Washback of an oral assessment system in the EFL classroom

Ana P. Muñoz
Universidad EAFIT, Colombia

Marta E. Álvarez
Universidad EAFIT, Colombia

Abstract
This article reports the results of a research study to determine the washback effect of an oral assessment system on some areas of the teaching and learning of English as a Foreign Language (EFL). The research combined quantitative and qualitative research methods within a comparative study between an experimental group and a comparison group. Fourteen EFL teachers and 110 college students participated in the study. Data were collected by means of teacher and student surveys, class observations, and external evaluations of students’ oral performance. The data were analyzed using descriptive statistics for qualitative information and inferential statistics to compare the mean scores of the two groups by One Way Anova. Results showed positive washback in some of the areas examined. The implications for the classroom are that constant guidance and support over time are essential in order to help teachers use the system appropriately and therefore create positive washback.

Keywords
classroom-based assessment, intended washback, positive washback, score gains, student progress, teacher change

Research literature abounds with studies concerning the influence that testing has on teaching and learning, or what has commonly been referred to as washback (Alderson & Wall, 1993). The majority of washback studies have focused on the positive or negative consequences of standardized tests on different areas of the curriculum. In general, these tests are considered high-stakes, that is, they are used for making important educational and professional decisions, such as admissions, graduation, employment, or promotions, and therefore affect people’s futures. However, there has been little investigation into the effect of classroom-based assessment on instructional and learning practices. Classroom assessment may be differentiated from standardized testing in the degree of formality...
and the purposes for which they are used. It has been argued that high-stakes tests have more power to modify teacher and learner behavior whereas low-stakes tests, such as classroom-based assessment, are not central to decision-making and therefore have fewer consequences. In any case, it is widely recognized that any type of assessment can act as a lever for educational change.

Aware of the influence that assessment may exert on classroom behavior, a group of researchers at a language center in a small private university in Colombia, South America, set out to investigate the effect of a classroom-based assessment system on instruction. The assessment system was developed in 2001 with the aim of improving the teaching and learning of oral language at the institution (Muñoz et al., 2003).

The components of the Oral Assessment System (OAS) are: 1) a set of rubrics for different proficiency levels, each of which provides a scoring scale from one (1.0) to five (5.0) along with performance descriptors for the different aspects of oral language to be assessed, namely communicative effectiveness, pronunciation, grammar, vocabulary, and task completion; 2) speaking standards for each course; 3) a set of suggested assessment tasks per course; 4) a scoring sheet for teachers to record assessment tasks and grades; 5) a report card that students receive during mid-term and final assessment feedback sessions; and 6) ‘Oral Assessment Guidelines,’ a document where oral assessment criteria and procedures for the Language Center are specified (Muñoz et al., 2004).

The principles behind OAS suggest that assessment should be done through:

1. Clear delineated assessment criteria that are understandable to teachers and students.
2. A variety of tasks aiming at different learning styles.
3. Authentic and meaningful tasks
4. Different grouping techniques to elicit interaction among the students and with the teacher.
5. Encouragement of self assessment.
6. Tasks derived from curriculum objectives that are consistent with instructional practices.
7. Ongoing assessments so that students can demonstrate the extent of their knowledge and abilities.
8. Assessment of different aspects of oral language where grammar is only one of many different aspects considered in the assessment of communicative competence.
9. Detailed and specific feedback.

These principles reflect current theories about communicative language testing in which aspects such as congruity between assessment and curriculum related objectives, authenticity of tasks, detailed score reporting, teachers’ and students’ understanding of the assessment criteria, learner self-assessment, and similarity between instructional and assessment tasks need to be considered when designing tests in order to produce beneficial effects in the classroom (Bailey, 1996; Hughes, 1989; Messick, 1996; Morrow, 1979; Shohamy, 1992). These variables were carefully considered in the design of the OAS. First, the main components of the OAS – speaking standards per course, rubrics, and
assessment tasks – were explicitly connected to guarantee congruity between assessment and instructional goals. Second, the speaking tasks were designed following authenticity criteria. Authenticity was defined as a quality of outcome arising from the task itself and the processing of input by students. Therefore, aspects such as situational authenticity, and interaction between the assessment tasks and students’ background were taken into account (Bachman, 1991; Bachman & Palmer, 1996; Douglas, 2000; Widdowson, 1979). Third, the rubrics were designed to measure separate components of language along grade level descriptors that allow for detailed scoring. Fourth, the teachers received a 30-hour training course to familiarize them with the appropriate use of the system. The course included three modules dealing with the following: 1) definition of the construct oral language ability and its importance for making appropriate assessment inferences; 2) design of assessment tasks (types and formats); and 3) consensus on scoring criteria ( calibration sessions and use of rubrics for student self-assessment). Afterwards the OAS was implemented in the instructional routines of the Language Center.

