Research Papers in Language Teaching and Learning
Vol. 3, No. 1, February 2012, 78-91
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

'Co-construction' in the B2 and C1 KPG oral exams: a comparison of examiners as a factor involved in candidates' performance

['Συν-οικοδόμηση' στις προφορικές εξετάσεις ΚΠΓ, Β2 και Γ1 επιπέδου: Σύγκριση των εξεταστών ως παράγοντα που εμπλέκεται στην απόδοση του υποψηφίου]

Xenia Delieza

Researchers who investigate oral testing invariably allude to the complexity of the procedure residing in the multitude of factors which influence its final outcome. One of these factors is the examiner who has been found to affect test takers' performance through his/her role both as interlocutor and rater. There has also been a long discussion on the characteristics of this role in the so-called oral paired-exam in comparison to the oral proficiency interview. The present paper looks into two oral tests of the same examination battery, both of which are paired in that two candidates go into an examination room where there are two examiners. However, only in one of the two tests do the candidates engage in a paired activity. This article aspires to describe the differences between the two tests in terms of the ways in which the examiner is involved in the candidates' language performance and discuss the implications of the findings for the two types of oral examination.

☙

Οι επιστήμονες που διερευνούν την προφορική δοκιμασιολογία αναφέρονται μονίμως στη πολυπλοκότητα αυτής της διαδικασίας η οποία και οφείλεται στο πλήθος των παραγόντων οι οποίοι επηρεάζουν το τελικό της αποτέλεσμα. Ένας από αυτούς τους παράγοντες είναι ο εξεταστής, ο οποίος έχει αποδειχτεί ότι επηρεάζει την απόδοση του υποψηφίου μέσω του ρόλου του και ως συνομιλητή και ως βαθμολογητή. Γίνεται επίσης μεγάλη συζήτηση για τα χαρακτηριστικά αυτού του ρόλου στην ονομαζόμενη εξέταση σε ζευγάρια σε αντιδιαστολή με τη συνέντευξη γλωσσομάθειας προφορικού λόγου. Το παρόν άρθρο εξετάζει δύο εξετάσεις προφορικού λόγου του ιδίου πιστοποιητικού γλωσσομάθειας, και οι δύο από τις οποίες γίνονται σε ζευγάρια καθώς δύο υποψήφιοι εισέρχονται σε ένα χώρο εξέτασης όπου βρίσκονται δύο εξεταστές. Παρόλ'αυτά, μόνο σε μία από τις δύο εξετάσεις εμπλέκονται οι υποψήφιοι σε δραστηριότητα για ζευγάρι υποψηφίων. Αυτό το άρθρο αποβλέπει στο να περιγράψει τις διαφορές μεταξύ των δύο εξετάσεων όσον αφορά

την εμπλοκή του εξεταστή στην γλωσσική απόδοση του υποψηφίου και να συζητήσει τη σημασία των ευρημάτων για τα δύο είδη προφορικής εξέτασης.

Key words: co-constructed performance, examiner involvement, categories of variation, level of proficiency, interlocutor

Introduction

High-stakes oral proficiency examinations have been synonymous with the oral proficiency interview (OPI), where candidates are usually individually examined by an examiner in a specially arranged room. One of the most important issues arising because of the OPI nature is that many factors or facets are involved in and interact during the assessment process (Ross, 1992; McNamara, 1995, 1997; Lazaraton, 1996; Milanovic & Saville, 1996; McNamara & Lumley, 1997; Brown, 2003, 2005; among others) such as tasks, examiners as 'interlocutors' and 'raters', candidates and assessment criteria.[1] For this reason, researchers have frequently used the term *co-construction* of the candidates' language output (Jacoby & Ochs, 1995; McNamara, 1995; He & Young, 1998; Fulcher, 2003; Brown, 2003, 2005; May, 2010, among others), first introduced by Kramsch (1986) proposing the so-called *Interactional Competence Theory* as an applied linguistics approach. The term co-construction captures what McNamara (1997, p. 459) calls 'the social dimension of interaction'. It refers to the fact that candidates do not speak alone, but their performance is collaboratively constructed, i.e. co-constructed by all the co-participants' contributions; this claim raises concerns for construct validity and reliability in oral testing.

Since the examiner is one of the co-participants, their role has been investigated in many empirical studies. Some researchers have emphasised the asymmetric nature of the interaction taking place in oral interviews; since the examiner is the one who controls the speech event, (that is the candidate cannot introduce topics or change the direction of the conversation), the validity of the interview as a representation of real life interactions is threatened (van Lier, 1989; Perret, 1990; Young and Milanovic 1992; Johnson, 2001; Csepes, 2002). Additionally, studies have looked into examiners' variation, their distinct or personal styles and the ways in which these might affect the candidates' performance (Ross, 1992; Ross & Berwick, 1992; Lazaraton, 1996; Brown & Lumley, 1997; Brown, 2003, 2005). Research on the issue of inter- and intra-examiner-as-interlocutor variation concludes that it might threaten the validity and reliability of any examination, and therefore, should be dealt with continuous monitoring of the examiners and examiner training.

