



International Journal of Languages' Education and Teaching
Volume 7, Issue 1, March 2019, p. 395-410

Received	Reviewed	Published	Doi Number
01.01.2019	11.03.2019	25.03.2019	10.18298/ijlet.3238

Correlation or Co-determination: An Analysis of Alternative and Traditional Testing Employed in English Language Teaching

*Seda AYDAN*¹

ABSTRACT

The aim of this study is to determine the correlation between alternative and traditional testing utilized in English language teaching classes and to investigate students' perceptions on these two methods. As several problems related to traditional testing have been pointed out, alternative assessment has started to be favored by teachers and learners. Portfolios are now being commonly used to assess English language skills of students as a replacement to tests. This study has been inspired from the need to see to what extent these two methods correlate. For this purpose, final grades and portfolio grades of students in a private university English preparation school in Ankara have been used and their Pearson Correlation Coefficient has been calculated. The results show that portfolio grades and final grades positively correlate which may indicate that they both serve to test the similar skills despite the differences in their design and implementation. In order to learn the students' opinions on these methods, 10 students have been interviewed and 7 students reported that they preferred portfolios rather than tests due to various reasons such as portfolios' allowing more space for feedback, not measuring performance at one time only and being less threatening for them. Also, the high positive correlation between these methods and students' support for alternative methods signaled the need for more integration of alternative assessment. Thus, the results of this study can be useful in designing and making decisions on assessment types.

Key Words: Alternative assessment, correlation between testing methods, students' perceptions on testing methods.

1. Introduction

Assessment has always been an indispensable component of teaching process and something that works like gears of education together with curriculum. However, issues such as what to test, when to test and how to test are among the challenges of the educators as there is no single method fitting into various contexts of teaching and learning. Language tests have been called even more challenging as practitioners are struggling to assess language skills rather than the knowledge of a specific content. Brown and Hudson (1998) underline that language testing pose more challenges for the educators since the number of educational decisions to make is much higher. Every method of testing has its own advantages and disadvantages, but tailoring assessment methods for the needs of the students has a great potential to contribute to learning process. According to Davison and Cummins (2007)

¹ Instructor, TED University, seda.aydan@tedu.edu.tr

assessment and evaluation judgments have generally been conveyed long after the event, formulated in often mysterious and non-negotiable terms, with a great dependence on technical terminology and statistics. As a result, assessment and evaluation have always been taken for granted in English language teaching, but most of the time misinterpreted by practitioners, barely added as a part in English language teacher training, and never deeply challenged by key stake-holders. However, the question of how to assess students is almost as old as learning itself and it keeps its importance as students spend a lot of preparing for the tests and teachers put a lot of effort on designing the assessment tasks (Marino, Pickering & McTighe, 1993). The statement: "Not everything that counts can be counted and not everything that can be counted counts" is said to be written in Einstein's office wall (Herman, Aschbacher, & Winters, 1992, p.6). The same thing goes for testing as there is no consensus among teachers and administrators on what performance to test for assessment. Traditional assessment has long been criticized for being too outcome oriented and it has been disapproved by many because of the exam stress that it imposes on the learners causing effective filter which hinders learning (Moeller & Reschke, 1993). There are serious concerns about the traditional testing in literature. They particularly question the outcomes of the learning that is assessed by such methods and their inability to provide data so as to shape instruction. Do improvements in test score performance really represent progress in learning or do they reflect an infertile curriculum with students being "drilled and killed" on expected test content? These questions are rightfully asked by many such as Linn (1994) and Shepard (2000). Furthermore, Huerta-Macias (1995) argues that traditional testing leads to some serious concerns regarding norming, linguistic, and cultural biases and alternative assessment includes valid and reliable procedures that avoid many of such problems inherent in traditional testing.