The alignment of the OAS with the curriculum establishes a distinction between classroom assessment and high-stakes tests which is important to consider for the current study. According to Goldstein (1989, cited in Burrows, 2004), the difference between large scale testing and classroom assessment is ‘its deliberate attempt to avoid connection with particular learning environments’ (p. 140). On the contrary, classroom assessment is school-based and is directly concerned with progress and learning processes. Since the OAS was closely connected to instruction its potential effect on the classroom may be different from the effect of a high-stakes exam. In a high-stakes situation, ‘in which the results of mandated tests trigger rewards, sanctions or public scrutiny and loss of professional status, teachers will be motivated to pursue the objectives that the test embodies’ (Popham, 1987, cited in Cheng & Couture, 2000, p. 6). In assessment connected to instruction, teachers focus their attention on the content of the test. Assessment may therefore lead teachers to pursue the objectives established in the curriculum.

The OAS was expected to exert some positive effects on the classroom. The aim of this study was therefore to explore the possibility of creating positive washback by focusing on some of the principles underlying the OAS development.

I Literature review

Many countries in the world have introduced different types of assessments in their educational system with the intention of motivating changes in teaching and learning (Alderson & Wall, 1993; Burrows, 2004; Cheng, 2004; Qi, 2004). The introduction of assessments for such aims has been described by different terms by different authors. For Frederiksen and Collins (1989), it is systemic validity; for Messick (1996) consequential validity; and for Bachman and Palmer (1996) test impact. All these terms refer to diverse traits of washback, but for the purposes of this study, we will use the term washback in its wider sense, that is, effect.

The majority of washback-intended studies have concentrated on the positive or negative effects of high-stakes examinations on such areas as course content, teachers’ methodology, teacher and student attitudes, and learning. The research findings support
some evidence of washback; however it has been reported that such examinations have a greater effect upon content and less upon the actual methodology employed by teachers. For instance, Cheng (1997) analyzed the effect that changes to the Hong Kong Certificate of Education Examination (HKCEE) in English had upon teaching and learning. She found that washback occurred, but only in terms of bringing about change to teaching materials. Evidence that the changes to the test brought about changes to the way teachers taught was not conclusive. In another study, Qi (2004) investigated the effect of the National Matriculation English Test (NMET), which was introduced in order to promote positive washback effect on secondary school English teaching in China. Qi’s findings are that there was a big discrepancy between the test designers’ intentions and classroom practice and that little positive washback was felt on students’ language use. Wall and Alderson (1993) conducted a study investigating the possible positive and negative effects of an examination (the 0-level examination) on English language teaching. The test was introduced in secondary schools in Sri Lanka with the aim of providing a means for gearing teaching less toward traditional practices and more toward communicative classrooms. The results of the study demonstrated that while the exam had no influence on teaching style, it did have an effect on the content of teaching. In another study, Burrows (2004) examined the washback effect of a classroom-based assessment in the Australian Adult Migrant English Programme. She found that the test effect varied from teacher to teacher depending on the teacher’s beliefs and attitudes.

Furthermore, studies examining effects of assessment on student learning appear to illustrate a lack of clear understanding as to how washback works (Alderson & Wall, 1993; Bailey, 1999). Both negative and positive attitudes about the effect of tests have been reported in different studies (Alderson and Hamp-Lyons, 1996; Alderson & Wall, 1993; Cheng, 2004; Shohamy et al., 1996). However, there have been few studies reporting verifiable gains in student learning. Some research has addressed the effect of coaching for examinations such as the SAT (Becker, 1990; Johnson et al., 1985), and the TOEIC (Robb & Ercanbrack, 1999) on gain scores. Yet it is not clear if gains in these tests are an indication of language competence or a skill to take tests (Amrein & Berliner, 2002; Klein et al., 2000).