The paired exam

Growing awareness of the aforementioned issues has oriented language testers towards the introduction of the paired testing pattern, during which there is one or more activities of interaction between the two candidates. The pair-work approach has its origins in classroom language learning throughout the world.

An increasing amount of literature provides theoretical and empirical data about the paired approach testing promoting its advantages. Ikeda (1998, p. 71) proposes the paired learner interview as 'an effective means to reduce communicative stress [...] and elicit authentic learner participation'. Iwashita (1996) suggests that peer-peer interaction creates a non-threatening environment and

generates similar scores to those obtained by the traditional interview (also in Norton, 2005). Együd and Glover (2001) argue that candidates can use language not as 'inferiors' (candidates) addressing 'superiors' (examiners), but as they normally use it in everyday speech situations. Saville and Hargreaves (1999) and Galazci (2003) argue that the paired format allows for more varied patterns of interaction, which is also advocated by Taylor (2000, 2001), who talks about a greater range of functions in the paired speaking test, and Brooks (2009) who found higher complexity in the test takers' interaction. Other researchers have looked into the paired format from the raters' perspective. May (2009) and Ducasse and Brown (2009) conclude that raters do recognise and assess the way(s) candidates contribute to successful interaction. And although discourse in paired activities is collaboratively produced by definition since there are two interlocutors-candidates who must be assessed separately, such co-construction is not regarded as a negative feature of the paired activity, but should be embraced as being more reflective of real world communication (Brookes, 2009, p.361). As long as interlocutor variability in peer-peer interaction is regarded as 'part of the ability construct we are interested in measuring' (Taylor & Wigglesworth, 2009, p.332) and appears to 'directly inform the assessment scale' (Nakatsuhara, 2003, p 22), assessment is achieved in a reliable and valid way.

This paper – being part of a more extensive PhD research – looks into the paired-type activity from the point of view of the examiner and the potential value of this testing pattern in the elimination of examiner variation. The study compares existing variation in examiners' performance in oral exams of two different levels by presenting a comparative study of the role of the examiner-as-interlocutor in the two tests. It also draws some conclusions concerning the examiners' role as this depends on the level and type of activity (paired or not paired). Finally, it reports on data from three different sources: observation of actual oral examinations, oral examiners' feedback forms and simulated oral examinations.

The context of the study

The present research was conducted by the Research Centre for Language Teaching, Learning and Assessment (RCEL) of the Faculty of English Studies, University of Athens, Greece[2], within the context of the Greek State Certificate of English Language Proficiency exams, known as KPG exams, and more specifically, the oral tests at B2[3] and C1 levels. These are the two levels the KPG battery started with, and also, in which the role of examiner as interlocutor has been defined quite differently by construct (see below). In both levels, two candidates enter the examination room and are examined by one examiner, while another examiner is also present. This second examiner is an observer who only assigns marks and does not participate in the speech event. At the end of the test, the examiner who conducts the interview also marks the two candidates. Both examiners use the same set of criteria, i.e. rating scale, to assign marks for separate skills or competences which are then added to produce two total marks (one from each examiner). The average of the two marks provides each candidate's final oral test mark. (See Appendix for the content and structure of the KPG oral tests at the B2 and C1 levels).

The role of the examiner as interlocutor differs in the two levels. At B2 level, the examiner reads out the questions (Activity 1) and tasks (Activities 2 and 3) to each of the two candidates, thus interacting with each one of them in turn. The candidates go into the examination room together, for reasons of organisation and time economy, but they do not speak to each other at all; thus the speaking test is not a paired test in the form traditionally discussed in the literature. On the other hand, at C1 level, the examiner reads out one question (for each test-taker) in Activity 1 for the purpose of which s/he

interacts with each of the two candidates; then in Activity 2 s/he becomes a listener while the candidates engage into interaction in order to reach a decision or solve a set problem on the basis of input presented in reading texts in Greek.

Regarding the rating scale, candidates are assessed on the basis of different types of criteria in the two levels, (see Table 1). Only at B2 level, they are evaluated for task achievement. Only at C1 level, test-takers are assessed for interactional competence because there is a task-type requiring interaction between the two candidates.[4] There are some more differences but they fall outside the scope of the present article.

B2	C1
Overall performance on task (i.e. the degree to which the candidate has responded to the requirements of the task)	
■ Dialogue (0-2)	
■ One sided talk (0-2)	
■ Mediation (0-2)	
Overall language performance (i.e. the quality and level of candidates' output in relation to certain criteria)	
■ Phonological competence (0-2)	■ Phonological competence (1-2)
■ Linguistic competence (0-4)	■ Lexical range and control (0-3) ■ Grammatical accuracy (0-3)
■ Sociolinguistic competence (0-4)	■ Appropriateness of language choices (0-3)
■ Pragmatic competence (0-4)	■ Cohesion/coherence/fluency (0-3) ■ Conversational competence (0-3)
	■ Mediation (0-3)

Table 1: The B2 and C1 Oral Assessment Criteria

The study

The present study is based on the many-relevant-research findings that inter- and intra-examiner variation might threaten the validity and reliability of the examination and therefore should be controlled. More specifically, it examines whether variation, also called *intervention* or *involvement*, differs according to the level of the exam as well as the design of it.

