Is testing speaking in pairs disadvantageous for students? A quantitative study of partner effects on oral test scores [1]

Csépes Ildikó

Introduction

New, two-level matriculation examinations (`intermediate' and `advanced') in all subjects will be introduced into Hungarian secondary schools in 2005. The new school-leaving exam in English is in the making, too. The test specifications for the exam have been drawn up and will be submitted to the Ministry of Education in order for it to become a decree in March 2002.

There are a number of innovative features that will characterise the new school-leaving exam in English in contrast to the old matriculation examination that is still in effect. A very important feature of the new exam is that it aims to test learners' communicative abilities using all the four language skills (Reading, Listening, Writing and Speaking). This is in marked contrast with the current school-leaving exam, in which students' listening skills are not measured at all, and many of the test tasks focus on eliciting learners' knowledge about the language rather than their ability to use language in order to accomplish a variety of communicative goals.

Unfortunately, the exam developers (members of the Hungarian Examination Reform Project sponsored by the British Council) have been unable to ensure maximum reliability and validity for the two levels of the exam. Although the productive skills (Writing and Speaking) at both levels will be assessed with the help of centrally-produced analytic rating scales and examiners will receive training, double marking will take place only in the case of the advanced Writing exam. According to a current governmental decree (100/1997), in the intermediate Writing exam, class teachers will be expected to do the marking on their own. It is even more worrying for language testers to find that only at the advanced Speaking exam will there be two examiners present: an interlocutor to conduct the exam and a silent assessor to rate candidates' performances. According to the same decree referred to above, the intermediate Speaking exam is not planned to be administered centrally. Therefore, it seems that it will be the class teachers' responsibility to design test tasks as well as to administer the whole exam on their own. Needless to say, such a flaw in the design of the intermediate Speaking exam seriously threatens the reliability and validity of the whole intermediate matriculation exam.

A strikingly new feature of the proposed school-leaving exam in English is that the Speaking exam is planned to be partly conducted in a so-called `paired mode' at both levels. The exam will consist of three tasks: a guided interview, an individual long turn based on picture prompts and a collaborative, two-way task. While the first two tasks of the Speaking exam will be conducted in the traditional, individual mode with the examiner controlling the flow of the conversation, in the third phase two candidates will be paired up to perform on a problem-solving or discussion task together. During the paired task the examiner is expected to withdraw him or herself but may intervene if specific assistance is required. The need for intervention may be prompted because there is a communication breakdown between the candidates, or one of the candidates clearly dominates the conversation, or because candidates have seriously misunderstood the instructions for the task. However, it is clear that the extent to which the communication in this phase of the exam will be successful primarily depends on the candidates, the peer-interlocutors themselves.

Although the paired speaking test appears to be a risky task type from the point of view of

eliciting candidates' `best performance', the test developers' rationale for including it in the new English matriculation examination is related to the alarming results of nation-wide classroom observations (Nikolov, 1999). At the dawn of the 21st century, in the era of communicative language teaching, it would appear that regular classroom interaction in L2 between learners is already an integral part of the English lesson since the extensive use of pair- and groupwork activities is believed to be an essential tool for developing learners' communicative language ability. However, the results of the classroom observations referred to above have revealed that the method used by teachers to develop learners' speaking skills fails to provide an equal chance for all students. The dominant classroom management format is still the traditional, lockstep fashion, where the teacher *initiates*, students *reply* and then the teacher gives *feedback*. Unfortunately, in a large number of cases this IRF cycle seems to engage only a few volunteers in speaking in L2. These findings suggest that oral language practice is very limited in the English language classroom, and there is very little scope for the average learner to develop his speaking skills. The test developers of the new English school-leaving exam believe that if the new exam has a task type that is based on pair-work, it will generate a positive washback effect on classroom teaching and learning. Teachers may feel the need for providing their learners with opportunities for more peer interaction in order to prepare them for the paired task of the oral exam. In a similar vein, students may take speaking activities done in pairs and small groups seriously as these classroom management formats are likely to be perceived by them relevant in terms of exam preparation.

Background to the paired speaking test

Paired oral may seem to be unfamiliar in the Hungarian context in spite of the fact that it has been used as part of large-scale, international, standardised oral proficiency tests since the late 1980s. For example, there are four UCLES oral proficiency exams that make use of the paired format: Key English Test [KET], Preliminary English Test [PET], First Certificate in English [FCE] and Certificate in Advanced English [CAE] (Saville & Hargreaves, 1999). Although the paired oral task type is no longer an experimental format because of its extensive use in international language proficiency examinations, its validity seems to have been researched minimally. One of the handful of studies investigated the effects of candidate acquaintanceship on performance scores (O'Sullivan, 2000). According to the author, familiarity between candidates tends to enhance test performance in the paired oral. He came to this conclusion after finding consistently higher scores given by independent trained raters to pairs where there was supposed to be close familiarity between the two candidates.

