight', and I think she said 'on second thoughts', which is nice expression at probably, even at B2 [L3], erm, however she didn't have an awful lot of er, range of vocabulary when it came to the, er, monologue or even...yes, particularly the monologue, no, the, the, er, picture description. Erm, the grammatical range was also not wide, erm, and these were the thoughts that led me to award a two. (SL2E14)

E7 uses a different marking strategy when assessing the qualitative item traits. The examiner gradually assesses these item traits in order to come to a final mark once the speaking test is completed:

(A)t the same time I'm thinking of the mark I should give for the sustained monologue, and at the same time I'm thinking about the sort of first idea to pencil in about fluency. So, I would have written in pencil sustained monologue – three, and fluency – three, but I'm poised to change that fluency mark up or down as the test goes on. (SL2E7)

E7 continues assessing Fluency:

(...) and I'm still... I check again, yes I gave.... I pencilled in a three for fluency overall and I think that's quite fair, so I'll leave that as it is, and then wait for the role play to see what else will be clarified. (ibid.)

Once the speaking test is completed the examiner awards a mark of three for Fluency.

The analysis of the verbal protocols, however, also revealed that (1) not all examiners base each scoring decision on the relevant marking criterion; (2) examiners differ in their interpretation of the marking criteria; (3) examiners face various extrinsic difficulties when applying the marking criteria; and (4) examiners make marking mistakes.

Regarding observation (1) the verbal protocols showed that especially E10 relied on his intuition and experience as an examiner when marking the speaking test. He makes no explicit reference to the new marking criteria or cites them to justify his marking decision. This can be expected to produce a 'halo effect' which occurs when a rater awards the same scores for multiple item traits, based on his/her overall impression, for all item traits. In other words, the rater does not use the analytic scales in an analytical manner. E10 states:

Just, just that it is difficult to, it is difficult to erm... to mark, erm, beyond your very, very narrow range when the candidate isn't really that forthcoming, when the candidate is shy, clearly quiet and shy, and not really giving anything other than the minimum of what's needed. However, I don't feel that I can mark her down for that, it's just a personality thing, it's just the way she's reacting in an exam situation. So, erm... but it does make, it does make it difficult, 'cause in another situation she may well actually have erm, have better language than, than just straight down threes but that's all I can, that's all I can mark her on. (SL2E10)

This observation is matched by the examiner report from the multi-faceted Rasch analysis, which showed that E10 is much more predictable (noticeably lower mean-square fit statistics), i.e., this examiner avoids the extreme scores (1, 5).

Regarding observation (2) the verbal protocols revealed that examiners give varying importance to or choose to ignore different aspects of individual marking criteria. This behaviour allows examiners to come to a scoring decision, but puts intra- as well as inter-rater reliability at risk. For example, the marking criterion *Range* for level L2 reads: *[The test taker] has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and*

*circumlocutions on topics such as family, hobbies and interests, work, travel and current events.* E14 gives the following retrospective verbal protocol on assessing *Range*:

On the range, (...), I was surprised by some of the fixed phrases that she actually had picked up, things like, erm, 'at first sight', and I think she said 'on second thoughts', which is nice expression at probably, even at B2, erm, however she didn't have an awful lot of er, range of vocabulary when it came to the, er, monologue or even...yes, particularly the monologue, no, the, the, er, picture description. Erm, the grammatical range was also not wide, erm, and these were the thoughts that led me to award a two. (SL2E14)

Even though the marking criterion states that a L2 test taker only needs to show *sufficient vocabulary to express him/herself with some hesitation and circumlocutions* and E14 gives examples of good expressions in the test taker's response, E14 bases her scoring decision on her interpretation of 'wide range'. E15, on the other hand, states:

And, range, 'has enough language to get by with sufficient vocabulary, some hesitation and circumlocutions', the B2 [L3] one, I think that it's not, definitely not completely a B2 [L3] because I feel like she didn't have a high enough range all the time, because it seemed like she was limited in what she wanted to say sometimes, but did have a fairly wide range vocabulary, so I'm going to give that a four. (SL2E15)

This examiner acknowledges that the test taker's range is limited in some places, but decides in accordance with the marking criterion that the test taker should not be penalized.

