A Research Study on Task-Based Language Assessment

Leonardo Herrera Mosquera
Universidad Colombo-Americana
Bogotá, Colombia

Abstract
The methodology of Task-based teaching (TBT) has been positively regarded by many researchers and language teachers around the world. Yet, this language teaching methodology has been mainly implemented in English as a second language (ESL) class rooms and in English for specific purpose (ESP) courses; and more specifically with advanced-level learners. The present research study aimed at proving the feasibility of a TBT approach in a different learning context: A beginning Spanish class. That is to say, contrary to the traditional TBT implementation, the experiment was conducted in a foreign language class with students bearing a low level of language proficiency. The result of the research experiment was quite positive.

Key words: stask-based teaching (TBT), English as a second language (ESL), English for specific purpose (ESP), foreign language learning

Resumen
La metodología de la enseñanza basada en tareas ha sido considerada positivamente por muchos investigadores y profesores de lengua alrededor del mundo. Incluso, esta metodología de la enseñanza de la lengua ha sido utilizada principalmente para los cursos de inglés como segunda lengua (ESL) y para los de inglés con fines específicos (ESP) con estudiantes de nivel avanzado. El estudio realizado en la presente investigación trató de probar la viabilidad de este tipo de metodología en un contexto de aprendizaje diferente: un curso de español para principiantes. Es decir, contrario a la implementación tradicional de la metodología estudiada, el experimento fue realizado en una clase de lengua extranjera con estudiantes con un nivel bajo de habilidad en esta lengua. El resultado del experimento fue muy positivo.

Palabras claves: enseñanza basada en tareas, inglés como segunda lengua, inglés para fines específicos, aprendizaje de una lengua extranjera

Recepción: 26-10-11 Aceptación: 5-12-11

Literature Review

In recent decades, *tasks* have become an important methodological tool within the language teaching and learning process, especially in the field of Second Language Acquisition (SLA). But what do researchers and teachers mean by task? A frequently cited author on this issue is Long (1985), who offers a general definition that ranges from a non-verbal event such as painting a fence or dressing a child to a more communicative event such as making a hotel reservation or borrowing a library book. According to Long, almost any action people perform in their daily life can be called a task. This definition has been repeatedly used as a starting reference point by other theorists. Nunan (1989) depicts a more communication-oriented definition stating that a *communicative task* is "a piece of classroom work which involves learners in comprehending, manipulating, producing or interacting in their target language while their attention is primarily focused on meaning rather than form. The task should also have a sense of completeness, being able to stand alone as a communicative act in its own right" (p. 10). Following the idea of interaction and meaning orientation, Lee (2000) claims that a task requires a structured workplan. This author emphasizes two fundamental elements for the accomplishment of a task: there must be a focus on meaning, and interaction must be the means through which the objective is met. Other authors like Bygate, Skehan, and Swain (2001) define *tasks* as activities that require learners to use language, with emphasis on meaning, to attain an objective. Ellis (2003) defines *tasks* as activities that call for primarily meaning-focused language use. As we can see from these definitions, tasks call for a focus on meaning in such a way that they resemble linguistic events that occur in real-life circumstances. That was my main concern when designing assessment tasks for the study herein described. Even if the object of study was a grammatical component, I adapted it in such a way that the assessment task would resemble a real-world communicative event. Teaching methodologies that incorporate such tasks as central components of the curriculum are denominated *task-based instruction* (TBI) or *task-based teaching* (TBT). However, the present study focuses on just one portion of the TBT process: Assessment or task-based assessment (TBA). That is, tasks have been designed and implemented in this study as a means to assess students' linguistic and communicative skills and to offer appropriate feedback.