The lack of research interest in the paired oral is in marked contrast with the bulk of studies focusing on the traditional, examiner-led format, which has received a considerable attention over the last decade (see e.g. Katona, 1996; Lazaraton, 1996; Luoma, 1997; Ross & Berwick, 1992; Young, 1995; Young & Milanovic, 1992). These studies have mostly highlighted the nature of discourse produced in the traditional oral interview format, and some have even questioned the validity of the interview as a measure of testing candidates' conversational ability (cf. van Lier, 1989). Kormos (1999), however, has found in her study that as opposed to the oral proficiency interview, in guided role-plays the examiner-interlocutor's control over discourse outcomes decreases to such an extent that candidates can be said to have similar rights and duties to the examiner's during the communicative exchange. If we accept Kormos's claim that the guided role-play is a more appropriate means to measure candidates' conversational ability than the proficiency interview, the paired oral can be assumed to measure candidates' conversational ability equally well since candidates performing on paired-tasks are often asked to do a simulation or role-play. The impact of working with a peer partner as opposed to an examiner in role-plays, however, is not understood well enough yet.

The most frequently raised criticisms against the paired speaking test seem to relate to various forms of mismatches between peer interactants. Foot (1999), for example, calls attention to a potential mismatch with respect to candidates' proficiency levels. He warns us that if the latter are markedly different, they are likely to affect both the performance and the assessment of the candidates. He points out that when a candidate has to work with an incomprehensible or uncomprehending peer partner, it may negatively influence the candidate's performance. Furthermore, he claims that in such cases it is quite impossible to make a valid assessment of candidates' abilities. These concerns may appear to be well-founded for some readers, however, it must be emphasised that no empirical study has been conducted to find supporting evidence for the above claims.

The main reason why the investigation of peer interlocutor effects should be included in the current research agenda is that the interlocutor, as a co-constructor of test performance, can be regarded as a potentially important source of variation that may either positively or negatively affect the discourse outcomes and eventually the assessment of candidates' proficiency. As McNamara proposes, "The age, sex, educational level, proficiency or native speaker status and personal qualities of the interlocutor relative to the same qualities in the candidate are all likely to be significant in influencing the candidate's performance" (1996, p. 86). However, it is not known yet whether all the above mentioned factors have a significant impact in terms of enhancing or spoiling candidates' test performance. Nor can we make any solid claims, based on empirical findings, about the nature of the assumed impact.

Research questions

In the present study, the basis for formulating the research question relates to the aforementioned criticism of the paired oral exam. First, the worries were translated into general questions such as:

- What impact does the partner's proficiency level have on a candidate's test score?
- Do candidates' scores vary if they have partners of different proficiency levels? If yes, what kind of variation characterises test scores?

The research question was narrowed down to investigating how the ratings of a specific group of candidates were influenced while talking to partners of different proficiency levels. Differing levels of proficiency between peer interactants were categorised at three broad levels. The three different levels were conceptualised in the following way:

1. higher level of proficiency of one of the candidates in comparison to the proficiency level of the other candidate in a given pair
2. lower level of proficiency of one of the candidates in comparison to the proficiency level of the other candidate in a given pair
3. similar level of proficiency of the two candidates in a given pair

In practice, these three categories were established by selecting a candidate (who will be referred to as a `core student' from now on) and three other candidates who acted as the first candidate's peer partners. The three peer partners were intended to belong to a specific level of proficiency in comparison to the core student. One of them was supposed to have a higher level of proficiency (s/he will be referred to as `top student'). Another one had to be at a lower level of proficiency (`bottom student'), but the third one was intended to be at a similar proficiency level as the core student, and s/he will be referred to as `middle student'. Based on the candidate pairing principle described above, the research question was formulated in the following way:

How do students' (core students) proficiency ratings vary when they perform in the paired

mode with a higher proficiency level partner (top student), a lower proficiency level partner (bottom student) and a similar proficiency level partner (middle student)?

According to the null form of the research hypothesis, there will be no difference in the core students' proficiency ratings resulting from them having to perform with different proficiency level partners. In other words, it was assumed that differing levels of proficiency of peer partners would not produce systematic, construct-irrelevant variation in the core students' performance that would either have a negative or positive impact on their ratings.

Research design

In order to design a systematic investigation of the research question, the contextual variables associated with oral proficiency testing had to be specified in order to make it possible to control for them later. Potential variation of test performance may be due to other variables than the candidate's language proficiency alone. In addition to task as a potential source of variation, test taker characteristics such as age, sex, educational level, proficiency and personal qualities in both candidates are all likely to be significant in influencing performance in the paired mode (cf. McNamara 1996). The way in which these variables were controlled for in the study will be discussed next.

The participants

The target population of this study was 17 and 18-year-old Hungarian secondary school leavers. The participants' educational level and age can be said to be comparable in all candidate pairings since students in each pair attended the same class in a Hungarian secondary school. By selecting students from the same class to form candidate pairings, it was also possible to control for another variable, which was familiarity between the interactants. The latter tends to enhance test performance, which in turn results in higher ratings for them as opposed to candidates who are unfamiliar with each other (cf. O'Sullivan, 2000).

Sex and personal qualities of the participants were not controlled for in the study. The reason why sex was not treated as a major source of variation in performance relates to findings by Alderson (2000), who examined a similar Hungarian student population and found that there was no statistically significant difference between male and female students' mean proficiency scores. Controlling for the subjects' personal qualities, on the other hand, was not found to be feasible for this study. If this variable had been controlled for, it would have made the whole investigation very difficult to carry out since the availability of suitable research subjects was already quite limited for the researcher. According to the main subject selection criteria, the selected candidates were supposed to have relative differences in their proficiency levels, which meant that on average, in one single class it was not possible to find more than four participants whose proficiency levels matched the selection requirements. If one more selection criterion had been introduced (that of candidates' personal qualities), it would have made the identification of suitable participants even more difficult, given the unfortunate circumstance that the researcher had to do the whole investigation single-handed. Therefore, the variable of candidates' personal qualities was not controlled for, and it must be acknowledged that this variable may have influenced candidates' test performance in ways that are difficult to distinguish from the impact of the intended dependent variable, which was candidates' proficiency level.