Regarding observation (3) the verbal protocols show that the state of mind the examiner is in, the test taker's behaviour or the behaviour of the interlocutor may be detrimental to the marking process. For example, E13 remarks half-way through the verbal protocol "I'm quite tired now, so it's quite ... I have to really ... trying quite hard to concentrate" (SL2E13). E10 reports on his struggle to assess the test taker's performance as he takes the test taker's personality into consideration:

(I)t's a little bit difficult to assess her because she's not terribly forthcoming (...) is difficult to, it is difficult to erm... to mark, erm, beyond your very, very narrow range when the candidate isn't really that forthcoming, when the candidate is shy, clearly quiet and shy, and not really giving anything other than the minimum of what's needed. **However, I don't feel that I can mark her down for that, it's just a personality thing, it's just the way she's reacting in an exam situation.** So, erm but it does make, it does make it difficult, 'cause in another situation she may well actually have erm, have better language than, than just straight down threes but that's all I can, that's all I can mark her on. (SL2E10, emphasis added)

E7 comments on how the interlocutor's behaviour, i.e. appropriate time keeping, made it more difficult to assess the test taker in a fair way. The examiner criticizes the fact that "the candidates are not given the time allowances that they're supposed to and sometimes you think, ooh that was very weak but when you check it's because they, you know, they haven't been given a full whack of it." (SL2E7)

Regarding observation (4) the verbal protocols reveal that examiners occasionally make mistakes while applying the new marking criteria. E13, for example, seems to confuse *Range* with *Fluency* when she states: "I've started to think about her fluency as well, erm, which I think would be a three because she's expressing what she wants to say even though it's at quite a basic level" (SL2E13). Unfortunately, on this occasion E13 does not consult the marking criteria which could have made her realize the mistake.

### 3.2.2 Results for the qualitative analysis of the writing test

The analysis of the verbal protocols shows that examiners use various strategies to help them with the assessment. For example, E15 exhibits a very structured approach to the marking process. She starts by re-reading the prompt. As the examiner is already familiar with the stimulus article, she takes notes of the task requirements to assist her throughout the marking process:

First thing is to remind myself of the prompt. So this is reading an article about children's worries. I've already read this article, so fine. Write a response to the article. So it is a response, I'm going to write this down [E15 takes notes], just cause it helps me, just a quick reference, opinion about things young people worry about, reaction to writer's points, reasons and examples to support your opinion. (WL3E15)

Examiners use the new marking criteria largely correctly and apply the marking scale appropriately as, for example illustrated when examiner E12 assess *overall written interaction*:

Overall Written Interaction [E12 quotes from marking criterion] erm, okay, clarity, precision, relating to the addressee flexibly and effectively. My view is it lacks clarity because of some of the other language problems involved. And I think there is a degree of ambiguity within the response. I think it has shades of a response at this level, but I don't think it is quite hiding the mark. So I would award a 2 for OWI. (WL3E12)

The verbal protocols were also analyzed with regard to any difficulties examiners experienced during marking. In general, examiners who marked the written responses rarely relied on their intuition, experienced fewer difficulties and made fewer marking mistakes than the oral examiners.