As can be seen from the above studies, merely instituting examinations is unlikely to bring about significant change. For meaningful educational change a host of elements beyond the examination itself needs to be considered. Following up on the Sri Lanka study, Wall (1996) reports on different factors which might have prevented the examination from providing positive effects: Teachers’ lack of understanding of the exam, resistance to change, and exam content. She also refers to other factors such as gap between test designers and teachers, and lack of well trained teachers.

Similarly, to overcome obstacles toward positive washback, many ELT specialists recommend improving examination systems by ensuring congruity between curriculum objectives and exams, authenticity of tasks, detailed score reporting, teachers and students’ increased understanding of the assessment criteria, and learner self-assessment (Bailey, 1996; Eisemon, 1990; Hughes, 1989; Messick, 1996; Shohamy, 1992). Other authors refer to meaningfulness of feedback (Bachman & Palmer, 1996; Black & Wiliam, 1998; Kulik & Kulik, 1991) and variety of test formats and tasks as powerful means for teaching and learning improvements (Kellaghan & Greany, 1992 in Wall, 1996; Messick, 1996).
Thus, there is a need for research that clearly establishes the connection between assessment and language development. Additionally, research needs to be done on the effects of classroom assessments on the depth and degree of student learning. The important role of classroom assessment in learning is highlighted by Glaser and Silver (1994, p. 26) when they state that ‘as assessment and instruction are more closely linked, achievement measurement will be integral to learning rather than imposed by some external examination on students’ fates.’

II  Research questions

The OAS was designed with the intention to promote positive washback in teaching and learning. This study intended to respond to the following research questions:

1) Do some of the principles that underlie the development of the OAS generate positive washback?
2) Is positive washback perceived by participants?
3) Do learners who participate in the OAS (the experimental group) make more progress than those who do not (the comparison group)?

III  Methods

In the examination of the research questions the principles that guided the investigation of washback were as follows:

Teaching

Congruence between curriculum objectives and instructional tasks  To investigate the connection between curriculum objectives and instructional tasks, teachers’ capacity to specify classroom objectives and conduct activities that directly targeted those objectives was examined. By stating lesson objectives, teachers make sure that students become conscious of and have a shared understanding of the proposed goals. Informing students of course and lesson objectives is important because students can focus their learning efforts toward specific targets. Multiple research studies show that students who perform better are those who know what the teacher intends to teach and how he or she intends to teach (Amigues & Guinard-Andreucci, 1981; Bonniol, 1981; Jorba & Sanmarti, 1994).

Variety of assessment tasks and task design  It has been widely recognized that a variety of assessment tasks benefits learning. The use of multiple tasks allows for more reliability of the assessment process and gives students more opportunities to show the extent of their knowledge. Likewise, for beneficial washback to exist, students need to be familiar with the format of the assessment so that the connection between assessment and educational goals may be established (Bailey, 1996; Messick, 1996). Moreover, students’ successful performance on assessment tasks greatly depends on how well teachers and test developers design those tasks. Therefore, teachers need to provide students with steps and clear instructions that indicate to students what they have to do and how they have to carry out the task.
Detailed and specific feedback  Black and Wiliam (1998) highlight the role of feedback in learning and elaborate on the ways feedback can be made effective for students’ learning. They claim that giving students only grades is not beneficial feedback. For feedback to impact learning, they say, it needs to provide each learner with specific guidance on strengths and weaknesses. Feedback is beneficial when it is given in a timely manner and in terms that students can understand. The best feedback is highly specific, or highly descriptive of what actually resulted, clear to the performer, and offered in terms of specific targets and standards (Wiggins, 1998). Such information can motivate students to improve if they are doing poorly or to maintain their performance if they are doing well. When feedback is provided in the form of grades, it needs to be detailed and meaningful. According to Shohamy (1992), to bring about change, exam results need to be translated into instructional activities and actual strategies for teaching and learning.