Task-based assessment (TBA) refers to "assessment that utilizes holistic tasks involving either real-world behavior (or as close as it is possible to get to this) or the kinds of language processing found in real-world activities" (Ellis, 2003, p. 285). Such classroom simulation of real-world behaviors corresponds to one of the dilemmas in the TBI literature and has consequently immersed linguists in a debate over the authenticity that TBI and TBA seek. Tasks involving these real-world actions seem to be more easily accomplished in contexts where the language is learned as a second language or in specific-purpose (ESP) curricula. In any other learning contexts, we may need to look at tasks in terms of thought processing. That is, classroom task performance might not mirror ex-

actly real-life events, yet such tasks require the thought and linguistic processing necessary for real-life communicative events. Some authors have designated those tasks that simulate real-world behavior as "performance-referenced" assessment and those tasks that do not involve real-world behavior as "system-referenced" assessment. System-referenced tests assess language proficiency in a general sense without reference to any particular use or situation (Baker, 1989). Listening to a radio advertisement and answering questions about it is an example of such a test. Performance-referenced tests, on the other hand, assess the ability to use language for specific purposes or in specific contexts. Ordering a meal at a restaurant, inquiring about the different dishes, and asking the server for suggestions can be an example of a performance-referenced test. Both types of tests are communicative in nature and can consequently be considered as good samples of TBA. McNamara (1996) argues that a communicative test needs to be both system-referenced and performance-referenced.

For the purpose of the present study, I designed communicative assessment tasks that combine elements of both types of assessment (system-based and performance-based). The themes, lexical items, and grammatical structures already stated in the Spanish 1 syllabus were incorporated into the design of the tasks in a way that better resembled real-world communicative events, also known as target-language use (TLU) tasks. I also want to add that most of the tasks designed for this study were integrative in that they integrate two or more language skills (listening and writing, reading and speaking, and the like) or two or more linguistic components (grammar, vocabulary, pronunciation).

Having described what linguists mean by task, task-based teaching, and task-based assessment, I will proceed to describe the methodology, results, and recommendations of this study.

Methodology

Research Design

As said earlier, the present experimental research sought to measure the impact of a task-based assessment approach in the learning of Spanish at a middle school level. The target population consisted of all Spanish 1 students in York County School Division in Virginia, and the accessible population consisted of four classes of Spanish 1 students at Tabb Middle School. These four classes were part of the teaching assignment of the author of this study during the 2009-2010 school year. Since the groups (accessible population) were already established, the experimental group design implemented is the static-group pretest-posttest.

An entrance test was administered to both the treatment and control groups, followed by a six-month treatment. An exit test was administered to both groups as well. Differences in performance scores between the treatment and the control groups demonstrate the effect of the treatment considering all the intervening variables and factors.

Sample

A sample of convenience was used in this study. Two out of the four Spanish 1 classes were randomly selected as the treatment group and the remaining two classes were designated as the control groups.

Composition of Groups

All four groups consisted of seventh and eighth grade students with ages ranging between 11 and 14 years. The characteristics to be analyzed in this study are group size, gender, grade level, and Spanish background for both the treatment and the control groups. The following table illustrates such features.

Table 1
Composition of groups

Groups	Group Size	Background in Spanish	Grade Level		Gender	
			7th	8th	Boys	Girls
A	22	5	11	11	10	12
B	23	3	14	9	14	9
C	18	8	3	15	8	10
D	18	2	9	9	9	9

Instrumentation

On the first day of class (September 2009), students filled out a form with information regarding personal data, their background in Spanish, and their motivation to learn Spanish. This data provided significant information for the interpretation of results. On the second day, a Spanish entrance test was administered to all four groups (accessible population). This test consisted of 30 questions of vocabulary and grammar. The questions for this test were taken from the textbook online self-tests (www.pasoapaso.com) for the first two chapters.

The treatment consisted of 10 task-based tests administered during the first semester, September 2009 to January 2010. These task-based tests were administered only to the two treatment groups. The two control groups were administered traditional quizzes, that is, multiple-choice, matching, and fill-in-the-blank type tests. The content for the task-based tests was dictated by the textbook syllabus and adapted according to real language demands. Students received a complete description of the task (workplan) and started preparing with

classmates in the allotted preparation time. The task handouts contained the task description, performance guidelines, and assessment rubric (if necessary).