Research questions

This paper seeks to answer the following questions.

- In what ways does examiner-as-interlocutor involvement differ between B2 and C1 KPG oral tests?
- Can the differences be attributed to the level, the type of activities involved, or both?

This study draws data from three different sources to shed light into the ways examiners-as-interlocutors affect the candidates' language output in two very different tests of oral proficiency. These three sources are separately described below along with their results, which are further discussed in the final section of this article.

The KPG Oral Examiners Observation project

The KPG observation project commenced as a pilot study in November 2005. Since then, it has been conducted in five more phases providing invaluable information about the efficiency of oral examiner conduct as well as other issues related to the test procedure and administration. As such, the observation project constitutes an on-going effort of the RCEL to control and monitor examiners' performance in the KPG oral tests. It is carried out in the biggest examination centres all around Greece, thus producing a representative amount of data in terms of both quality and quantity (see also Delieza, 2008 and Karavas & Delieza, 2009)

Data for this paper are drawn from the third and fourth phases of the observation project, carried out in May and November 2007. These phases aimed to collect information in relation to the type and frequency of interlocutor intervention in each of the test activities. Through previous piloting observation phases, the assigned observers (the writer herself among them) had detected and listed various types of interlocutor interventions, which are presented in Table 2 below; these types were classified into two general categories, namely *change or interference with the questions/tasks rubrics* and *interruption of candidate or interference with their language output*.[5]

Changes to or interference with the rubrics	**Interruption of the candidate or interference with his/her language output in order to**
• Use of an introductory question • Change of one-two words from the rubric • Supplying a synonym for a word • Expansion of the original exam question • Explanation of the rubric (through the use of examples) • Repetition of the rubric (more slowly)	• redirect the candidate because s/he misunderstood something by repetition of the rubric or part of it • make some kind of correction • supply one or more words the candidate was unable to find • add something

Table 2: Types of examiner intervention identified by observers.

Since this paper looks into the differences between the two levels in terms of the role of the examiner-interlocutor, I only present results directly relevant to this issue. Thus, in May 2007, 32 observers observed 156 examiners who examined 588 candidates for the B2 level oral test and 105 examiners examining 342 candidates for the C1 level oral test. In November 2007, 42 observers observed 133 examiners examining 514 B2 candidates and 66 examiners examining 232 C1 candidates.

Table 3 presents results (percentages and numbers) for the two general types of intervention for each Activity in the two levels for the two observation phases.[6] Percentages of interventions have been calculated on the basis of the total of candidates observed, since instances of intervention have been counted per candidate.

Comparing the results from the two phases it is obvious that C1 examiners generally tend to intervene less than B2 examiners. It becomes clear that C1 examiners make changes to the question/task rubrics much less frequently than B2 examiners; this is seen in both sets of findings (for instance, May 2007, Activity 1 – B2=57% as opposed to C1=22.5%). Examiners, therefore, appear to feel the need to provide support to and/or facilitate candidates of B2 more than those of C1 level. Furthermore, percentages in November 2007 are generally higher in all B2 Activities and in C1

Activity 1, in which the examiner is the candidate's interlocutor, than in C1 Activity 2, in which the two candidates interact, the examiner being a mere listener.

May 2007 **Intervention per activity**	**B2** (588 in total)			**C1** (342 in total)	
	Activity 1	**Activity 2**	**Activity 3**	**Activity 1**	**Activity 2**
Changes to the rubrics	57.5% (338)	31% (184)	33% (194)	22.5% (77)	23.5% (80)
Interruptions and/or interferences	20.5% (121)	33.5% (198)	28.5% (168)	***36%*** (123)	***55.5%*** (188)
November 2007 **Intervention per activity**	**B2** (514 in total)			**C1** (232 in total)	
	Activity 1	**Activity 2**	**Activity 3**	**Activity 1**	**Activity 2**
Changes to the rubrics	30% (155)	23% (117)	22% (112)	21.5% (53)	13.5% (31)
Interruptions and/or interferences	25% (128)	40% (204)	30% (152)	***34.5%*** (80)	21% (49)

Table 3: Results of examiner intervention per activity in the B2 and C1 level oral tests

There are three percentages (depicted in bold letters in the table) which appear to contradict the assumed preference of examiners to intervene in B2. Firstly, *36%* of the C1 examiners in May 2007 interrupted the candidates or interfered with their language output in some way in Activity 1 and *55.5%* in Activity 2. Karavas and Delieza (2009) attribute these percentages to the tendency of examiners to a) expand on the given opinion question in Activity 1, where the candidates sometimes do not produce enough assessable language; and, b) intervene in Activity 2 in order to remind the candidates of test procedures and requirements which they sometimes forget, since the two of them are in a process of information exchange and interaction, which the examiner is supposed to listen and only *monitor* if necessary. In Activity 1, where examiners are interlocutors, they appear to be *facilitators* of the talk, while in Activity 2, where they are listeners, they act as *instructors*, ensuring the procedure is conducted according to regulations.

