On the Validity of the Group Oral Test: A Correlation Experiment

By
Murtadha J. Bakir
And
Falih Al-Emara

Abstract

The testing of speaking skill has constituted a problematic area in most foreign language programs. The Oral Interview technique, which is the most widely used one, has proved its superior validity in comparison with the other techniques of testing this skill. However, its main weakness is its low practicality which lies in the fact that the technique is time-consuming. Therefore, a serious need for alternative techniques emerges; techniques which could claim an equal degree of validity and at the same time overcome the practical weakness of the oral interview. In this connection, the Group Oral Test (a rather novel technique first suggested by Folland and Robertson 1976) looks like a good candidate for such an alternative. In addition to its function as an appropriate motivational device that encourages learners to participate in spontaneous and creative discourse, it is more practical than the oral interview.

In this connection, this paper sets to establish the concurrent validity of the group oral test against the oral interview through empirically testing the two techniques against each other. It is conducted by correlating the scores on the output of thirty advanced learners of English on both tests, simultaneously assessed by two independent skilled raters.

The results of the experiment show that the group oral test is concurrently valid with the oral interview, as revealed by the highly positive correlation coefficients ranging up to .88, .87 and .83. Unintentionally, the results proved a high degree of rater's reliability in the measurement. As for practicality, it is revealed that the group oral test is more practical than the oral interview. Timewise, it took about only one fourth of the time that the oral interview required for testing the same group. These attributes recommend the use of the group oral test as a reliable and practical supplement or alternative to the oral interview.

On the Validity of the Group Oral Test: A Correlation Experiment

1.0 Introduction

The testing of speaking skill in foreign language learning has always constituted a problem for both teachers of foreign language and testing specialists. Speaking skill refers here to the "ability to communicate informally on everyday subjects with sufficient ease and fluency" (Harris, 1969: 82). The recent utilization of modern technological devices, such as the audio and video recording, has done much in the way of improving our testing techniques in terms of validity and reliability. However, it is recognized that this skill does not lend itself easily to the professed objective techniques either. The credibility of those objective techniques has been put to question and generally perceived inadequate for testing the speaking skill (Harris, 1969: 85-89 and Mullen, 1978: 302). The assessment of this skill is commonly carried out on individual basis by testing each learner separately and assessing his output manually by the examiner.

Taking all these into consideration, the Oral Interview Test (OIT) is still considered the most valid and reliable of all techniques used for testing this skill, despite all its obvious weaknesses (Clark, 1972: 42 and Wilds, 1975: 39). This has given rise to widespread dissatisfaction and to the search for alternatives that can claim an equal degree of reliability and validity. One of such alternatives has been the Group Oral Test (GOT), first suggested by Folland and Robertson (1976) which retains the same merits of the OIT and overcomes its drawbacks.

This paper reports on a rather straightforward experiment that was conducted to establish the concurrent validity of the GOT against the OIT. The aim is to see whether this technique meets the criterion of validity to the extent that it could supplement or replace the OIT. Later on in this paper, the practicality of this technique will be discussed.

2.0. The Procedure

In order to run a correlation experiment of the two test techniques (GOT and OIT), thirty participants were recruited to take both tests. They were all fourth (final) year students in the English Department, University of Basra. The tests were taken as discussed below.

2.1. The GOT: In this test, the thirty subjects were divided into five groups of six subjects each. The authors acted as examiners. The sessions were arranged so as to resemble normal situations where students engage in the discussion of their life concerns to reduce, as much as possible, the unnaturalness and tension of the test atmosphere. The materials for test discussions were brief dialogues and statements about everyday issues recorded on tape by a native speaker. The recordings lasted for less than two minutes each, and a separate one was used for testing each group so as to prevent test compromise. The members of each group

alternated in commenting on what they heard and in discussing their own views and the views of others.

The discussion lasted long enough to allow the examiners decide the score of each subject. The examiners' role was restricted to ensuring that everything ran smoothly and to very few cases of interference when they felt that the flow of conversation was hindered. The testing sessions lasted between 20-23 minutes each, including the listening to the recorded materials.

2.2. The OIT: An OIT was carried out for the same subjects two weeks later. In this test each subject was individually interviewed by the same examiners. A number of question sets were prepared in advance for this purpose. The questions, which the examiners alternated in putting to the subjects, were drawn from their own experience and everyday life topics and affairs. Two scores were separately given, one by each examiner, in assessing the subjects' performance, using the same scoring system and assessment chart used for assessing the GOT. The interviews took between 15-20 minutes each, allowing enough time to adequately rate each interviewee.