The participants' proficiency had to be carefully monitored since the main principle for selecting subjects had to do with the level they had reached at the time of data collection. In fact, this variable had to be manipulated in order to answer the research question. It was essential that all candidates' proficiency should be measured in a comparable way, using the same set of instruments, henceforth referred to as `the composite measure'. It consisted of three instruments:

a cloze test, a self-assessment questionnaire and a teacher-assessment questionnaire. An English version of the speaking self-assessment measure was already available thanks to Brian North (1996), and thus it only had to be translated into Hungarian (see Appendix 1). The same scale with minor modifications (1st person singular changed for 3rd person singular) was then turned into a teacher-assessment instrument, while the 38 item cloze test was constructed by the researcher (see Appendix 2). The validity of the Hungarian version of the speaking assessment scale was checked by getting a small group of Hungarian teachers of English to do two rounds of assessment of ten of their students' speaking ability: once using the English and once the Hungarian version of the speaking scale. The comparisons between the two sets of assessments showed no significant variation, which was taken as a confirmation of the validity of the Hungarian translation of North's original speaking assessment scale.

In order to allow for a quantitative study of the central research question, a group of 30 core students had to be identified together with their peer partners (three for each core student), who were expected to fit into the categories described earlier as `top', `bottom' and `middle student'. Thus, altogether 120 students (30 x groups of 4) had to be selected with a wide range of levels of proficiency in English. Ten secondary schools across Hungary (in Debrecen, Nyíregyháza, Békéscsaba and Budapest) were contacted where the author found English teachers who volunteered to participate in the research. As can be seen in Table 1 below, the majority of the participating schools were grammar schools. However, every effort was made to involve students with a variety of language learning backgrounds in order to make the student sample as heterogeneous, that is, representative of the Hungarian student population, as possible.

Eventually, more than 130 final-year students were identified with the help of the composite measure, but only 120 of them were used in the study due to factors such as subjects' failure to meet the pre-specified selection criteria and their absence at the time of administering the oral exam. Table 1 gives an overview of the 120 participants' background with respect to location of school and school type.

Table 1 Overview of the 120 participants' school background

Location	No. of grammar schools	No. of vocational schools	No. of students
Debrecen	4	-	48
Nyíregyháza	1	1	8 + 4
Békéscsaba	1	-	16
Budapest	3	-	44
Total	9	1	120

After assessing the participants' proficiency with the help of the composite measure (cloze test, self- and teacher assessment questionnaires), the students were rank-ordered within their groups (30 groups of four), based on their scores on the three instruments. Students who came out first at least twice in the three rank-order lists were assigned to the top rank students' group. Students who took the fourth rank at least twice on the three rank-order lists were treated as bottom rank students. The two other remaining students in each group of four were assigned to the middle rank students' group. Finally, core subjects were selected randomly from the two middle students in each group.

The tasks

In this research the tasks had an important function: they were the only means by which the researcher was able to ensure that adequate quality and quantity of language is elicited from the

candidates. In contrast to this, in the individual mode (examiner vs. candidate) quality control can be built into the performance framework through examiner training, which means that examiner conduct can be monitored and standardised. In the paired mode the procedure of setting up the task had to be standardised. For this purpose, the following procedure was designed.

The examiner had to read out the task for the candidates, who were allowed to ask for repetition only once in case they did not understand something. Preparation time before commencing the task was specified (maximum 30 seconds) as was the time limit for the whole task (4-5 minutes). The examiner was allowed to intervene if the candidates seemed to get stuck, misunderstood the task or one of the candidates seemed to dominate the conversation. In such cases, the examiner was expected to give a brief prompt, such as

- repeat all or part of the rubric
- invite candidates to talk about one specific aspect of the task
- invite the candidate whose contributions seem to be unsatisfactory (i.e. too short) to talk about one specific aspect of the task.

The guidelines for examiner conduct were based on the interlocutor guidelines developed by the Hungarian Examination Reform Project.

The study was designed in such a way that core students were expected to perform three times with three different proficiency level partners, using parallel speaking tasks. The tasks were developed taking into account some of the findings from task-based research (cf. Skehan, 1998). The most important considerations are summarised below.

The three paired speaking tasks

- were designed to measure candidates' *interactional* skills, providing them with opportunity to bridge an opinion gap between them;
- were designed to be *symmetric* as such tasks seem to generate more interaction and negotiation;
- were *structured* tasks with four different word prompts for each candidate in order to help them produce greater fluency and accuracy by creating expectations and activating prior knowledge and/or experience;
- *allowed for both agreement and disagreement* between the interactants so that they could produce both short and long turns (the former seems to generate less complex language while more complex and varied language may result from the latter), although an agreed outcome was expected by the end of task completion (agreeing on the three most important/useful options);
- related to *life-like situations* that candidates could identify with;
- required candidates to assume *familiar roles* only;
- were designed to be *parallel* in the sense that all the three of them were intended to be of the same level of difficulty in terms of code complexity (the language required) and cognitive complexity (the thinking required).