The third somewhat odd percentage is *34.5%* in C1 Activity 1, in November 2007, since examiners are not expected to intervene to such an extent at this level. Again, it can be explained by the fact that examiners as interlocutors in Activity 1 often expand on the candidate's answer in order to help them produce longer or more complete answers.

Feedback from Examiners

All KPG oral examiners participating in the oral examination are asked to complete *anonymously* the Oral Examiner Feedback Forms at the end of each examination day. Some of the content of these forms has varied from some examinations to others, but their core questions always refer to the efficiency of the questions and tasks. With the introduction of an Interlocutor Script[7] in the English KPG oral test which was introduced in November 2007, some questions relevant to the examiner conduct were added in the feedback forms. These questions concerned the use and efficiency of the newly introduced Interlocutor script; they also asked the examiners to state how often they changed or interfered with the questions or tasks rubrics and how often they interrupted the candidates while they were talking. Table 4 presents the questions (4a, 4b and 6a, 6b) from the feedback form which relate to this study, along with the results for the two levels. For B2 level, 258 Feedback forms

and for C1 level 180 Feedback forms were collected and analysed.[8] Table 4 shows the frequencies of what examiners themselves stated they did in terms of activity rubrics and interruptions. The last column shows the percentage of forms in which the questions were answered, as in a few cases they were left blank.

B2 level (total of feedback forms=258)				
4. Did you change or interfere with the rubrics in any of the following ways?	**VERY OFTEN**	**SOMETIMES**	**NEVER**	**% of forms**
a. Change one-two words and/ or supplying a synonym for a word.	2.71% (7)	66.67% (172)	29.46% (76)	98.84% (3 not answered)
b. Expand the question and/ or use examples to explain.	2.71% (7)	50.78% (131)	43.02% (111)	96.51% (9 not answered)
6. Did you generally interrupt the candidates or intervene while they were talking in order to:	**VERY OFTEN**	**SOMETIMES**	**NEVER**	**% of forms**
c. correct or add information?	0% (0)	20.54% (53)	72.48% (187)	93.02% (18 not answered)
d. help the candidate by repeating the whole or part of the question?	3.88% (10)	71.32% (184)	21.71% (56)	96.91% (8 not answered)
C1 level (total of feedback forms=180)				
4. Did you change or interfere with the rubrics in any of the following ways?	**VERY OFTEN**	**SOMETIMES**	**NEVER**	**% of forms**
a. Change one-two words and/ or supplying a synonym for a word.	2,22% (4)	38,33% (69)	57,78% (104)	98.33% (3 not answered)
b. Expand the question and/ or use examples to explain.	32,22% (58)	0,56% (1)	62,22% (112)	95% (9 not answered)
6. Did you generally interrupt the candidates or intervene while they were talking in order to:	**VERY OFTEN**	**SOMETIMES**	**NEVER**	**% of forms**
c. correct or add information?	0,56% (1)	14,44% (26)	77,78% (140)	92.78% (13 not answered)
d. help the candidate by repeating the whole or part of the question?	46,67% (84)	3,33% (6)	45,56% (82)	95.56% (8 not answered)

Table 4: Results for questions 4 and 6 of the November 2007, B2 and C1 Oral Examiner Feedback Forms

Almost 67% of the B2 examiners who answered question 4 sometimes make some kind of change to the question or task rubric given in the Examiner Pack, and almost 51% expand the question or use examples to explain it. Because the questions are verbalised in a generalised way, it is not clear how the examiners exactly interfere with the given rubrics (henceforth the need for discourse analysis). However, more than two thirds of the examiners admit that they sometimes 'tamper' with the questions or tasks, which is a major threat to the reliability of the examination.

In question 6, findings are more encouraging. Almost 21% of the examiners, who answered this question, state that they sometimes corrected the candidates or added information while they were talking, while 72% did not do so at all. 21% is not negligible, this being a practice which examiners have been advised to avoid. Although the use of correction or elaboration can be explained by the

fact that KPG examiners are also EFL teachers, it cannot but be eliminated as a source of examiner variability which may threaten the validity and reliability of the test. Additionally, around 71% of the examiners stated that they sometimes used repetition of the whole or part of the questions or tasks. Repetition has been promoted by KPG oral exam designers and examiner trainers as an efficient way of helping the candidate out of a trouble situation (e.g. being 'stuck' or showing lack of understanding etc.) without providing the linguistic means to do so; i.e. without supplying the candidate with language which s/he is expected to produce.