Learning

Understanding of the assessment criteria  Familiarity with evaluation and assessment helps students establish a direction for learning, become critical of their own progress, plan for assessment in order to perform better, and develop the ability to self-assess. In order for beneficial washback to be fully realized, according to Hughes (1989, p. 46), ‘the rationale for (a) test, its specifications, and sample items should be made available to everyone concerned with preparation for the test’.

Learner self-assessment  Self-assessment can be related to beneficial washback because it helps learners develop internal criteria for progress and success, and thus develop learner autonomy. In referring to autonomy as positive washback, Bailey (1999) points out that ‘ownership and self-regulation are thought to develop greater locus of control and deeper processing of the material at hand (van Lier, personal communication, 1994). Thus, the issue of learner autonomy and responsibility is directly related to Alderson and Wall’s (1993) 10th washback hypothesis: ‘A test will influence the degree and depth of learning’ (p. 120).’

1 Design

The study was designed to combine quantitative and qualitative research methods, comprising a comparative study between an experimental and a comparison group. The study covered 15 months, corresponding to 15 courses in the intensive schedule of the Adult English Program.

All the teachers used the OAS. Included was a set of rubrics proposed for beginner and intermediate students, but only the teachers of the experimental group received supplementary training on the use of the rubrics and on how to teach their students to use them for self-assessment. Additionally, they participated in periodical discussion meetings where proper assessment practices were reinforced. In contrast, assessment in the comparison group was largely left to the teachers who decided what, when, and how to assess, obviously following the communicative approach established by the Language Center.
2 Participants
Fourteen EFL teachers took part in the study, seven in the experimental group and seven in the comparison group. All the teachers were non-native speakers who had been teaching at the Language Center for more than three years. Four of these teachers had undergraduate degrees in teaching foreign languages from local universities. The other teachers had participated in regular in-service programs offered by the institution. The participating students were 110 EFL college students (beginner to high-intermediate), 55 in each group. Most of these students enroll in English classes as a requirement for graduation as stipulated by the university’s ‘bilingualism policy.’ New students were assigned to each group if any student dropped out, thus maintaining the sample size. Neither students nor teachers changed groups – from experimental to comparison and vice versa.

3 Data-gathering tools
a Students’ survey: 10 questions, including closed item and open-ended questions, to examine student perception regarding potential washback effects. Students were asked about their familiarity with classroom objectives and their congruence with assessment and instructional practices, the format of assessment tasks, their familiarity with assessment criteria and procedures, their self-assessment practices, the type of feedback received from teachers, and their perception of improved language areas.

b Teachers’ survey: aimed at exploring teaching and assessment practices. The survey included five closed items to ask teachers about specification of lesson objectives, the format of assessment tasks, the assessment and feedback procedures, and students’ involvement in self-assessment. Two open-ended questions were also included to ask teachers if they perceived any improvements in students’ oral skills and if they considered their teaching methodology had changed after the implementation of the OAS.

c Class observations: to match teachers and students’ perceptions of washback, as revealed by the surveys, with observers’ perception. Fifty-four observations – 29 of the experimental and 25 of the comparison group – were conducted by two academic coordinators from the Language Center. Observation sheets – filled out by the observers – required the following information: specification of class objectives, types of assessment and instructional tasks, feedback procedures, and self-assessment practices.

d External evaluations: (by a teacher trained in assessment) to determine score gains. Five hundred and fifty-four evaluations were conducted with 110 students who were tested at six different moments along the study: Two hundred and ninety-nine evaluations of the experimental group students and 255 of the comparison group. In order to establish the appropriate balance in the number of evaluations conducted along the observed period a test for independence between month of evaluation and group was conducted. Results ($\chi^2 = 4.96; p > 0.05$) indicated a desired balance. Evaluations comprised three parts which increased in difficulty according to the proficiency level. In part one, students had to provide factual information (name, age, family, hobbies). In part two they interacted with the interviewer either by asking questions (using prompts) or reaching agreement. In part three,
they were asked to describe a picture or photograph. They were evaluated using the corresponding rubrics for each level and received grades for communicative effectiveness, pronunciation, grammar, vocabulary, and task completion.