The same context provided the background to the three parallel tasks: candidates had to imagine

that a foreign student had just arrived in Hungary at their school. In the three different tasks this foreign student had different plans. In Task 1, she wanted to learn Hungarian. In Task 2, she wanted to make friends in Hungary, while in Task 3 she wanted to learn about the Hungarian way of life. Before administering the tasks, four testing specialists of the Hungarian Examination Reform Project were asked to evaluate them. The judges agreed that the three tasks appeared to be parallel, as they seemed to represent a similar level of difficulty. This claim, however, was later verified empirically, too, as task difficulty was examined in the light of candidates' proficiency scores. (see discussion of results below).

Procedures for data collection

All the exams were conducted and audio-recorded by the author during March and April of 2001. The exams were scheduled in school time, when the students had a timetabled English lesson. Data collection from each group of 4 students from the same class was designed to take no longer than a 45-minute lesson. In this way, it was hoped that a minimal amount of disruption in their daily work would be caused for both teachers and students.

Each group of four students always took the exam in the same order, i.e. the paring pattern of the candidates was the same for all the exams.

- First pair: core student with top student
- Second pair: core student with bottom student
- Third pair: core student with middle student.

However, the speaking tasks were rotated in a principled way in order to eliminate an order effect of the tasks. The task sequences were repeated with every fourth group taking the oral exams. Furthermore, the three performances of all core students were recorded on a different cassette in order to prevent the assessors from trying to rate core students' oral performances on the basis of cross-comparisons, that is, by comparing the three performances of a given core student in three different candidate pairings.

The 90 audio-recorded performances were rated by two independent assessors, both female teachers of English in secondary schools in Hungary. The speaking assessment scale used in rating candidates' performances had been developed within the framework of the Hungarian Examination Reform Project (Appendix 3). Because the two assessors had been working in the Project for some years, they were thought to be adequately familiar with the assessment scale and properly trained to apply it.

Results

Subject selection measures

The corner stone in the design of the whole investigation is the subject selection procedure, that is, whether students with appropriate levels of proficiency were chosen and allocated to the different student ranks (top, middle and bottom ranks). Unless there is sufficient evidence to support the validity of subject selection, the findings will automatically lose their validity. Although information (rank order lists) from three proficiency measures (cloze test, self- and teacher assessments) had been triangulated before the final selection of subjects was made, it is possible to investigate the validity of subject selection further. Therefore, a closer look will be taken at the results of the composite measure.

In order to check the validity of the students' allocation to the three proficiency ranks (top,

bottom and middle), the mean scores for each subgroup in the case of each instrument of the composite measure can be compared with one another, using a one-way between-group ANOVA procedure. Table 2 below shows the results of the one-way ANOVA procedure in relation to all the three instruments of the composite measure. The F ratios reflect a statistically significant difference in the scores across all the three instruments.

Table 2 Results of the one-way ANOVA comparing subgroup mean scores for each instrument of the composite measure

	Max score		N	Mean	F	Sig.
Cloze test score	38	Top	30	29.10	74.17	<.0001
		Middle	60	21.87		
		Bottom	30	12.90		
Teacher-assessment score	84	Top	30	61.57	87.74	<.0001
		Middle	60	47.17		
		Bottom	30	30.97		
Self-assessment score	84	Top	30	55.90	49.28	<.0001
		Middle	60	44.18		
		Bottom	30	31.47		

In order to locate where the differences lie, a post hoc comparison of means was also carried out. The results of the Tukey HSD post hoc test showed that there was a significant difference in the mean scores across all the three student rankings for each part of the composite measure. This means that the subject selection and rank allocation based on the cloze test, self- and teacher assessments were appropriate in the sense that students at three markedly different levels of proficiency were selected.

Rater reliability

The reason why two independent raters were employed is that the reliability of scoring can be ensured only if independent assessments of the same set of performances show a significant agreement. The usual way of checking the degree of overlap between two raters' assessment of oral performances is by correlating the two rank orders of candidates on the basis of their test scores. The result of the Spearman rank order correlation between candidates' oral test scores shows that there is significant correlation between the two rank orders, although the correlation coefficient (rho = .806) represents the lower bound of the usual acceptable level.

Task difficulty

The three speaking tasks were intended to be parallel in the sense that they were designed to be of a similar level of difficulty. Approximately the same number of bottom, middle and top students performed on each task in order to eliminate task effect that would have resulted from one particular rank of students performing on one particular task only. Similarity in terms of difficulty level across the three tasks was checked by comparing how the three student groups performed on their own task.

Since the oral test scores given by the two raters failed to show normal distribution, the researcher could only employ non-parametric statistical procedures to perform the data analysis. The Kruskal-Wallis test showed that there was no significant difference between the medians (mean ranks) of the three student groups across the speaking task (Task 1, 2 and 3). As a result, it was confirmed empirically that the speaking tasks could be regarded as parallel versions indeed.

Partner effects

The research question of the study focuses on how ratings are influenced if a candidate has to engage in a pair-task performance with different proficiency level partners. In order to answer the research question, it was hypothesised that there is no difference between the mean ranks of core students, calculated from the scores given by Rater 1 and Rater 2, across the three performance conditions.