Herman, Aschbacher, and Winters (1992) provide a guide to assessment and underscore the importance of many aspects to be considered in assessment. One of the vital points they raised is the importance of the match between tasks and intended student outcomes for effective assessment. In traditional testing, we expect students to circle the correct answer or fill in the blanks correctly. However, the actual outcome we expect in language classes are speaking fluently and correctly and writing well-organized texts with well-supported contents. Thus, there is a clear mismatch between testing and learning outcome. At this point it is also appropriate to mention the washback effect of tests. Washback effect is defined as the effect of the tests on teaching and learning and it is important to be aware of the washback effect of any kind of assessment which may force teachers and learner to do the things that they do not necessarily do otherwise (Pearson, 1988). There is also research indicating that tests might greatly affect what teachers teach and how teachers teach causing exam oriented teaching methods which can be identified as a form of washback effect as well (Alderson & Wall, 1993). Based on this notion, it is appropriate to report that if we can shape testing we can lead to positive effects of it on teaching as well. If we are expecting students to perform communicative tasks, to be creative or improve their higher order skills, we need to design our assessment tools accordingly. Since testing will always be a part of instruction, it requires careful thought to make it a facilitator of the curricular objectives. It is also intermingled with how we learn and discussions of constructivism.

2. Review of Literature

2.1. A Constructivist Perspective on Testing

Constructivism has contributed to what we know about the process of learning denoting the importance of self-regulated learning through experience. Notion of constructivism that is based on the studies of Piaget (1970), Vygotsky (1978) and Bruner (1986) tells us that based on a synthesis of current work in cognitive psychology, philosophy, and anthropology, constructivism identifies knowledge as a developmental socially and culturally shaped and so non-objective process of resolving internal cognitive conflicts which can frequently be seen via concrete experience, collaborative discourse and reflection (Brooks & Brooks, 1993). Education from a constructivist perspective is not about presenting information to the students and testing if they have mastered the same version of knowledge or not. It is about helping students to master how to obtain knowledge. Fosnot (1996) states that we as human beings are not able to reach an objective reality because we are constructing our version of it, while simultaneously transforming it and also ourselves. This idea makes plenty of room for creativity and personalization of the information. At this point, it is essential to highlight the importance of assessing students by allowing enough room to reflect the knowledge which they have constructed.

Brooks and Brooks (1993) presents five far-reaching principles of a constructivist pedagogy: (1) posing problems of emerging relevance to learners; (2) structuring learning around 'big ideas' or primary concepts; (3) seeking and valuing students' point of view; (4) adapting curriculum to address students' suppositions; and (5) assessing student learning in the context of teaching (p.viii). When we take standardized testing practices into consideration within this frame, it is justifiable to say that traditional methods of testing ignore most of these principles by not providing any opportunity for students to shape their learning process and making the process as standard as possible. Discourse, debate and inquiry are cornerstones of constructivist classrooms (Anderson & Piazza, 1996). Nevertheless, they are not given place in most of the assessment methods. Furthermore, as how we test directly affect how we teach, such practices are likely to be shadowed by traditional way of presenting information and expecting students to take notes and memorize information without making it their own by questioning or critical thinking. Such practices are bound not to go beyond classroom activities done rarely if they are done at all. In this case, students are not given any opportunity to take the responsibility for their own learning as well. In addition, in constructivism, changing behaviors or improving skills of students are the main objectives of teaching. Constructivism focuses on deep understanding, concept recreation and structuring active student reorganization (Brooks, 1990). However, traditional form of assessment does not evaluate or assess this form of instruction. With these concerns in mind, as Alderson (1991) indicated there have been obvious changes in the content of testing. So as to obtain a more positive washback from the tests, testing is moving into the new "alternative" one rather than the traditional one just like the teaching itself.

2.2. The Rise of Alternative Assessment

As grading globally seen as the major indicator of success, tests inevitably mingle with learning process. Therefore, it is vital to create assessment tools which facilitate higher-order skills to ensure the development of such skills within the curriculum. In traditional tests, students' final outcomes of students' efforts are considered as the summary of their learning. The focus is on mastering