4 Data analysis

A comparative analysis was made between the experimental and comparison group students’ answers by a chi-square test of independence for each item of the survey. To analyze teachers’ opinion descriptive statistics were used. Class observations were analyzed by means of absolute and relative frequencies. The analysis of the external evaluations was conducted by examining both the total scores of the evaluations and the scores for each aspect of the rubrics. The analysis required descriptive statistics and inferential statistics to compare the mean scores of the two groups by One Way ANOVA and validation tests for the latter (fitting a normal distribution for total by groups and a variance check for similar variances of the groups).

IV Results

Results are presented for the aspects examined in teaching and learning:

Teaching

Specification of lesson objectives and their congruence with instructional tasks Teachers of the experimental group reported specifying lesson objectives by writing them down on the board at the beginning of the lesson. Sixty per cent of these teachers said that they also stated objectives orally. Teachers in the comparison group claimed to state objectives orally (49.5%), or to start the lesson by checking homework (50.5%). However, class observations showed that the examined category was highly independent of the groups ($\chi^2 = 27.64; \ p < 0.05$). That is, 24% of the teachers in the comparison group did not specify the objectives, but the tasks to be done, and 76% did not state the objectives or tasks. On the contrary, 86.2% of the teachers of the experimental group announced lesson objectives in both oral and written form, while 13.8% did not.

Upon examining the congruence between class objectives and instructional tasks, the observations revealed a significant difference between the groups ($\chi^2 = 20.49; \ p < 0.05$). While objectives and content in the experimental group classes were related (63%), there was little relationship between these two aspects in the comparison group classes (14.81%). Additionally, the analysis of students familiarity with assessment tasks, showed a significant differences between the groups ($\chi^2 = 8.76; \ p < 0.05$). Whereas teachers of the experimental group used assessment tasks which were similar (56.1%) rather than identical (4.9%) or apparently not related (39%) to instructional tasks, teachers of the comparison group seemed to use tasks which were completely new (47.8%) rather than similar (35.8%) to instructional tasks. The percentage of tasks used for assessment which had already been used for instruction was also significantly higher in the comparison group (16.42% as opposed to 4.9% in the experimental group, $\ p < 0.05$).
Variety of assessment tasks and task design

Teachers in both groups reported using: (1) role-plays, (2) giving opinions, and (3) presentations. Class observations showed that teachers in the experimental group also used narration and description tasks. Teachers in the comparison group focused exclusively on grammar-oriented activities, such as repetition or controlled question and answer where the focus was practicing a specific language form. There was no concern for using the language in a meaningful communicative context where the language used was not totally predictable. Other types of activities observed were word definition and memorized dialogues.

All the teachers in the experimental group agreed on following specific steps when designing tasks: (1) making sure the topics to be assessed had already been taught, (2) selecting the speaking standards to be assessed, (3) selecting the task, (4) selecting oral language aspects from the rubric, (5) deciding on the group arrangement, and (6) preparing the feedback forms. The majority of the steps that teachers in the comparison group said they followed were related to the explanation of how to carry out the task.

Feedback procedures

When describing feedback procedures, teachers in both groups claimed to give individual feedback. However, all the teachers followed a different routine during the sessions. An example of a feedback session may or may not include all the relevant materials, namely feedback form, rubrics, speaking standards, and score record sheets. However, score reporting in both groups was different ($\chi^2 = 59.21; p <0.05$). While teachers of the comparison group tended to assign a global grade, teachers of the experimental group broke it down into the different aspects, that is, communicative effectiveness, pronunciation, grammar, vocabulary, and task completion.

A final question was asked to the teachers of the experimental group concerning their perception of changes in the teaching and learning process due to the use of the OAS. All the teachers in this group claimed to have perceived positive changes in relation to: (1) improvement in students’ oral production, (2) lesson planning, and (3) planning for oral assessment. Additionally, they commented that the descriptors in the rubric had allowed them to measure more accurately students’ oral production and had helped them support their decisions. Another positive aspect that they mentioned was their awareness in defining and sharing class objectives with students.