As has already been mentioned, the scores by the two raters failed to show normal distribution, and thus only a non-parametric repeated-measures statistical procedure was possible to apply. The Friedman test compares the mean ranks of one group of subjects based on three or more measures. The measures in our case are the three different performance conditions in which the core students worked with different proficiency level partners. Table 3/a shows the mean ranks of core students in the three performance conditions.

Table 3/a Friedman test, comparing the mean ranks of core students based on scores by Rater 1and Rater 2 across the three different performance conditions

	Mean Rank Rater 1	Mean Rank Rater 2
With TOP student	2.02	2.25
With BOTTOM student	2.07	1.90
With MIDDLE student	1.92	1.85

Table 3/b Test of significance in relation to the Friedman test statistics in Table 3/a

	Rater 1	Rater 2
N	30	30
Chi-Square	.393	3.226
df	2	2
Asymp. Sig.	.822	.199

As can be seen in Table 3/b, there is no statistically significant difference between core students' mean ranks across the three performance conditions: the level of significance of the Chi square figures is higher than .05 in the case of both raters. Thus, we can conclude that the null hypothesis has been confirmed. This means that there is no statistically significant variation in the assessments of oral test performance of the selected group of 30 candidates. The scores given by Rater 1 and Rater 2 suggest that their perceptions of core students' proficiency were neither positively nor negatively influenced by the fact that the level of proficiency of core students' partners showed considerable variation.

As was pointed out earlier, the negative impact on test performance due to mismatching partners in terms of proficiency has been emphasised by Foot (1999). The finding that there is no significant difference in the ratings of core students suggests that when a candidate has to work with a peer partner whose proficiency level is lower than that of the candidate, the less able partner does not necessarily influence the given candidate's ratings negatively. Nor can we say that the same candidate's ratings would tend to be positively influenced by a peer partner whose proficiency level is higher than that of the candidate. It seems that partner effects related to proficiency differences between the interactants in the paired mode cannot be predicted as either

harmful or beneficial on the basis of the assumed differences between their proficiency levels.

The finding also suggests that there is no superior performance condition among the three types of conditions examined in the study that is likely to result in higher ratings for candidates in the paired mode. When the candidates were assumed to match each other with respect to proficiency (candidate pairing = core student with MIDDLE student), the ratings did not seem to reflect that this performance condition was more favourable than the other two conditions involving mismatching proficiency level partners. It was found that core students' scores did not differ significantly from scores gained in the other two performance conditions. Based on this statistical evidence in relation to the research question, this study disconfirms the assumption that mismatching proficiency level partners will exert negative influence on performance ratings.

Discussion

The most important implication of the results of the study relates to the validity of the paired speaking exam. Validity is usually interpreted as the appropriateness of a given test for a specific purpose. Oral proficiency tests, irrespective of the fact whether they are conducted in the traditional, individual mode or the paired format, usually aim to measure candidates' ability to perform on interactional and transactional tasks. In constructing oral test performance, it seems that there is a complex interplay between three major variables: candidate - task - interlocutor. The present study was designed in order to minimise the effect of one of the three variables (task) and to manipulate another variable (interlocutor's proficiency) in order to examine its impact on candidates' test scores.

First, task effect was controlled for with the help of carefully designed tasks so that variation in test scores could only be attributed to characteristics either in the candidate or in the interlocutor. Then specific variables concerning the candidates and their interlocutors (peer partners) were controlled for: their age, education background, L1, and familiarity between them. For reasons discussed above candidates' sex and psychological traits (personality), however, were not controlled for. Therefore, it has been acknowledged that candidates' personality may have had an impact on their performance in ways that could not be identified in this study. Finally, one specific interlocutor variable (proficiency level) was manipulated in order to examine its impact on candidates' test scores. Since candidates' oral proficiency was measured on three consecutive tasks without any break in between them, their ratings were expected to be similar as the tasks were of the same level of difficulty. Any potential variation in candidates' test scores was therefore expected to be due to variation in their partners' proficiency level.

Contrary to beliefs in the negative impact of mismatching partners with respect to proficiency, it seems that proficiency gaps between candidates in the paired mode do not generate non-systematic or unwanted variation in their performance ratings. The results of the study showed that there was no statistically significant difference between candidates' performance ratings across different candidate parings, formed on the basis of differing degrees of mismatch or overlap between candidates' proficiency levels. Thanks to the lack of significant variation in scores, we can claim that peer partners' differing levels of proficiency did not affect candidates' oral assessments. This finding in the Hungarian school-leaving exam context, therefore, should be treated as validity evidence in support of the paired speaking exam format. We can conclude that the paired speaking exam seemed to measure candidates' actual oral ability and not something else. We must also add, however, that the lack of peer partner influence on scores only applies if assessments are made with the help of a reliable rating scale, if the degree of inter-rater reliability is acceptable, and if there is evidence that the tasks are comparable in terms of level of difficulty.

When interpreting the findings from this study, we must also consider the limitations of the

research. Although the study was designed to be a quantitative one, when selecting the subject sample, it could not be checked whether the selected group was representative of the Hungarian school-leaver population. The reason for this is that there has been no national survey conducted in Hungary that would make it possible to estimate the extent to which the subjects of the study are representative of the whole student population. Due to the aforementioned limitations of the study, the interpretation of the main finding above is not readily available for generalisation. Moreover, the author does not claim that there is universal lack of partner effect on test scores in relation to mismatching proficiency level partners. In other testing situations different results might be obtained.