2.3. The Scoring system: Evaluation of the subjects' performance was based on judgements on the mastery of the speaking skill. Excessive subjectivity was lessened by dividing the skill into five components (pronunciation, vocabulary, grammar, fluency and comprehension), and allocating a special score to each of these components in each test. Each participant was tested and evaluated for each of these five components. A chart was designed for this purpose which contained five scoring cells for each component. The scores stretched between 5-1 representing the various levels of mastery of these components; the highest score was five and the lowest 1. The reader is referred to the appendix for details about the scoring system.

The scores which were entered in the analysis were the total of the five components for each subject (These will be labelled "scores" hereafter). Thus, each subject would have two scores each representing the total mark he/she got in each of the two tests.

3. The Results

In order to determine whether the two tests yielded similar results, the *mean* and *standard deviation* of the scores given by each examiner in both tests were compared. Table 1 below shows that the mean and standard deviation of the scores are very close for both examiners, and for the overall results either.

Table 1 Mean and Standard Deviation of the GOT and OIT

	Mean		Standard Deviation	
Test	GOT	OIT	GOT	OIT
Examiner A	16.4	16.6	2.77	3.34
Examiner B	17.56	17.66	2.57	2.76
Overall *X* and *SD*	16.98	17.13	2.67	3.06

It is evident from Table 1 that the two tests yielded almost identical results.

The nature of the relationship between the GOT and the OIT results was examined by the use of Pearson Product Moment Correlation ® (Adopted from Guilford and Fruchter 1978: 83).

Five ®'s were obtained as follows:

Examiner A	GOT vs OIT	® = 0.88
Examiner B	GOT vs OIT	® = 0.87
Examiner A vs	Examiner B GOT	® = 0.83
Examiner A vs	Examiner B OIT	® = 0.83
Examiners A & B	GOT vs OIT	® = 0.83

All the correlation coefficients were highly significant at $p<.001$. This means again that the ratings of examiner A and B for GOT and OIT are highly related.

The most striking finding here is that the GOT and the OIT scores were so positively highly related that they could be treated as identical. In addition, the comparison of the scores of the two tests given by each examiner and also the overall scores of the two tests show high validity coefficients which range to 0.88, 0.87, 83 respectively. Hence, it can be safely claimed that the GOT is concurrently valid with the OIT.

4. Conclusion

The experiment has revealed that the GOT fares well on the test of concurrent validity with the OIT which is the most valid and reliable technique for testing the speaking skill so far. On the other hand, the preference of the GOT over the OIT for practical considerations is quite obvious. Time wise, in the experiment, it took only about one fourth of the time that the OIT required. Thus, the GOT comes as a relief to those teachers who cannot devote as much time in testing as required by the OIT. Furthermore, the human input needed for the running of the GOT is less than that required for the OIT. Accordingly, if the results of this experiment can be of any indication, it is that of suggesting the high rater's reliability of the GOT. Thus, one examiner would be sufficient to run the test as the scores of the GOT given by the two examiners were nearly identical. So, it is recommended that the GOT is to be used as a supplement to or replacement of the other techniques of questioned validity and reliability or the ones that suffer practicality weaknesses such as the OIT.

References

Clark, John. 1972. *Foreign Language Teaching: Theory and Practice*. Philadeliphia: Centre for Curriculum Development.

Folland, David and Robertson, David. 1976. “Towards Objectivity in Group Oral Testing”, in *English Language Teaching Journal, vol. xxx.*

Guilford, J. P. and Fruchter,---- 1978. *Fundamental Statistics in Psychology and Education*. McGraw Hill.

Harris, David. 1969. *Testing English as a Second Language*. New York: McGraw Hill.

Mullen, Karen. 1978. “Direct Evaluation of Second Language Proficiency: The Effect of Rater and Scale in Oral Interviews”, in *Language Learning, Vol..28 (2).*

Wilds, Claudia. 1975. “The Oral Interview Test” in Randall Jones and Bernard Spolsky (eds). *Testing Language Proficiency*. Arlington, Virginia: Centre for Applied Linguistics.

Appendix I

Sample Oral-English rating Scale: Behavioural Statements and Their Numerical Values for Measuring the Speaking Skill of English Learners

Pronunciation

5. Has few traces of foreign accent.
4. Always intelligible, though one is conscious of a definite accent.
3. Pronunciation problems necessitate concentrated listening and occasionally lead to misunderstanding.
2. Very hard to understand because of pronunciation problems. Must frequently be asked to repeat.
1. Pronunciation problems so severe as to make speech virtually unintelligible.