By the end of the semester (January 2010), all groups took the midterm written exam, which had already been designed by the Foreign Language Department of the school. This test consisted of 139 multiple-choice questions, with 40 listening comprehension questions, 40 vocabulary questions, 40 grammar questions, and 19 textual comprehension questions. Along with the midterm written exam there was a midterm speaking exam. On this speaking exam, all students from both the treatment and the control groups were interviewed by an external interviewer, another Spanish teacher invited to participate in the research. This ten-question interview constituted the tenth assessment task and was titled "Entrevista de trabajo 2" (Job Interview 2). Students were questioned about personal information, personality traits, favorite indoor and outdoor activities, school information, and other general topics such as dates and times. Lastly, a ten-question survey was carried out. Through this survey students expressed their opinions regarding assessment preferences in Spanish class.

All instruments in this study were administered to all groups except the 10 assessment tasks which correspond to the treatment. Similarly, all instruments were administered by the classroom teacher (researcher) except the semester speaking test. The personal information form and the midterm multiple-choice exam were designed by the Foreign Language Department of Tabb Middle School. The entrance test was adapted by the teacher from the website above mentioned. The ten assessment tasks and the semester speaking test were designed by the researcher. However, as mentioned previously, the speaking test was administered by a guest teacher who had the possibility to choose among a wide range of questions—that is, there was not a fixed 10-question interview.

It was pivotal to invite an external interviewer for the speaking exam as well as to take the above-mentioned steps in order to make the study more valid. In this regard Norris et al. (1998) states that validity is one of the main threats to the effective implementation of a Task-Based Language Approach.

Data Collection

The grading software "Gradequick" was used to enter grades and calculate the groups' averages (mean and median) for the entrance test and the midterm multiple-choice exam. The semester speaking exam was graded by the tester (guest teacher) using the rubric on the task handout. All ten assessment tasks were graded by the classroom teacher, but these grades were not used for the final analysis of this study. The teacher's journal of observations on the pre-task, during, and post-task phases were added to the analysis of the quantitative information. Thus, qualitative and quantitative information were combined in the analysis.

Data Analysis

Scores obtained on the semester written and speaking exams were compared to the scores obtained on the entrance test for each group. Both the group means and the medians were calculated in order to have a more reliable evidence of the results. The comparison between the entrance test and the exit test averages provided relevant information regarding the effect of the treatment (implementation of a task-based assessment approach). This analysis was accompanied by the teacher's observations and the result from the assessment survey. Students' opinions and preferences served as complementary arguments to the analysis of quantitative results.

Results

As previously indicated, an entrance test, a midterm speaking exam, and a midterm multiple-choice exam were the principal assessment tools used in the present study to measure the impact of a task-based assessment approach on the learning of Spanish as a foreign language in Tabb Middle School in Yorktown Virginia. The information from the above-mentioned survey was solely used to complement the analysis made from the different tests.

Entrance Test

This test was administered on the first week of the 2009-2010 school year (September). It was administered to four Spanish 1 classes in order to measure their initial level of Spanish language proficiency.

The test consisted of 30 multiple-choice questions assessing vocabulary and grammar. Table 2 shows the scores for each class in a grouped frequency distribution with intervals of ten. The table also shows the size of each class, as well as the medians and the standard deviations.

Table 2
Grouped frequency distribution of entrance test scores

Raw scores (intervals of ten)	Frequency			
	Class A	Class B	Class C	Class D
0-10	0	0	0	0
11-20	6	7	2	1
21-30	8	5	4	4
31-40	5	7	4	9

41-50	1	3	3	3
51-60	2	0	0	0
61-70	0	1	4	0
71-80	0	0	1	0
81-90	0	0	0	0
91-100	0	0	0	0
n	**22**	**23**	**18**	**17**
Median	**30**	**30**	**35**	**37**
SD	**11.23**	**13.37**	**17.75**	**6.91**

Groups B and D correspond to the treatment groups and A and C to the comparison groups. As we can see, groups C and D show the highest averages and the smallest class size class as well. Group D also shows the smallest spread in its data (standard deviation).

All four groups obtained scores in the 30-40 range on the entrance test. Group D (one of the treatment groups) obtained the highest score (37/100); group C obtained the second highest score (35/100); groups A and B obtained 30 out of 100. These numbers correspond to the initial group data and are intended to give us an idea of students' level of language proficiency and background knowledge before undertaking the treatment for the present study.