Critics doubtful of these findings might feel that since the study was conducted in simulated circumstances, the candidates did not have to take high risks like in a real matriculation examination. Researching testing issues, including partner effects in paired speaking tests, in real examination circumstances, however, is very difficult to accomplish. There is a serious ethical concern that researchers may find difficult to tackle. How fair is it to collect data from candidates who are anxious enough already since a lot is at stake for them depending on the result of the exam: successful completion of secondary school education and for some of them university admission too?

It is encouraging to know, however, that the Ministry of Education has plans for organising large-scale piloting in 2003 and 2004 before introducing the new `Érettségi' in 2005. Thus, it will be possible to investigate partner effects on a larger scale. Nevertheless, it must be noted that pilot exams can hardly be designed to be as controlled as an experimental study like the one described in this paper, and so findings in relation to the impact of differing proficiency level peer partners may be confounded with effects of unintended or uncontrolled variables too.

Conclusion

The results of this study are intended to provide empirical evidence in support of the paired speaking exam to be introduced in the new Hungarian school-leaving exam in English. Although the study has its own limitations, it managed to investigate one of the most debated aspects of the paired mode: does it matter from the point of view of candidates' measured language ability whether their partners' proficiency level is similar or different? The answer to this question, based on the findings of the quantitative study described in this paper, seems to be NO. The paired exam is an innovative feature of the new school leaving exam. It is a valid and reliable measure of the candidates' oral abilities.

Notes

1 This paper is the written version of the talk given at the 11th IATEFL-Hungary Conference at Nyíregyháza, October 7, 2001. **(back)**

Acknowledgements

I would like to thank Prof. Charles Alderson for his invaluable comments on the design of this study. Thanks are due to Brian North for giving me permission to use the speaking self-assessment scale he calibrated in relation to the Council of Europe Framework. Finally, I would like to thank colleagues of the Hungarian Examination Reform Project for their assistance in scoring the audio-taped performances and the project management for giving me permission to use the Intermediate Speaking Assessment Scale developed by the project.

References

Alderson, J. C. (2000). Exploding myths: Does the number of hours per week matter? *novELTy, 7* (1), 17-33.
Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53,* 36-41.
Katona, L. (1996). Do's and dont's: recommendations for oral examiners of foreign languages. *novELTy, 3* (2), 21-36.
Kormos, J. (1999). Simulating conversations in oral proficiency assessment: a conversation analysis of role-play and non-scripted interviews in language exams. *Language Testing 16*, 163-188.
Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing 13*, 151-172.
Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking*. Unpublished Licentiate Thesis. Jyväskylä: University of Jyväskylä.
McNamara, T. (1996). *Measuring second language performance.* Harlow: Longman.
Nikolov, M. (1999). Classroom observation project. In H. Fekete, É. Major, & M. Nikolov (Eds.), *English language education in Hungary*. Budapest: The British Council Hungary.
North, B. (1996). *The development of a common framework scale of language proficiency based on a theory of measurement*. Unpublished Ph.D. Thesis. Thames Valley University.
O'Sullivan, B. (2000). *Towards a model of performance in oral language testing*. Unpublished Ph.D. Reading: The University of Reading.
Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition 14*, 159-176.
Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal 53*, 42-51.
Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.
Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly23*, 489- 508.
Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition14*, 403-424.
Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning 45*, 3-42.

APPENDIX 1. Beszédkészség önértékelés: Részletezõ skála

Jelöld be x-szel azokat az állításokat, amelyeket angolul meg tudsz valósítani, azaz tükrözik angol nyelvi tudásodat.

ANNYI ÁLLÍTÁST JELÖLHETSZ BE, AMENNYIT IGAZNAK TARTASZ.

1	Tudom kezelni az egyszerû számokat, mennyiségeket, árakat és az idõt.
2	Fel tudok tenni és meg tudok válaszolni egyszerû egyenes kérdéseket magamról és más emberekrõl, és arról, hogy hol élek, hogy milyen embereket ismerek és azokról a dolgokról, amelyek az enyémek.
3	Ha megkérdeznek, tudok válaszolni olyan egyszerû egyenes kérdésekre, amelyeket nagyon lassan és tisztán mondanak, és a velem kapcsolatos dolgokra vonatkoznak.
4	Tudok vásárolni egyszerû helyzetekben, ahol rá tudok mutatni a megvásárolandó tárgyra vagy gesztikuláció segítségével ki tudom fejezni magam.
5	El tudok kérni emberektõl dolgokat és megértem, ha tõlem kérnek el dolgokat.
6	Be tudom mutatni magam és tudok alapvetõ köszönési és búcsúzkodási formulákat használni.