Findings for C1 are different from B2. First of all, almost 58% and 62% of the examiners, respectively, stated that they never changed the rubrics or expanded them etc. (questions 4a and 4b). This could be attributed to two factors, the first being the level of the candidates' linguistic competence. The second is the amount of questions and tasks in the C1 test. The examiner is supposed to ask *one* opinion question in Activity 1 (as opposed to *two* to *four* personal questions in Activity 1 in B2). Moreover, the examiner assigns one task in Activity 2 (as opposed to two tasks, one for Activity 2 and one for Activity 3 in B2) in which s/he is a mere listener, allowing the candidates to engage in a long conversation (as opposed to his/her being each candidate's interlocutor). However, percentages are quite high for 'sometimes' (38%) in 4a and for 'very often' (32%) in 4b. Observation of actual exams has shown that these types of intervention may be connected with the Activity 1 opinion question which is sometimes a source of 'trouble' in two ways. Either a word or phrase is incomprehensible for the candidate or the candidate fails to answer fully, both of which cause explanation and/or expansion of the given question.

In relation to question 6c, the answers are similar with B2 level: almost 78% of the examiners stated that they never correct candidates or add information to their language production. Concerning question 6d, however, almost half of the examiners state that they *very often* use repetition to help the candidates or that they *never* do so. This can be explained as follows. Firstly, repetition is one of the strategies examiners have been advised (by the KPG exam developers) to utilise with candidates in their effort to assist language production. Secondly, drawing from experience, C1 candidates sometimes ask for repetition of the opinion question in Activity 1 and also need to be reminded of part of their task while they have been interacting for some time in Activity 2. When repetition is not used, it is probably because higher proficiency candidates may not ask or need to be reminded of the question or task as often as lower proficiency ones.

To conclude, it is evident that examiners at B2 level tend to intervene more often than in C1, by changing rubrics, or expanding the question/task or through correction and addition of information and also repeating the whole or part of the question/task. It appears that examiners at B2 level seem to try to facilitate candidates more than at C1, either because of the candidates' level or because of the role they are supposed to play in the communicative event, or for both of these reasons.

Simulated Oral Tests

From October to December 2006, 14 simulations of the actual KPG oral exams were conducted (7 for each level) with learners preparing to take the KPG exams in May 2007. These simulations were conducted by highly experienced KPG examiners but without the Interlocutor Script – since this was only introduced in November 2007. These simulations were video- and audio-recorded and then transcribed. The transcribed data analysis produced valuable findings both in quantity and quality. The coding of the types of intervention was done through careful study of the transcripts; different categories arose inductively through analysis of the raw data and also deductively through use of the

findings from observation and relevant studies (Ross, 1992; Ross & Berwick, 1992). Within the limits of the present article, I will present some quantitative data which can be compared to the data presented in the two sections above. Table 5 presents the most frequently used types for each activity in the two levels.

The categories of intervention outlined in Table 5 are not directly comparable to the ones produced though observation and feedback forms. This comes as a result of the discourse analysis and coding of the transcribed data. Close examination of the language used by examiners revealed similarities among some types, as these were defined in observation and feedback collection, but rendered their 'boundaries' unclear. Therefore, types of involvement were reconsidered and re-defined. On the basis of the new analysis, *expansion* includes any addition to the candidate's language production, i.e. providing words or phrases and asking further questions. *Repetition* was defined as a general type including repetition of a question, a task or part of them and repetition of the candidates' word(s). *Explanation* is a type which includes mainly explanation of a word/words in the given question or task, and this differentiates it from expansion. Finally, *comment/evaluation* is a type which came to light through this study and had scarcely been detected by observers.

B2	No of interventions per activity	Total no of interventions = 279			
		expansion	repetition	explanation	comment /evaluation
Act 1	133	8.6% (24)	*12.2% (34)*	4.3% (12)	3.9% (11)
Act 2	79	9.3% (26)	10.8% (30)	0.7% (2)	0.7% (2)
Act 3	67	*12.5%* (35)	4.7% (13)	0.7% (2)	0.4% (1)
C1	No of intervention per activity	Total no of interventions = 86			
		expansion	repetition	explanation	comment /evaluation
Act 1	50	*15.1%* (13)	7% (6)	10.5% (9)	5.8% (5)
Act 2	36	*18.6%* (16)	5.8% (5)	2.3% (2)	0.0% (0)

Table 5: Most frequent types of intervention in simulated oral tests in the B2 and C1 levels

As shown in Table 5, examiner intervention is much more frequent in B2 than in C1, which was also indicated by the observation and feedback forms results. Additionally, it is in B2-Activity 1 that examiners mostly prefer to involve themselves in the candidates' language output. It could be argued that examiners tend to facilitate candidates because Activity 1 is the introductory activity to the whole examination and consists of a series of (two to four) personal questions asked by the examiner-interlocutor. Repetition is the most frequently used type of intervention (12.2%) in this activity, while expansion also lies among the most preferable choices of examiners (8.6%).

Explanation and comments/evaluation are also frequent. It appears that examiners involve themselves in more ways than they do in Activities 2 and 3 in B2 level, in which intervention is much less frequent in the first place (79 and 67 as opposed to 133). In Activity 2, examiners most frequently choose to make repetitions (10.8%) but also expansions (9.3%), the frequency of these two types being very close, while the percentages for explanation and comment/evaluation are almost negligible. In Activity 3, expansion (12.5%) is much more frequent than the rest, some of which are scarce. Activities 2 and 3 are more cognitively challenging (according to their construct definition) requiring the candidates to carry out a task on the basis of one or more pictures and

Greek input text respectively. Given that these tasks are sometimes long and consisting of sub-questions (see example in Appendix), examiners tend to facilitate the candidates either by repeating part(s) of them or expanding them in order to prompt the candidates to speak.