One area that caused difficulty and led to marking mistakes is related to weighing components of marking criteria differently. Even though the response does not fully meet the requirements relating to Orthographic Control, E11 decides to award a score of three and justifies the decision by placing varying emphasis on the different components of the marking criterion:

Orthographic control: There are no paragraphs. Punctuation, let's have a look [E11 reads response]. Punctuation seems ok. And the spelling seems fine, apart from some words 'bombardious', but generally, well the odd slip. It's got no paragraphs. So we're going back to the question, how important if it's got no paragraphs. I would tend to think that's just one out of three criteria, and give it a 3, because it fulfils the other criteria. So I give it a 3. (WL3E11)

E8 also allows one strong component (lexical range) of the marking criterion Range to compensate for a weaker one (grammatical range), but additionally the examiner overrates the strong component awarding a score that exceeds the level requirement:

I don't have the feeling that this person is restricted herself at all. I don't think the range of structure is as wide as the range of vocabulary. However, what they have used is perfectly effective. It doesn't strike me that they just don't know any more grammar, they just use vocabulary to get there point across. So does that mean this is C1 level [E8 re-reads descriptor]. I think it does – I got a few issues about coherence with the length of the sentence. But I think the range of what they got is at the next level. I'll give it a 5. (WL3E8)

However, in the course of the marking E8 realises this mistake and corrects her marking decision:

Do you know what, I'm going back to Range and mark it down to 4, because I'll think it is a case they got lots of vocabulary but they don't know how to necessarily put it into the right sentence. So I think range is more of a 4 than 5. (ibid.)

For E9, the fact that the test taker uses run-on sentences determines the assessment process throughout. The examiner finds it difficult to decide which marking criterion to consider when penalizing the overuse of run-on sentences. Initially, E9 states that run-on sentences cause problems with coherence:

I know what's not right with this, no, it's not 'effectively in writing and relate to those of others', no, because what you've got, is a series of unrelated ideas, problems with coherence (...) okay, that's er, what that should come under is cohesion because it's a run-through sentence, that's it. (WL3E9)

In the course of the assessment, the examiner hesitates whether to penalize run-on sentences assessing Accuracy or Coherence and Cohesion:

[E9 quotes from response] 'We have the responsibility to give them the tools to grow up with the feeling that they can make a change.' So once again that's a run-through sentence. So we've got run-through sentences, it's actually not right, the run-through sentences, so do you count those as grammatical or errors of, erm, cohesion? I think you should count those as basically it's the clause structure, sentence clause structure I think that would come under, yes that would come under (...) accuracy. (ibid.)

E9 then goes on to penalize the feature under both, *accuracy or coherence and cohesion*:

Accuracy, erm, not happy with the accuracy, er, two because of the run through sentences (...). So, coherence, 'can use a variety of linking words efficiently to mark clearly the relationships between ideas, can use a limited number of cohesive devices to link his or her...into a clear coherent discourse', no, it's not a clear coherent discourse 'though there may be some jumpiness in a long contribution', there is definite jumpiness in long contributions. (ibid.)

### 3.2.3 Examiner feedback regarding 'Thinking aloud'

The paper-based post-marking questionnaire revealed that examiners found marking while thinking aloud as difficult as or more difficult than regular marking. The examiners who found it more difficult stated that thinking aloud while marking got easier in the process.

Examiners of the written responses experienced thinking aloud for *overall written interaction* and *coherence and cohesion* as most difficult and for *orthographic control* and *accuracy* as least difficult. Oral examiners experienced thinking aloud for *sustained monologue* as most difficult and for *phonological control* as least difficult. Examiners stated that clear evidence from the test taker for a specific item trait and straightforward marking criteria made thinking aloud while marking easier.

When asked if thinking aloud altered the way examiners thought while marking, E13 comments: "It increases concentration and makes you more aware of detail." E9 shares this opinion: "Thinking aloud can increase awareness of how one is applying the criteria" and adds "I believe it will give an insight into how markers justify marking decisions". E10 suggests that thinking aloud while marking "clarifies what you think – makes one feel more confirmed in one's opinion." E15 believes "(i)t's useful because you can design marking materials which give clearer guidance to the marker".

In general, oral examiners who produced retrospective verbal protocols after each section of the speaking test found giving think alouds more difficult than the examiner giving concurrent verbal protocols. E14 states: "By speaking in between the sections it was difficult to remember the test elements as a whole for awarding the 'overall' marks for fluency, interaction, accuracy and range. The longer my 'thinking aloud' response, the more difficult it was.