7	Fel tudok tenni és meg tudok válaszolni egyszerû kérdéseket. Meg tudok fogalmazni és válaszolni egyszerû állításokat olyan témákban, amelyek közvetlenül érintenek vagy nagyon ismerõsek a számomra.
8	Tudom üdvözölni az embereket, meg tudom kérdezni tõlük, hogy hogy vannak és tudok válaszolni egyszerû dolgokra, amiket mondanak nekem.
9	Egyszerû formában meg tudom beszélni, hogy mit csináljunk, hova menjünk és hogy találkozzunk.
10	Részt tudok venni rövid társalgásban, de ritkán sikerül mindent olyan jól megértenem, hogy a társalgást hosszabb ideig is fenntartsam.
11	Egyszerû formában meg tudok vitatni hétköznapi gyakorlati kérdéseket, amikor a beszélgetõpartnere(i)m tisztán és lassan beszél(nek).
12	Egyszerû formában el tudom mondani, hogy mit gondolok dolgokról, amikor hivatalos közegben egyenesen engem kérdeznek, és ha szükséges, kérhetem, hogy ismételjék meg a fontosabb pontokat.
13	Meg tudom kérdezni, hogy az emberek mivel foglalkoznak a munkahelyükön és mit csinálnak a szabad idejükben.
14	Tudok meghívni embereket és bocsánatot kérni tõlük, és tudok meghívásokra és bocsánatkérésekre válaszolni.
15	Tudok javaslatokat tenni és válaszolni javaslatokra.
16	Ki tudok fejezni röviden olyan érzelmeket mint meglepõdés, boldogság, szomorúság, érdeklõdés és közömbösség, illetve tudok ezekre az érzelmekre válaszolni is.
17	Folyamatosan részt tudok venni a társalgásban vagy vitában, de néha lehet, hogy nehéz követni azt, amit mondok, amikor pontosan akarom megfogalmazni azt, amit gondolok.
18	Tudok kezdeményezni, lefolytatni és lezárni beszélgetést, ami olyan dolgokról szól, amelyeket jól ismerek vagy személyesen érdekelnek.
19	Részt tudok venni olyan beszélgetésekben, ahol ismerõs témáról van szó, mégha nem is volt idõm arra, hogy felkészüljek, hogy mit is fogok mondani.
20	Elboldogulok majdnem minden olyan szituációban, ami utazási irodán keresztül történõ utazásszervezéshez vagy magához az utazáshoz kapcsolódik.

21	Elboldogulok a tömegközlekedéshez kapcsolódó kevésbé megszokott helyzetekben, mint például ha egy utast meg kell kérdezni, hogy hol kell ahhoz leszállni, hogy egy ismeretlen célállomáshoz eljussak.
22	Amikor nem jut eszembe egy szó, tudok helyette más, egyszerû, de hasonló jelentésû szót használni vagy meg tudom kérdezni, hogy mi lenne a helyes szó, amire szükségem van.
23	Röviden el tudom magyarázni és meg tudom indokolni a terveimet, szándékaimat és tetteimet.
24	El tudok mesélni saját szemszögömbõl nézve különbözõ élményeket, némiképp részletezve a hozzájuk fûzõdõ érzelmeimet és a reakcióimat.
25	Tudok anyanyelvi beszélõvel beszélgetni anélkül, hogy szórakoztat-nám vagy bosszantanám õt, ha ez nem áll szándékomban, illetve nem kényszerítem õt arra, hogy másképp viselkedjen mint amikor egy másik anyanyelvi beszélõvel társalog.
26	A nyelvet természetesen, folyékonyan és hatékonyan beszélem ismerõs szituációkban.
27	Vita közben ki tudom fejteni és alátámasztani a véleményemet olyan dolgokról, amelyek számomra ismerõsek, azáltal, hogy odaillõ magyarázatokkal, érvekkel és megjegyzésekkel szolgálok.
28	Részt tudok venni hosszabb beszélgetésekben, amelyek az általános érdeklõdési körbe beletartoznak.
29	Le tudom írni vagy definiálni az olyan konkrét dolgokat, amelyek neve nem jut eszembe. Például “parkolójegy”, “kreditkártya”, vagy “biztosítási rendszer”.
30	El tudok magyarázni aktuális témához kapcsolódó nézõpontot úgy, hogy különbözõ nézetek elõnyeit és hátrányait is részletezem.
31	Ismerem a nyelvet olyan jól, hogy megoldást tudjak találni vitás helyzetekben (például: érdemtelenül kapott közlekedési büntetõcédula, lakásban okozott kárért való pénzügyi felelõsség, felelõsség balesetokozásért).
32	Ki tudom magam fejezni folyékonyan és spontán módon, szinte erõlködés nélkül, különbözõ témák igen széles körében.
33	Tudom a nyelvet rugalmasan és hatékonyan használni a társas érintkezésben úgy, hogy a nyelvhasználatom érzelemkifejezésre és viccelõdésre is kiterjed, és jól ismert irodalmi vagy más jellegû forrásokra vonatkozó utalásokat is tartalmaz.

34	Az elképzeléseimet és véleményemet világosan és pontosan meg tudom fogalmazni, és meggyõzõen tudok elõvezetni összetett érveléseket illetve meggyõzõen tudok reagálni azokra.
35	Széles a szókincsem, aminek a segítségével bármilyen nehézséget át tudok hidalni, ami számomra esetleg ismeretlen szó miatt adódik.
36	Folyamatosan részt tudok venni vitában, mégha az olyan elvont vagy bonyolult dologról szól is, amely számomra nem túl ismerõs.
37	Olyan gördülékenyen tudok visszatérni és újrafogalmazni pontokat, hogy a beszélgetõpartnere(i)m alig veszi(k) észre.
38	Tudok világosan és gördülékenyen fogalmazni úgy, hogy a mondandóm hatásosan és logikusan van felépítve.
39	Magabiztosan tudok beszélni bonyolult témáról olyan hallgatók elõtt, akiknek a téma nem ismerõs. Úgy építem fel azt, amit mondani akarok, hogy a hallgatóim igényeihez rugalmasan illeszkedjen.
40	Minden nehézség nélkül, helyesen beszélem a nyelvet, és a nyelvtudásom egyáltalán nem akadályoz abban, hogy a társas érintkezésben és a magánéletben céljaimat elérjem.
41	Majdnem olyan jól beszélem a nyelvet mint a saját anyanyelvemet.
42	Úgy tudom kifejezni magam, hogy még anyanyelvi beszélõk is azt gondolják, hogy bizonyára hosszabb ideig éltem az adott idegen nyelvi környezetben.