information about something rather than developing strategies on how to learn. On the other hand, in real life, the tasks that students have to complete or achieve are predominantly on how to evaluate, analyze or synthesize information rather than recalling it. Most of the time, traditional assessment just monitor students' learning and makes a classification of students. They are divided into the groups of the ones who know the content and those who do not (Berlak, 1992). In this type of an environment, students focus on their individual performance and evaluate it in relation to the others. Thus the learning environment gets more and more competitive and outcome oriented (Johnston, 1992). Also, traditional assessment focuses on cognitive skills ignoring students' areas of interests, specific skills, values and background (Raven, 1992). Students' use of social skills and other competencies are given little attention. The instructor has the full power over what to teach and test, so students do not play any roles in what is taught or assessed (Heron, 1988; Sessions, 1995). Although this might seem satisfying or promising for the instructor, this type of an approach might prevent students from developing skills to evaluate what is essential to learn and to what extent they are making progress. However, if we consider the issue in terms of practicality from teachers' point of view, traditional assessment is much more practical to administer and to grade. To elaborate, there are limited choices when it comes to traditional testing methods namely multiple choice questions, fill in the blanks and matching types of questions and so on. Therefore, the teacher can make a choice easily or can pick up from one of each question types to prepare a test. As these testing methods are the ones that most teachers are used to and also were exposed to, it is very practical for them to apply and grade these tests. Besides, many books offer ready-made multiple choice tests and online tools that create such tests according to the needs of the teacher. For this reason, such tests might be both time-saving and convenient for the teacher. Last but not least, it is easier to provide variety in such testing methods.

On the other hand, when teachers decide to utilize alternative assessment methods, they have to be creative and careful. They have to consider the characteristics of the context of teaching and design the tasks accordingly. The tasks should both foster learning and also embrace multiple intelligence and learning types of the students. As they need to provide an alternative to what students are accustomed to, teachers need to give more thought to offer an effective one. Teachers are supposed to tailor such assessment tools for the very specific contexts that they are teaching. For this reason, it takes relatively more energy and labor to create alternative assessment tools. What is more, evaluating or grading alternative assessment methods is another issue of discussion as teachers need to develop their own rubrics or evaluation criteria. This requires not only time but also a lot of effort. To address this issue, Anderson (1998) suggests developing the rubric together with students and presents an outline for this purpose. She also adds that by formulating criteria before negotiating with students, teachers can feel confident that their expectations will be addressed and also with this method students can have a little guidance to judge a task. Self-assessment is another method which can be benefited by both teachers and students so as to evaluate the performance and to enhance learner autonomy (Javaherbakhsh, 2010).

Knowledge has a collection of realities with accompanying multiple meanings just like suggested in constructivism (Roderick, 1991). Unlike traditional assessment, alternative assessment has the capacity to provide a platform in which students can create and share their own interpretation of situation. It takes what, how and why students learn into consideration (Hutchings, 1993; Johnston, 1992). Anderson (1998) illustrates the comparison between traditional and alternative assessment with the help of the following figure:

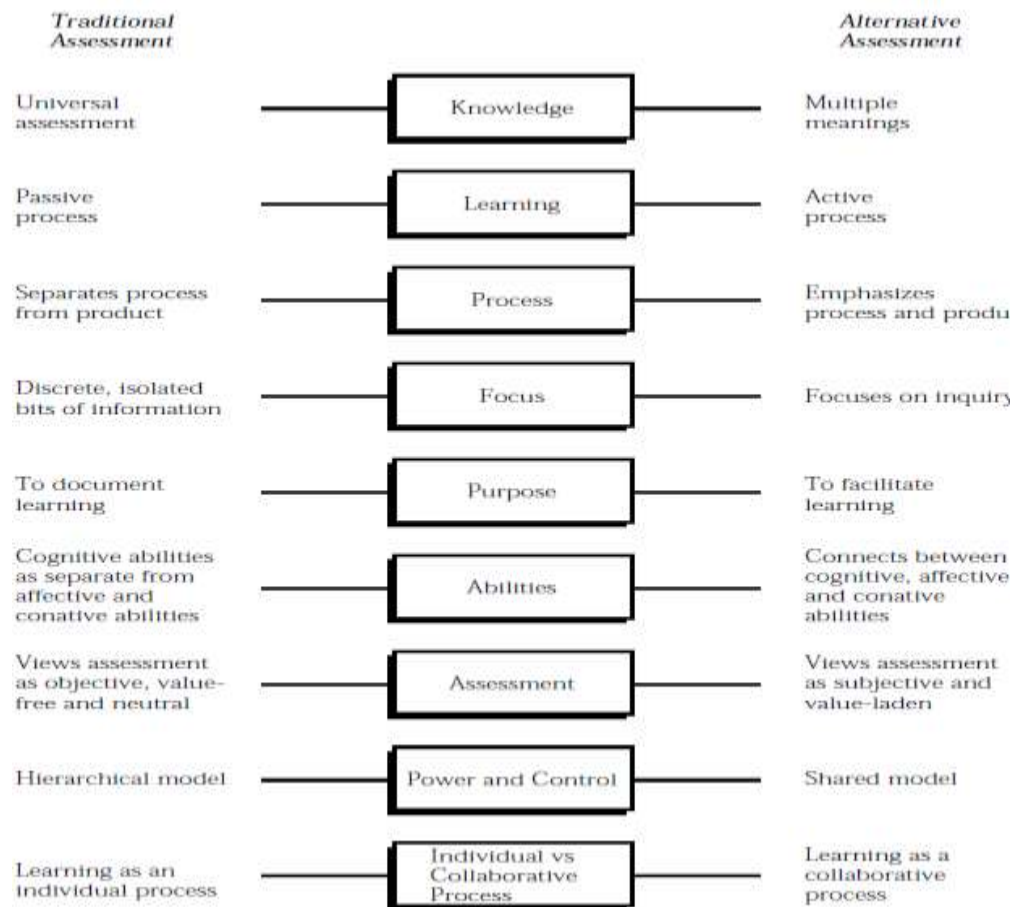


Figure 1. Anderson's Comparison of Philosophical Beliefs and Theoretical Assumptions of Traditional and Alternative Assessment

As seen in Figure 1, traditional method of assessment and alternative assessment differ from each other in various radical areas. To elaborate, traditional assessment focuses on product and cognitive abilities while keeping process and affective domain and conative abilities aside. Learners are passive and teacher is the decision maker in the hierarchical process. On the other hand, alternative assessment addresses affective domain as well by taking students' interests and values into consideration. Learner is active in the process and shares the power with the teacher. S/he collaborates with the educator and shares the responsibility rather than being passive test taker. Furthermore, as students create their own content with their own pace, the environment is collaborative rather than competitive. Besides, as Heron (1988) denotes, it is likely to enhance a democratic decision making process embracing both learners' and teachers' opinions. They are partners or co-learners in this context (Bintz, 1991).

Bearing these ideas in mind, for the last decades, tests have moved from being "discrete-point", knowledge-focused to assessing the complex process of learning. Now, there has been an increase in the number of educators who take more integrative approaches to testing making use of portfolios, oral production inventories, cooperative student-student techniques, and authentic testing rubrics (Richards & Renandya, 2002). These tasks are designed to be more holistic and found interesting for students as they offered an intellectual challenge (Carlson, 1991). Also, they focus on the progress rather than the result (Short & Burke, 1991). As Alderson and Banerjee (2002) stated alternative testing

methods are likely to have beneficial washback as they welcome creativity and encourage learner autonomy. These methods have also been found more communicative, less stressful and more enjoyable for students (Haggstrom, 1994).

With the help of these assessment tools, teachers take the opportunity to assess students' performance during the course of the semester rather than testing their performance only via a traditional exam style in one shot (Cheng, 1999). This can be particularly helpful in testing language skills as it takes much more time to acquire and internalize a language compared to mastering a specific content. Language acquisition is a complex and diverse process that requires distinctive methods of testing (Cheng, 1997). Alternative testing subsumes different aspects of language acquisition. Since language development is complex, assessing proficiency at one point in time might not give us reliable data on how far students have progressed or even what they are capable of learning (Danon-Boileau, 1997). Alternative assessment provides more room for these complexities while traditional assessment is likely to ignore many processes in language learning. However, traditional assessment methods are still preferred very often as they are found objective and valid by many institutions especially so as to make important decisions (Alderson, Clapham & Wall, 1995). In addition to the advantages of traditional testing, this is also about how these types of testing methods are graded. These can easily be graded objectively as there is one single correct answer and they do not have much room for human mistakes. On the other hand, grading alternative assessment products are challenging and demands more effort to be objective. For this reason, while alternative testing methods seem promising, traditional testing methods are considered indispensable. As a result, there are many institutions just like the university that this research has been conducted which employs both of these methods to have a better, clear and detailed understanding of students' knowledge, progress and performance and also to measure their skills as objectively as possible. Thus, students are assessed from various aspects with this combination.