Grammar

5. Makes few (if any) noticeable errors of grammar or word order.
4. Occasionally makes grammatical and/or word order errors which do not however obscure meaning.
3. Makes frequent errors of grammar and word order which occasionally obscure meaning.
2. Grammar and word order errors make comprehension difficult. Must often rephrase sentences and/or restrict himself to basic patterns.
1. Errors in grammar and word order so severe as to make speech virtually unintelligible.

Vocabulary

5. Use of vocabulary and idioms is virtually that of a native speaker.
4. Sometimes uses inappropriate terms and/or must rephrase ideas because of lexical inadequacies.
3. Frequently uses the wrong words; conversation somewhat limited because of inadequate vocabulary.
2. Misuse of words and very limited vocabulary make comprehension quite difficult.
1. Vocabulary limitations so extreme as to make conversation virtually impossible.

Fluency

5. Speech as fluent and effortless as that of a native speaker.
4. Speed of speech seems to be slightly affected by language problems.
3. Speed and fluency are rather strongly affected by language problems.
2. Usually hesitant; often forced into silence by language limitations.
1. Speech is so halting and fragmentary as to make conversation virtually impossible.

Comprehension

5. Appears to understand everything without difficulty.
4. Understands nearly everything at normal speed, although occasional repetition may be necessary.
3. Understands most of what is said at slower-than-normal speed without repetitions.

2. Has great difficulty following what is said. Can comprehend only “social conversation” spoken slowly and with frequent repetitions.
1. Cannot be said to understand even simple conversational English.

Appendix II

Oral-English Assessment Chart

Component	5	4	3	2	1
Pronunciation					
Grammar					
Vocabulary					
Fluency					
Comprehension					

ملخص باللغة العربية

طالما شكلت عملية إختبار مهارة الكلام مسألة إشكالية في معظم برامج تدريس اللغات الأجنبية. وقد أثبت اختبار المقابلة الشفوية، وهو الإختبار الأكثر إستعمالا في هذا المجال، مصداقية رفيعة مقارنة بالإختبارات الاخرى لهذه المهارة. إلا أن نقطة الضعف الأساسية في هذا الإختبار تكمن في التطبيق العملي، إذ أن فيه مضيعة كبيرة للوقت، ولذلك تبرز الحاجة الماسة للبحث عن بدائل تقدر على تقديم مستوى مساوي من المصداقية وتتغلب في الوقت ذاته على نقطة الضعف التطبيقية. ويبدو أن الإختبار الشفوي المجموعاتي الذي إقترحه فولاند وروبرتسون عام 1976 يمثل بديلا مرشحا بقوة في هذا السياق. فبالإضافة إلى الوظيفة التحفيزية التي يمتلكها في تشجيع المتعلمين على المشاركة في الحديث بصورة عفوية وخلاقة، فهو أكثر عملية من المقابلة الشفهية من حيث الوقت المطلوب لإجرائه.

يحاول هذا البحث، إنطلاقاً من هذا السياق، إثبات المصداقية التزامنية للاختبار الشفوي المجموعاتي مقارنة بالمقابلة الشفوية من خلال إجراء إختبار عملي للإثنين معاً. ويجري ذلك عن طريق كشف الترابط بين الدرجات التي حصل عليها ثلاثون مشاركا من المتعلمين في المرحلة الجامعية المتقدمة في اللغة الإنجليزية في الإختبارين، وقام بالتقويم اثنين من المقومين كلا على إنفراد وفي الوقت نفسه.

وقد أظهرت نتائج التجربة أن الاختبار الشفوي المجموعاتي يتمتع بمصداقية تزامنية عالية مع المقابلة الشفوية، كما يتضح ذلك من معاملات الارتباط الموجبة العالية التي تصل الى 0.88، و0.87 و0.83. كما أثبتت النتائج عرضياً وجود درجة عالية من الموثوقية في المقياس. أما بالنسبة للتطبيق العملي، فقد تبين أن اختبار المجموعات هو أكثر واقعية من المقابلة الشفهية، إذ استغرق فقط نحو ربع الوقت الذي استغرقته المقابلة الشفوية لإختبار المجموعة نفسها. هذه الصفات تشجع على التوصية باستخدام الاختبار الشفوي المجموعاتي مكملاً أو بديلاً موثوقا به عن المقابلة الشفوية ومتفوقاً عليها من الناحية العملية التطبيقيةمن حيث الإقتصاد بالوقت ومن حيث القدرة التحفيزية على الاداء.