Megjegyzés:

A fenti állítások az Európa Tanács következõ szintjeire vonatkoznak:

1 - 7 - A1
8 - 15 - A2
16 - 24 - B1
25 - 30 - B2
31 - 36 - C1
37 - 42 - C2
(translated from North, 1996 and reprinted with his permission)

APPENDIX 2 Cloze tests

Complete the following text about *language dialects* by writing the missing words on the lines provided (1-19). Use ONLY ONE WORD in each gap.

The fact that English has been spoken in England (1)_____ 1,500 years but in Australia for only 200, explains (2)_____ we have such a great wealth of regional dialects (3)_____ England that is more or less totally lacking in (4)_____. There has not been enough time for

changes over (5)_____ last 200 years to bring about much regional variation, it (6)_____ almost impossible to tell where someone comes from (7)_____ all, although very small differences are now beginning to (8)_____. It is very unlikely, however, that there will ever (9)_____ as much dialectal variation in Australia as there is (10)_____ England. This is because modern transport and communication conditions (11)_____ very much different from what they were 1,500 years (12)_____ even 100 years ago. Even though English is now (13)_____ in many different parts of the world many thousands (14)_____ miles apart, it is very unlikely that English will (15)_____ break up into a number of different non-intelligible languages (16)_____ the same way that Indo-European and Germanic did. German (17)_____ Norwegian became different languages because the ancestors of the (18)_____ of these two languages moved apart geographically, and were (19)_____ longer in touch and communicating with one another.

TOTAL: 19 points / Student's score:_____

Complete the following text about *revision* by writing the missing words on the lines provided (1-19). Use ONLY ONE WORD in each gap.

In an ideal world revision would be part of (1)_____ learning. At the end of each lesson you would (2)_____ a few minutes reading over what you had just (3)_____, thinking about it, and perhaps discussing it with your (4)_____. But it is far more likely that you escape (5)_____ the classroom with a sigh of relief, and simply (6)_____ home.

The importance of small doses of regular revision (7)_____ stressed by teachers, who may round off lessons by (8)_____ up what they have said, test you regularly on (9)_____ you know, or distribute copies of the key points (10)_____ each lesson or topic.

The art of summarising, which (11)_____ usually known as `taking notes', is very important when (12)_____ comes to revision. The best notes are those that (13)_____ take yourself because they are the ones most likely (14)_____ be related to the way you see things and (15)_____ sense of them. A helpful technique for subjects that (16)_____ hard learning is to note down things you really (17)_____ know on separate cards, which you may keep in (18)_____ pocket to learn at odd moments. Vocabulary, for example, (19)_____ easily be learned on the bus to school.

TOTAL: 19 points / Student's score:_____

Appendix 3 Intermediate level speaking assessment scale

	Communicative impact	**Grammar and coherence**	**Vocabulary**	**Sounds, stress, intonation**
	Candidate…	Candidate…	Candidate's vocabulary…	Candidate
7	makes entirely natural hesitations when searching for ideas, requires no additional prompting,	uses wide range of structures, uses mostly accurate grammar, makes coherent	has wide range, is fully appropriate	is understood with ease, uses mostly accurate and appropriate sounds and stress, uses a wide range of intonation to convey

	makes relevant contributions	contributions		meaning effectively
6				
5	makes hesitations when searching for language, requires no additional prompting, generally makes relevant contributions	uses adequate range of structures, makes frequent minor mistakes only, makes mostly coherent contributions	has adequate range, is generally appropriate with isolated inappropriacies	is understood easily with isolated difficulties, makes mistakes in sounds and stress which occasionally affect comprehensibility, uses an adequate range of intonation to convey meaning mostly effectively
4				
3	makes frequent hesitations, requires some additional prompting, occasionally makes irrelevant contributions	uses limited range of structures, makes occasional major and frequent minor mistakes, makes contributions with limited coherence	has limited range, is generally appropriate with occasional disturbing inappropriacies	is understood with some strain, makes mistakes in sounds and stress which seriously affect comprehensibility, uses a limited range of intonation to convey meaning
2				
1	makes long intrusive hesitations, requires major additional prompting, makes irrelevant contributions	uses very limited range of structures, makes frequent major and minor mistakes, makes mainly incoherent contributions	has very limited range, is frequently inappropriate	is understood with constant strain, mostly uses sounds and stress that are difficult to understand, makes little use of intonation to convey meaning
0	no assessable language	no assessable language	no assessable language	no assessable language

(reprinted with the permission of the Hungarian Examination Reform Project)

Csépes Ildikó was a member of the Hungarian Examination Reform Project and also worked in the team commissioned by KÁOKSZI to produce the Test Specifications for the new Matriculation Examination in English. She works at the Institute of English and American Studies of Debrecen University.