C1 examiners get remarkably less frequently involved than B2 examiners (86 interventions as opposed to 279 respectively), which can be attributed to their expectations of the candidates' proficiency level as defined by the specifications as well as the nature of the role they play in the process. C1 examiners ask each candidate one opinion question (Activity 1) and then assign a task to both candidates to be conducted on the basis of Greek texts (Activity 2), thus restricting the role of the examiner to that of a listener. As can be seen in Table 5, examiners get more involved in Activity 1 than in Activity 2 (50 as opposed to 36 interventions). In Activity 1, they most frequently expand on or explain the question and much less frequently repeat it or make comments/evaluations. This can be attributed to two reasons: a) the opinion question itself causes trouble because of a word or phrase which creates difficulty[9] and b) according to the examiner, the candidate does not fully answer the question. In Activity 2, although the percentage appears to be the highest (18.6%), in fact, it is only 16 times that expansion was used in all 7 simulated tests. These 16 interventions were actually against test-conduct instructions and can be attributed to the lack of Interlocutor script and the personality and choices of the specific examiner, who opted for a more 'interventionist' role in this Activity. Such tactics are not uncommon in actual exams, thus threatening the validity and reliability of the procedure and its outcome.

Discussion and conclusions

This article presents data from three studies investigating the differences in interlocutor conduct and the extent to which they depend on the level, the paired activity format or both. Although the three studies are not directly comparable due to differences in methodology, they all offer insights into interlocutor variation.

First of all, examiners tend to use different types of intervention, and therefore vary in the way they conduct the test, more frequently in the B2 exam than the C1. It appears that when they are given the role of the sole interlocutor of the candidate, they have the tendency to use more facilitative techniques –even if they have been advised to be mere deliverers of questions and tasks. On the other hand, they prefer to intervene mainly in a more instructive role when they are listeners of a paired-type task. This is especially evident in the C1 exam, where examiners get more involved in Activity 1. Examiners themselves ask each candidate an opinion question while in Activity 2, they remain silent, listening to the candidates' interaction and interfere mainly for instructive reasons.

Additionally, comparing all activities in both levels, examiner intervention is more frequent in Activity 1 of the B2 exam since it consists of personal questions which seem to be more prone to changes or expansions by the examiner-as-interlocutor. Although this activity is called 'Dialogue' and it is not supposed to be conducted as any dialogue in a real-time situation, examiners seem to be undertaking a 'freer' and/or more accommodating kind of interlocutor role. It is, finally, possible that, because it is the very first activity, examiners tend to use facilitation strategies in order to encourage the candidates. In contrast, examiners generally refrain from getting involved in C1 Activity 2, which is the only peer-peer interaction activity. Examiners mainly intervene in an instructive role, to repeat the task or remind the candidates of procedure details.

The analyses also showed that repetition is a strategy frequently used by examiners. This is a very positive fact for the way KPG exams are conducted, because repetition can be a type of involvement which will not provide the candidates with the linguistic means to continue thus affecting their language output.

The results presented in this paper contribute to the literature on examiner variation in oral tests by shedding light into differences between levels of oral proficiency (KPG B2 and C1) as well as between two activity patterns (paired and non-paired). It also provides evidence that the involved candidates' performance seems to be co-constructed by examiners in cases where the latter intervene. Evidently, further study into the effect of types of variation on candidate language output as well as on the final score could further support this idea of co-construction. Nevertheless, it could be stated that examiners tend to be factors of co-construction more frequently in the B2 exam than in the C1 exam and more frequently in the non-paired activities than in the paired ones. This has implications for oral examiner training in the assessment of proficiency at different levels and also supports the value of the paired activity itself.

In conclusion, examiner variability in oral proficiency tests has always been a major concern for high-stakes proficiency testers. Continuous research into and monitoring of oral tests provides information about the high complexity of the oral test procedure. This article shows in what ways examiners vary in the practices they use when conducting two tests whose level and interactional pattern differ. It has offered some insight into the ways examiners collaboratively construct the candidates' performance and highlighted a positive aspect of paired oral activities.

However, more research into oral tests is required, since many variables affect such procedures – whatever the types of activities involved in them. Additionally, the co-constructed nature of oral performance in both paired and non-paired activities is an issue awaiting more empirical research, because safe conclusions should be drawn on the ways co-construction is or can be included in assessment scales or criteria and then internalised, interpreted and applied by raters.

Evidently, such research results provide material for continuous and vigorous training of examiners and raters aiming at dealing with the complexity and elusiveness of oral communication in tests of oral proficiency.