Midterm Multiple-choice Exam

The midterm multiple-choice exam consisted of 139 questions distributed as follows: 40 listening comprehension questions, 35 vocabulary questions, 40 grammar questions, and 24 questions for textual comprehension. This exam was designed by the Foreign Language Department of Tabb Middle School in previous years. Students took this exam on the last week of January 2010. It took them between 50 and 70 minutes to complete this test. Table 3 shows the scores for each class in a grouped frequency distribution with intervals of ten. This table also shows the size of each class, as well as the medians and the standard deviations.

Table 3
Grouped frequency distribution of the midterm multiple-choice exam scores

Raw scores (intervals)	**Frequency**			
	Class A	**Class B**	**Class C**	**Class D**
51-60	0	0	0	0
61-70	1	2	2	0
71-80	10	9	4	5

81-90	8	10	8	9
91-100	3	2	4	3
n	**22**	**23**	**18**	**17**
Median	**79.50**	**83**	**81**	**87**
SD	**7.93**	**7.86**	**7.94**	**7.20**

As we can see, the two treatment groups (B and D) show the highest averages (83 and 87 respectively), and the spread of the data for all groups is quite similar. Even though the treatment groups were never assessed throughout the semester using multiple-choice type tests, they did very well on this midterm test to the point of obtaining higher scores than the groups who were always assessed using multiple-choice tests. These scores show us a very positive result in terms of the assessment treatment applied to the two groups. As it will be analyzed later, this difference in averages may not be statistically significant, according to the tools of statistic analysis used, yet it shows a practical significance. As said, students did better even though they were not trained on this type of test.

As already mentioned, the treatment groups obtained the highest scores in the midterm multiple-choice exam even though they were never assessed through multiple-choice tests and quizzes during the semester. These students were assessed through communicative tasks which usually consisted of dialogues, interviews, or written letters, among others. All of these tasks were intended for students to produce Spanish in a more contextualized fashion either by speaking or by writing.

Midterm Speaking Exam

The midterm speaking exam consisted of a 10-question interview covering topics such as personal information, likes and preferences, personality traits, school subjects, and other general questions about dates and times. These were the topics covered during the first semester by the four groups. The tester was a Spanish teacher from the school who was invited to assess students in order to make the research more valid. This teacher has great experience teaching Spanish 1 students as well as teaching the Spanish curriculum of Tabb Middle School. The tester was instructed to select the questions she wanted from a long list or even to ask the questions she considered suitable for the conversation. These interviews were administered during two 90-minute class periods. Table 4 shows the scores for students in each class, the size of each class, the median, and the standard deviation for each group.

Once again, the treatment groups obtained the highest averages, 92 and 96 respectively. These results are consistent with the way students in these two groups were assessed throughout the semester. Most of the tasks were aimed at speaking; therefore students were better trained to take this type of test.

The difference in averages between the treatment and the comparison groups is both statistically and practically significant, which enables us to suggest the assessment treatment as a great alternative to use in the teaching of Spanish. Yet this suggestion as well as others will be more deeply explained in the following section, along with conclusions.

Table 4
Grouped frequency distribution of the midterm speaking exam scores

Raw scores (intervals)	Frequency			
	Class A	Class B	Class C	Class D
51-60	0	0	0	0
61-70	1	1	1	1
71-80	10	7	5	2
81-90	4	1	7	3
91-100	7	14	5	11
n	**22**	**23**	**18**	**17**
Median	**82**	**92**	**86**	**96**
SD	**11.01**	**10.16**	**8.59**	**8.51**

Statistical Analysis

This statistical analysis consisted of an independent samples t-test using the statistics program SPSS. What this t-test did was to compare the difference in means obtained in both midterm exams (multiple-choice and speaking) by the treatment and comparison groups. From this operation, it was determined if the difference in means was statistically significant. Also calculated was the probability of obtaining the same result by chance. In order to make this conclusion, the .05 level of significance was used. That is, if the difference in means was of .05 or less, it would be considered statistically significant. If the difference in means was bigger than .05, the probability of obtaining the same result by chance is bigger and therefore the research hypothesis may not be well-supported.