Learning

Understanding of the assessment criteria Students in the experimental group seemed to be more informed about the oral aspects to be assessed than were students in the comparison group ($\chi^2 = 13.26; p <0.05$). The use teachers made of the corresponding rubrics for oral assessment became evident when the surveyed students listed the aspects evaluated through oral assessment activities. In both groups, students listed grammar and vocabulary, but only students of the experimental group listed aspects such as communicative effectiveness and task completion. As for pronunciation, students of both groups mentioned it, but students of the experimental group referred to it more often and named different aspects such as intonation and stress.
Learner self-assessment  A difference was found between the two groups related to the use that students made of the corresponding rubrics for self assessment purposes ($\chi^2 = 31.71; p < 0.05$). Students of the experimental group reported to use them significantly more (73.2%) than students of the comparison group (34.5%) as a tool for self-assessment. In contrast, students of the comparison group only mentioned discussion of the aspects to improve and feedback, but made no reference to the rubrics ($\chi^2 = 8.86; p < 0.05$). Teachers were asked whether they involved students in self-assessment practices. Eighty per cent of the teachers in both groups claimed to promote self-assessment. However, only the teachers of the experimental group reported that self-assessment was done using the rubrics. Teachers in the comparison group claimed to involve students in self-assessment through the discussion of aspects to improve as a whole class activity.

Score gains  Table 1 shows 95% confidence intervals for the score average of the two groups in the external evaluations. The mean score of the comparison group was 3.41 (for 255 evaluations) and for the experimental group was 3.53 (for 299 evaluations). The confidence intervals for the means show a minimal overlap between them (3.47–3.48). Additionally, a One Way ANOVA indicates that the mean score of the experimental group is greater than the mean of the comparison group ($t = 2.54; p = 0.006$). Although the mean scores for the two groups look very close, there is a significant difference between them because the sample size was large and the standard error was small.

These results were further analyzed taking into consideration the rubric and the students’ proficiency level. As mentioned above, the data was checked for normality to validate the results of the One Way Anova test. Since $p > 0.05$, the normality of the data was established ($\chi^2 = 8.03$ with 8 d.f.; $p = 0.43$). Likewise a variance check analysis (Table 2) was performed to further validate the ANOVA test. All p-values >0.05 indicating that variances are similar in the two groups.

**Table 1** Mean score gains by group

<table>
<thead>
<tr>
<th>Group</th>
<th>Evaluations</th>
<th>Mean</th>
<th>SD</th>
<th>95% confidence intervals</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>255</td>
<td>3.41</td>
<td>0.57</td>
<td>3.34–3.48</td>
<td>0.036</td>
</tr>
<tr>
<td>Experimental</td>
<td>299</td>
<td>3.53</td>
<td>0.54</td>
<td>3.47–3.59</td>
<td>0.031</td>
</tr>
</tbody>
</table>

**Table 2** Validation of the ANOVA test

<table>
<thead>
<tr>
<th>Test</th>
<th>Value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cochran</td>
<td>0.53</td>
<td>0.35</td>
</tr>
<tr>
<td>Bartlett</td>
<td>1.00</td>
<td>0.35</td>
</tr>
<tr>
<td>Levene</td>
<td>0.05</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Table 3 reports the results of a comparative analysis of the scores for each aspect of the rubric by groups was conducted in order to determine which aspect(s) contributed to establishing the differences found in the total scores between the groups.

As shown in Table 3, there were no significant differences between the groups on task completion and vocabulary. However, communicative effectiveness, grammar, and pronunciation had p-values <0.05 revealing higher gains in the experimental group. It is important to note that the students’ answers in the survey related to perception of gains referred to communicative effectiveness and pronunciation as the areas in which they felt they made most progress.

The last question of the students’ survey made reference to their perception of their own gains in the different aspects of oral language as they appear in the rubrics. No significant differences were found in relation to grammar and vocabulary. In contrast, for both communicative effectiveness ($\chi^2 = 7.68; p <0.05$) and pronunciation ($\chi^2 = 10.97; p <0.05$) more students of the experimental group felt that their progress was excellent.