Acknowledgements
I would like to extend my honest gratitude to Professor Bessie Dendrinos, main PhD supervisor and RCEL Director, for her continuous supervision and essential guidance in my study. I am also thankful to Assistant Professor Evdokia Karava (PhD supervisor and RCEL Deputy Director) for her kind co-operation and invaluable assistance in my research. Finally, my very special thanks go to Dr Dina Tsagari and Dr Spiros Papageorgiou, for the review of this article, as their insightful comments greatly contributed to its improvement. This study has been partly supported by the RCEL research funding programme. However, opinions expressed herein belong to the author and do not reflect formal policy of the centre.

Author's email: xdelieza@enl.uoa.gr

Notes

1. Milanovic and Saville (1996, p. 6) propose a most comprehensive diagram of variables interacting with each other and affecting assessment.
2. The writer – in her capacity as a research assistant at the RCEL – coordinated (and *participated as real-time observer* in) the observation project with KPG candidates from November 2005 to November 2008 examination periods. This project is part of a larger ongoing research project, directed by Prof. V. Dendrinos and Assist. Prof. E. Karavas, investigating the KPG oral exam.
3. Since May 2011, B1 and B2 have become an integrated exam.
4. B2 – Activity 1 is a Dialogue only to the extent that examiners deliver 2-4 personal questions and test-takers answer them – they do not engage in some kind of conversation and, for this reason, are not assessed for interactional competence (for details see Karavas, 2009).
5. Karavas and Delieza (2009) present the results from the May 2007 phase with special reference to a) their impact on the improvement of the oral test by means of the use of an interlocutor script – which was introduced in all English KPG oral exams in November 2007; and b) on the training of the oral examiners – who were re-trained and supplied with a list of 'acceptable and non-acceptable types of intervention' and were further evaluated on the applicability of these instructions through observation again in November 2007.
6. For detailed results of different types of intervention per activity for May 2007, see Karavas and Delieza (2009).
7. An Interlocutor Script or Frame is a set of written instructions which examiners read out to candidates guiding them through the examination; it is similar to the scripts actors use when rehearsing their lines.
8. C1 feedback forms were fewer than the B2 ones because the C1 level usually has at least 50% fewer candidates than the B2 level.
9. They sometimes contain sayings or difficult words – see example in Appendix.

References

Brooks, L. (2009). 'Interacting in pairs in a test of oral proficiency: Co-constructing a better performance.' *Language Testing,* 26/3: 341-366.

Brown, A. (2003). 'Interviewer variation and the co-construction of speaking proficiency.' *Language Testing,* 20/1: 1-25.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt am Mein: Peter Lang.

Brown, A., & Lumley, T. (1997). 'Interviewer variability in specific-purpose language performance tests.' In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: proceedings of LTRC 96*. Jÿvasskÿla: University of Jÿvasskÿla and University of Tampere, 173-191.

Csepes, I. (2002). 'Measuring oral proficiency through paired performance.' Unpublished PhD Dissertation, Eötvös Loránd University.

Delieza, X (forthcoming). 'Monitoring KPG examiner conduct.' *Directions,* 1. RCEL publications, University of Athens.

Ducasse, A.M. & Brown, A (2009). 'Assessing paired orals: raters' orientation to interaction.' *Language Testing*, 26/3: 423-443.

Együd, G. & Glover, P. (2001). 'Oral testing in pairs: a secondary school perspective.' *ELT Journal*, 55/1: 70-76.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Galazci, E.D. (2003). 'Interaction in a paired speaking test: the case of the First Certificate in English.' *Cambridge ESOL Research Notes*, 14: 19-23.

He, A. W. & Young, R. (1998). 'Language proficiency interviews: a discourse approach.' In R. Young, & A. W. He, (Eds) *Talking and testing. discourse approaches to the assessment of oral proficiency.* Amsterdam: John Benjamins, 1-23.

Ikeda, K. (1998). 'The paired learner interview: a preliminary investigation applying Vygotskian insights.' *Culture and Curriculum,* 11/1: 71-96.

Iwashita, N. (1996). 'The validity of the paired interview in oral performance assessment.' *Melbourne Papers in Applied Linguistics*, 5/2: 51-65.

Jacoby, S. & Ochs, E. (1995). 'Co-construction: an introduction.' *Research on Language and Social Interaction*, 28/3, 171-183.

Johnson, M. (2001). *The art of non-conversation: a re-examination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Karavas, E. (Ed.) (2009). *The KPG speaking test in English: a handbook*. University of Athens: University of Athens RCEL Publications.

Karavas, E. & Delieza X. (2009). 'On site observation of KPG oral examiners: implications for oral examiner training and evaluation.' *Apples – Journal of Applied Language Studies*, 3/1: 51-77.

Kramsch, C. (1986). 'From language proficiency to interactional competence.' *Modern Language Journal*, 70, 366-372.

Lazaraton, A. (1996). 'Interlocutor support in oral proficiency interviews: the case of CASE.' *Language Testing* 13: 151-172.

May, L. (2009). 'Co-constructed interaction in a paired-speaking test: the rater's perspective.' *Language Testing*, 26/3: 397-421.

May, L. (2010). 'Developing speaking assessment tasks to reflect the 'social turn' in language testing.' *University of Sydney Papers in TESOL,* 5: 1-30.