Table 5
Group statistics for the multiple-choice exam

	VAR00002	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	1.00	40	82.4250	7.74890	1.22521
	2.00	40	80.9000	8.04730	1.27239

The first row contains the data for the treatment groups and the second row corresponds to the data for the comparison groups. Both groups have the same number of students (40). The mean for the treatment groups is 2 points higher than the mean for the comparison groups.

The statistical result shows that the difference in means was not statistically significant at the 0.05 level. That is, the significance (2 tailed) obtained (.391) is bigger than the 0.05 reference level. In other words, the possibility of obtaining the same difference in means by chance is 39.1 out of 100 (too large probability). However, as previously stated, this difference in means has a practical significance taking into account that the treatment groups were never assessed through multiple-choice type tests during the semester. Yet, they obtained a two-point higher score in their means. It also suggests that language production-oriented teaching will also prepare students to take more structured type tests such as multiple-choice tests.

Table 6
Group statistics for the midterm speaking exam

	VAR00002	N	Mean	Std. Deviation	Std. Error Mean
VAR00004	1.00	40	89.0000	9.70540	1.53456
	2.00	40	84.5000	10.13499	1.60248

On this test, the mean for the treatment groups is 4.5 points higher than the mean for the comparison groups. The statistical operation showed that the difference in means was statistically significant at the 0.05 level. That is, the significance (2 tailed) obtained (0.024) is smaller than the 0.05 reference level. In other words, the possibility of obtaining the same difference in means by chance is 2.4 out of 100 (a small probability). This difference was expected because the treatment groups were mostly assessed through oral type tests during the entire semester. To sum up, the treatment groups obtained higher scores in both semester exams, but only in the speaking exam was the group score (means) considered statistically significant according to the independent samples t-tests.

Conclusions and Recommendations

The treatment applied in this study yielded positive results in students' performance on both semester exams. Even though students in the treatment groups were never assessed through multiple-choice type tests, they obtained higher scores on the midterm multiple-choice exam than the control groups. This result indicates that students can be taught a language, in this case Spanish, following a communicative approach and assessed through communicative tasks, and still be prepared for psychometric type tests (i.e. multiple-choice tests). As it

is known, most of the standard national and international tests (TOEFL, GRE, SOLS, SAT, AP, etc.) are designed following the psychometric tradition. Thus, teachers do not necessarily have to teach students through a grammar-based approach in order to prepare them for standard language tests, which as said, are usually designed in the multiple-choice format.

Regarding the midterm speaking exam, the treatment groups also obtained higher scores than the comparison groups. This result was not only significant to the teacher and researcher (face validity), but also it was determinant to have a statistical significance according to independent sample t-test illustrated in the previous section.

In short, students in the treatment groups scored higher than students in the control groups on both semester exams. It is important to remember that on the entrance exam, the higher scores were for one of the treatment groups and one of the control groups.

The main conclusion that this study depicts is that task-based language assessment in a middle-school Spanish class is not only possible but also effective. This language assessment approach is highly recommended in terms of high achievement on speaking tests (shown through the speaking semester exam scores) and in terms of high achievement on multiple-choice type tests. In other words, the six-month treatment applied in the two Spanish classes at Tabb Middle school produced higher levels of communicative and linguistic performance. However, for further studies in this specific area (TBA in a foreign language class), it is suggested to have a more extensive period of teaching and assessment in order to confirm the results of this treatment. If future research studies confirm these results, the implementation of a TBA can be generalized to other levels of language proficiency as well as to other language learning processes, as foreign or second language.

If TBA is to be adopted, I propose the following recommendations that can make task-based language assessment smoother and the learning process more effective. It was observed and also stated in the assessment survey that students get very nervous when taking oral tests or performing oral tasks, especially those in which they are to speak in front of the class. Then, it is highly recommended that teachers do significant preparation on the pretask stage so that students' level of anxiety, one of the main threats to performance-based language assessment, can be lowered and consequently their language performance can increase in spontaneity and fluency. That is why on the pretask stage students need to be given sufficient time to comprehend the task, to ask for clarifications, to practice, and to rehearse interactions. According to Krashen (1985), once the anxiety variable is controlled and the motivation is enhanced, students are better prepared for language learning success (Affective Filter Hypothesis).