V Discussion

This study examined the effect of the OAS on different areas underlying the development of the system. Teachers’ and students’ surveys revealed that teachers in both groups specified class objectives. However, observations showed that whereas the experimental group teachers stated objectives verbally and in writing, teachers of the comparison group wrote up the names of the activities on the board, but not the objectives. This may indicate a lack of clarity about the term ‘activity’ versus ‘objective’ or lack of knowledge on how to state objectives. Moreover, teachers of the comparison group conducted a series of activities in the classroom that did not have a clear connection with the objectives to be accomplished.
While goal setting is beneficial to students’ learning, carrying out activities aimed at the attainment of those goals is clearly more helpful. Failing to establish the objectives of the lesson may be detrimental to students because they do not see a clear direction for learning. Failing to develop activities related to lesson objectives creates confusion on students which will eventually impact negatively upon achievement. A plausible consequence of such failure might be students’ lack of attention leading to loss of motivation in learning. Establishing a clear connection among objectives and instruction can help learners feel more motivated to participate in class, be more aware of expected results, develop self-assessment strategies, detect their own progress or deficiencies, and establish a plan of action.

Washback was also perceived on the design of assessment tasks. Teachers of the experimental group structured assessment tasks mainly by considering content validity and management of the task. Task design is of utmost importance because students’ success during assessment largely depends on how well teachers structure activities. Establishing clear steps to develop a task implies determining which details and information teachers want the student to include during their learning process. This can certainly make the assessment more valid and reliable because it allows the teacher to elicit extended chunks of speech from students. ‘Requiring extended chunks of speech, with support from the inherent structure of a specific task, will give the student experience in being in charge of the speech situation and responsible for effective communication taking place’ (Brown & Yule, 1983, p. 118).

Class observations showed that assessment and teaching practices are limited to a few tasks. However, the use that teachers make of those tasks differs in the two groups. The experimental group students were assessed through tasks in which they faced new but analogous situations to instructional tasks. In other words, students were familiar with the assessment format, but the situations or contexts in which they were required to apply their knowledge and skills were new. In contrast, the students of the comparison group were assessed through tasks which were identical to instructional ones. A possible explanation might be the teachers’ lack of planning or the teachers’ deliberate interest in students’ success. When planning for assessment, teachers need to consider that tasks must not be completely new to the student or identical to those used for instruction because both provide for incorrect assessment. Using assessment tasks that require application of concepts has beneficial effects on learning. When teachers use tasks in which students need to apply their linguistic background and skills to novel situations, they can determine if students have a real understanding of the material presented and if they are able to synthesize concepts – aspects which are essential to authentic assessment and therefore beneficial to learning. On the other hand, when students are assessed through tasks that are identical to instructional tasks, students end up repeating memorized pieces of information (rote learning). Consequently, it will be difficult to determine whether they have understood the concepts taught. As highlighted by Wiggins (1998, p. 116), ‘understanding is best revealed when students are faced with new application situations.’

Although washback on feedback procedures could not be clearly established, score reporting presented differences between the groups. While teachers in the experimental group provided scores for each aspect of the rubric, teachers of the comparison group assigned a global score. In relation to score reporting, Shohamy (1992, p. 515) claims
that to promote positive washback, assessment information must be ‘detailed, innovative, relevant and diagnostic’ and that it must ‘address a variety of dimensions rather than being collapsed into one general score.’ Teachers in the experimental group provided students with information about their performance in terms of grades for communicative effectiveness, grammar, vocabulary, and task completion.

In general terms, we can say that students in the experimental group have a better understanding of the assessment criteria and procedures. They are more knowledgeable about oral language components in the rubric. In fact, they were able to name all aspects and specify important features. The fact that teachers provided students with separate scores for each aspect of the rubric allowed students not only to become familiar with the rubrics and the assessment criteria, but also to notice improvements in communicative effectiveness and pronunciation and, consequently, to obtain higher oral score gains. Possible explanations for this might be the emphasis the teachers give to these language components during instruction and assessment; the students’ awareness of the importance of these aspects for communication purposes; or their developed ability to measure aspects which were less tangible before the use of the rubrics.

Regarding self assessment, both teachers’ and students’ surveys indicated a significant difference between the groups in relation to the use that students made of the rubrics for this practice. The difference between the groups may be accounted for by the fact that teachers of the experimental group trained their students to use the rubrics for self-assessment. Self-assessment can be related to beneficial washback because it helps learners develop internal criteria for progress and success, and thus develop learner autonomy.