E15 states: "(...) I found it hard to remember what the candidate had said while speaking about what mark I was going to give." E8 even warns that "there is a danger that the thinking aloud distracts form the actual marks".

### 3.2.4 Summary

The analysis of the verbal protocols showed that to a large extent both groups of examiners use the marking criteria and marking scale correctly. Examiners are, however, faced with the difficulty of interpreting the marking criteria correctly and consistently when these do not match all features of individual performances. In addition, as most examiners rely on their experience and intuition to some degree, they need to reconcile their first impression or intuitive assessment with the marking criteria and scale.

In this regard, the findings from the verbal protocol analysis provide a valuable source for improving examiner standardization materials and training as they highlight the difficulties examiners encounter while using a particular marking scheme. To prevent examiners from using coping strategies such as resorting to intuition, relying on previous marking experience or overrating certain components of marking criteria while disregarding others, they must be trained in utilising marking materials correctly and adopting marking strategies such as 'best-fit' and consulting the marking criteria of the level above and below in addition.

## 4. Conclusion

This paper described the development process of the new marking criteria for the speaking and writing component of PTE General. The CEFR descriptors served as the starting point for the creation of new analytic marking schemes. The approach adopted included identifying CEFR scales that relate to the types of language functions test takers have to perform when engaging in various language situations during PTE General examinations and the quality of the language produced. Accordingly, descriptors that illustrate language functions (individual item trait) as well as descriptors representing qualitative aspect of language use (qualitative item traits) were selected as marking criteria for the speaking and writing components.

In cases where the CEFR did not provide descriptors alternatives were introduced. If parts of the CEFR descriptors were not relevant to the test specifications, they were abbreviated and in rare cases edited. To support the consistent interpretation of the CEFR-derived marking criteria commentaries for each criterion were created. These commentaries serve to resolve ambiguities in terminology, enhance consistency and provide definitions in the test-specific context, i.e. to explain, clarify and if appropriate exemplify the individual and qualitative item traits in relation to the six levels of PTE General.

The study then investigated the usability of the marking criteria and marking scale and illustrated how qualitative and quantitative analyses can complement each other. The quantitative findings from the multi-faceted Rasch analysis showed that the marking scale discriminates different degrees of test takers' language proficiency successfully. It also showed that the item traits/marketing criteria are of comparable difficulty. The analysis further revealed that examiners exhibit different levels of severity. However, this is recognized as unavoidable in the literature (see e.g., Lunz and Stahl 1990, McIntyre 1993, Lumley and McNamara 1995), but that efficient standardization training can succeed in yielding consistent marking, i.e. high intra-rater reliability.

The results from the verbal protocol analysis provided valuable insights into the marking strategies of examiners, ascertained whether examiners use the marking criteria and marking scale correctly and consistently, and identified difficulties in the application of marking criteria.

Verbal protocols are therefore a useful source of information and are recommended to be used as part of the process of construct validation. Both, quantitative and qualitative findings helped improve training materials before large-scale examiner standardization took place.

## References

- Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly*, 3 (1), 3 - 30.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Council of Europe (2003). *Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Preliminary Pilot Version*. Strasbourg: Council of Europe.
- Ericsson, K.A. & Simon, H.A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.
- Greatorex, J. & Sütő, I. (2008). 'What do GCSE examiners think of 'thinking aloud'? Findings from an exploratory study', *Educational Research*, 50: 4, 319-331.
- Green, Alison. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: CUP.
- Kuusela, H. & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology* 113(3), 387-404.
- Lumley, T. & McNamara, T.F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing* 12 (1), 54-71.
- Lunz, M.E. and Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions* 13, 425-44.
- McIntyre, P.N. (1993): The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples. MA thesis, University of Melbourne.