The post-task is another learning momentum for both teachers and students. On the one hand, the teacher can benefit from students' reflection and feedback (metacognitive reflection) and make adjustments to the task or the task process. On the other hand, students can benefit from the teacher's feedback regarding linguistic, communicative, or interactional performance.

These two task stages will undoubtedly help students lessen their level of stress as well as reflect on their own language learning process. On the assessment survey, students expressed preference for group task and interview type tasks when assessed. Group tasks can serve well the demands of some real-life tasks (due to their interactive nature) at the same time that help students cope with pressure. However, in group tasks teachers must ensure that all students have the same load of communicative information to produce during the task. Another assessment alternative according to the students' opinions is one-to-one oral interviews, preferably with the teacher. Thus, a good combination of group conversations, dialogues, and interviews may compose a good repertoire of task-based language tests.

According to the assessment survey, learners continue to prefer multiple-choice tests over other types of assessment. As just mentioned, it is due perhaps to the fear of speaking in front of an audience. That is why it is imperative that teachers offer clear guidance and encouragement during the pretask stage so that students can see a meaningful and functional purpose behind task performance. Students need to understand that a task-based language methodology not only prepares them for academic achievement but also for real-life linguistic challenges.

Since oral tests are generally more time-consuming than written tests, it is recommended that teachers design rubrics that can be easy to use by the teacher and easy to understand by students. These rubrics must be incorporated into the task description sheet so that students can refer to them before and after accomplishing the task. It is suggested that teachers use the same rubric, if possible, for each task so that both teacher and students get familiar with them. Familiarity with the rubric will allow the teacher to maintain good control of time in order to not fall behind in the development of syllabus content.

As stated throughout this article, despite the fact that task-based language teaching models have been traditionally applied in second language classrooms and special language programs (immersion, specific purpose, etc.), this experiment showed that a task-based language assessment model is possible in a foreign language classroom.

This experiment also showed that a task-based language model can be implemented in beginning language classes with positive results. The implementation process may not be as smooth as that in higher proficiency groups or that in specific purpose programs, yet a significant degree of communicative and linguistic competence can be achieved. As Willis and Willis (2007) put it, "This is one of the most valuable things we can give a learner: the confidence and willingness to have a go, even if their language resources are limited" (p. 2).

Based on the positive results this task-based language produced, it is recommended that not only the assessment portion of the language learning process be implemented but also the whole syllabus. A task-based language syllabus can be designed, depending on the circumstances, by carrying out a linguistic needs analysis or by following established standards of learning. However, a syllabus based on learners' linguistic needs may yield more real-life type tasks.

Finally, this study intends to serve as the foundation or complement of further studies on task-based language learning processes in foreign language classrooms. The results obtained in this research project are satisfactory, yet it is my expectation that more teachers and researchers will carry out similar experiments that confirm not only these results but also the benefits of the task-based language teaching methodology.

Bibliography

Baker, D. (1989). *Language testing: A critical survey and practical guide.* London: Edward Arnold.

Bygate, M., P. Skehan, & M. Swain (2001). *Researching pedagogic tasks, second language learning, teaching, and testing.* Harlow, England: Longman.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

Krashen, S. (1985). *The input hypothesis: Issues and implications.* Beverly Hills, California: Laredo Publishing Co.

Lee, J. (2000). *Tasks and communicating in language classrooms.* Boston: Mc Graw- Hill.

Long, M. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam and M. Pieneman (Eds.), *Modelling and assessing second language acquisition (pp. 83-96).* San Diego, California: College-Hill Press.

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

Norris, J., J. Brown, T. Hudson, & J. Yoshioka(1998). *Designing second language performance assessment. Technical report.* Hawaii University, Manoa: Second Language Teaching and Curriculum Center.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. New York: Cambridge University Press.

Willis, D.,& J. Willis (2007). *Doing task-based teaching*. Oxford: Oxford University Press.