In relation to score gains, the experimental group had higher gains than the comparison group mainly in communicative effectiveness, grammar, and pronunciation. These results matched students’ perception of gains in communicative effectiveness and pronunciation. Improvements in oral language may be due to the emphasis teachers placed on assessing not only linguistic competence but communicative competence as a whole. The use that teachers of the experimental group made of the assessment system may have helped students become more conscious of assessment goals, criteria, and techniques, thus leading them to target their learning efforts to better performance.

Finally, teachers and students were also asked about their perception of the OAS. All the teachers of the experimental group perceived positive washback due to the use of the OAS. They observed improvements in their teaching and assessment practices and in students’ oral production. The constant guidance teachers received during the development of this study helped them make appropriate use of the system. They were able to reach a shared understanding of the system and convey it to their students. Students also perceived that their progress regarding language aspects was excellent especially for communicative effectiveness and pronunciation.

VI Limitations

A limitation of this study lies in the generalizability of the results. Although positive results were found the question arises to whether teachers in the experimental group performed in accordance to the requirements of the OAS because they were participating in a research study. It is not clear if the instructional practices found are the consequence
of teachers’ own conviction about the benefits of effective assessment procedures or are the product of reform initiatives by the Language Center.

The research literature suggests that beliefs and practice are inevitably related, and that teachers may hold beliefs that are not compatible with the practices called for in institutional plans (Bliem & Davinroy, 1997; Borko et al., 1997). Therefore, while a number of educators recommend change through improvements in assessment methods, it seems that this is not sufficient for creating meaningful change. In the case of change by means of new forms of assessment, it is possible that these forms present teachers with theoretical and practical conflicts, mainly the conflict between the underlying assumptions of assessment systems promoted in current assessment research and teachers’ pre-existing beliefs about instruction (Black & Wiliam, 1998; Duschl & Gitomer, 1997; Shepard, 2000). Therefore, for meaningful change to occur, teacher educators and reformers need to understand the beliefs that teachers bring to instructional practices. Based on this, professional development programs oriented toward reflection could be very beneficial to foster desired changes.

Additionally, this study is limited in that it relied on students’ and teachers’ self-reported data and the researchers had no opportunity to carry out follow-up interviews and therefore any conclusions drawn must take this into consideration. It does, however, show that valuable insights can be obtained about washback.

VII Conclusions

We could tentatively say that among the different aspects of teaching and learning observed a degree of positive washback occurred. This effect most likely took place because of the ongoing training on assessment practices provided to teachers. Constant guidance and support over time are, therefore, essential in order to help teachers understand and make appropriate use of the system. It is possible that the introduction of an assessment system generates effects that may cease if systematic training is not guaranteed. It is also imperative to understand the beliefs that teachers hold about assessment to identify possible conflict between institutional policies and teachers’ conceptions.

From the results of this study it can be concluded that washback may be fostered, first, by informing students of assessment procedures and scoring scales, specifying objectives, and structuring assessment tasks. When students are well-informed of such practices, they can focus their learning on specific goals. Learners will perform better and increase their language store if they understand the assessment system. Second, positive washback will be promoted when both teachers and students clearly establish the connection between educational goals and assessment. As stated by Buck (1988, p. 17) ‘most educators would probably agree that the content of classroom instruction should be decided on the basis of clearly understood educational goals, and examinations should try to ascertain whether these goals have been achieved. When the examination does that, it forces students and teachers to concentrate on these goals, and the washback effect on the classroom is very beneficial.’

Third, the use of self-assessment mechanisms fosters washback to the learners because they can take control of the assessment. Through self-assessment students become responsible for their own progress because they may be able to diagnose strong or weak areas, identify current proficiency level, and become more goal-oriented and consequently more self-directed learners.
Finally, the current study has yielded valuable information for curricular decisions. Washback on students and teachers will help the language program administrators to determine whether the Language Center as a whole performs well, whether the teaching methodology is effective, and whether the program goals are realistic. Based on such information, decisions can be made as to curriculum modifications.

**Note**

1 New forms of assessment or innovative forms refer to classroom assessments that use challenging tasks to elicit higher order thinking, and look not just at results but at developing learning processes (Shepard, 2000).

**VIII References**