McNamara, T. F. (1995). 'Modelling performance: opening Pandora's box.' *Applied Linguistics* 16/2: 159-179.

McNamara, T. F. (1997). 'Performance testing'. In C. Clapham & D. Corson (Eds.), *Language testing and assessment.* Dordrecht: Kluwer Academic Publishers, 131-139.

McNamara, T. F., & Lumley, T. (1997). 'The effect of interlocutor and assessment variables in overseas assessment of speaking skills in occupational settings.' *Language Testing,* 14: 140-156.

Milanovic, M., and Saville, N. (1996). *Performance testing, cognition and assessment: selected papers form the 15th language testing research colloquium*. Cambridge: Cambridge University Press.

Nakatsuhara, F. (2006). 'The impact of proficiency-level on conversational styles in paired speaking test.' *Cambridge ESOL Research Notes*, 25: 15-20.

Norton, J. (2005). 'The paired format in the Cambridge speaking tests.' *ELT Journal*, 59/4, 287-297.

Perret, G. (1990). 'The language testing interview: a reappraisal.' In J. de Jong & D. K. Stenenson (Eds.) *Individualising the assessment of language abilities*. Philadelphia. PA: Multilingual Matters, 225-238.

Ross, S. (1992). 'Accommodative questions in oral proficiency interviews.' *Language Testing* 9: 173-185.

Ross, S., and Berwick, R. (1992). 'The discourse of accommodation in oral proficiency interviews.' *Studies in Second Language Acquisition,* 14: 159-179.

Saville, N & Hargreaves, P. (1999). 'Assessing speaking in the revised FCE.' *ELT Journal*, 53: 42-51.

Taylor, L. (2000). 'Investigating the paired speaking test format.' *Cambridge ESOL Research Notes*, 2: 14-15.

Taylor, L. (2001). 'The paired speaking test format: Recent Studies.' *Cambridge ESOL Research Notes*, 6: 15-17.

Taylor, L. & Wigglesworth, G. (2000). 'Are two heads better than one? pair work in L2 assessment contexts.' *Language Testing*, 26/3: 325-339.

van Lier, L. (1989). 'Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversations.' *TESOL Quarterly,* 23/3: 489-508.

Young, R. & Milanovic, M. (1992). 'Discourse variation in oral proficiency interviews.' *Studies in Second Language Acquisition,* 14: 403-424.

Appendix

The B2 and C1 levels exam structure and content accompanied by *examples* for all activities.

B2	**C1**
Duration of test	
15-20 minutes	20 minutes
Pattern of participation	
Candidates are tested in pairs but do not converse with each other.	Candidates are tested in pairs and converse with each other in Activity 2.
Content of oral test	
a) **Dialogue** (3-4 minutes) between examiner and each candidate who answers questions about him/herself and his/her environment posed by the examiner. For instance, the candidate is asked questions such as '*Do you have a lot of friends or just a few close ones? Tell us about them',* and/ or, '*Do you prefer listening to music at home or going to live performances? Why?*' and/ or, '*What would be the ideal school environment/ working environment for you? Why?*' (November 2007 – English KPG – B2 – Module 4, Examiner Pack, page 2). b) **One-sided talk** (5-6 minutes) by each candidate who develops a topic on the basis of a visual prompt. For instance, the candidate is shown a page with pictures depicting 'People's emotions' (November 2007 – English KPG – B2 – Module 4, Candidate Booklet, page 6) and the relevant task is '*Look at photos 1 & 2. Tell us how you think the people are feeling, what has happened and what is going to happen next*' (Examiner Pack, page 3). c) **Mediation** by each candidate who develops a topic based on input from a Greek text. (6 minutes for both) For instance, the candidate is given text in Greek about 'how to take care of contact lenses'. (November 2007 – English KPG – B2 – Module 4, Candidate Booklet, page 12) and the relevant task is '*Imagine I am going to wear contact lenses for the first time. Using information from Text 1, give me some advice on how to take care of them*' (Examiner Pack, page 4).	a) **Warm-up** (not assessed – 1 minute) Examiner asks each candidate a few ice-breaking questions (age, studies/work, hobbies) b) **Open-ended response** (4 minutes): The candidate responds to a single question posed by the examiner expressing and justifying his/her opinion about a particular issue/topic. For instance, the candidate is asked questions such as, **'***Do you think that some professions are more appropriate for men and some for women?*' or '*Do you think that the saying "Hard work never did anyone harm" is always true? Explain why or why not*' (November 2007 – English KPG – C1 – Module 4, Candidate Booklet, page 2. c) **Mediation and open-ended conversation** (15 minutes): Candidates carry out a conversation in order to complete a task using input from a Greek text. For instance, the candidates are given two different but related texts in Greek about 'ways of saving energy'. (November 2007 – English KPG – C1 – Module 4, Candidate Booklet, pages 6 & 11) and the relevant task is '*Imagine that you have been asked to design campaign leaflets on how to save energy at home. Read your texts, and together decide on the two most useful pieces of advice you would include*' (Examiner Pack, page 3).