This study has its genesis in a need to know the correlation between traditional and alternative testing methods mentioned above. Revealing the relation between these two methods can contribute to the educational practices in two main ways. To start with, it can give us a better framework of these two methods providing us the chance to test if they are opposite to each other. They seem to be aside in their roots and in the theories behind them, but they might be very similar in measurement and even in practice. Moreover, this study might help us analyze these methods in depth to make appropriate decisions on when to use them, where to use them and whether to use them interchangeably or not. For this reason, it is possible to say that this research has potential to put one more brick into the construction of valid and trustworthy assessment methods. In the context of the study, these two methods are supposed to measure the skills of the students differently. Thus, the origin of this study takes us to the discussion of the construct validity of the testing methods used in the place of the study. As Messick (1996) argued construct validity is a multifaceted but unified and overarching concept and it can be studied from various aspects. This study has been inspired from the notion that correlation between two testing methods can actually be a sign of construct validity. Chapelle (1999) offers reasons for the changes in educators' understanding in validity and states that there is no single answer to the question "What does our test measure?" or "Does this test measure what it is supposed to measure?" In the context of this study, questions to be asked are "Is it likely that alternative and traditional testing methods used are measuring the same thing?" or "Do they really measure the things that they are supposed to measure?" Namely, the alternative testing method, which is the

portfolio in this study, is designed to test progress during the course of the semester and is supposed to focus on how far a student has advanced. If students cannot perform well in one task; they always have the chance to improve their performance in the next one, so they have enough opportunities to do better. Traditional assessment methods, on the other hand, are created to assess performance at one time and provide no room for compensation. If they cannot give correct answers, they are very likely to fail without a second chance. Based on this notion and the differences between these methods mentioned above, it is possible to say that they can be clearly distinguished from each other. However, this study investigates the correlation between these two methods and the researcher of this study proposes that if there is a very strong correlation between these two methods, it means that they are not as away from each other as they seem. Alternative assessment aims to assess progress and appreciates productivity and creativity of the students. Although these things are kept separate from exam performance, positive strong correlation between these two methods may be an indicator of their parallelism.

Therefore, interpretations of the scores taken from these two methods are crucial. This is also strongly related to the validity issues of these assessment methods. New concept of validity advocates the characteristics of the inferences made on the basis of test scores and the purpose that these tests are used. It is reasonable to validate not a test, but 'a principle for making inferences (Cronbach & Meehl,1955). Outcomes of testing should be carefully analyzed and justified by studying the consequences and validity issues. Figure 1 created by Messick (1996) illustrates new facets of construct validity and kinds of arguments that should be used to examine test results.

	Inferences	Uses
Evidence	Construct Validity	Construct Validity + Relevance/Utility
Consequences	Construct Validity + Value Implications	Construct Validity + Value Implications + Relevance/Utility + Social Consequences

Figure 2. Messick's progressive matrix on construct validity

With respect to Figure 1, it can be noted that construct validity is intertwined with various concepts. Validation also depends on relevance and utility. Therefore, seeing the correlation between alternative and traditional testing methods might help us to see how relevant they are with each other, how they actually affect each other and how we can interpret their results. Thus, this study has been inspired from the need to measure the correlation between these two methods and uses Pearson correlation coefficient to see the covariance between these two assessment methods. In order to see this correlation –if it exists or not– Pearson correlation coefficient of final grades of 37 elementary, 37 intermediate, 37 upper upper students at an English preparation school of a private university in Ankara, Turkey and the portfolio grade of the same students have been calculated. The researcher expected to have a positive correlation between these two variables hypothesizing that students who regularly and diligently complete their portfolio tasks are likely to perform well in the exam.

3. Aim of the Study

This study investigates the correlation between two different testing methods implemented at a private university in Ankara, Turkey. Students' opinions on these two methods are also included in the study to see these methods from a wider perspective. Research questions of the study are as following:

- Is there a correlation between the final grades and portfolio grades of students at TED University English preparatory school?
- Are traditional testing method that is finally exam and the alternative testing method that is portfolio assessing different or same skills?
- Which one of these methods namely exams or portfolio tasks are stated to contribute to their learning better by students?

4. Method

4.1. Participants

Participants of the current study are 37 students from each level namely elementary, intermediate and upper. This means that 111 students participated in this study. Portfolio and final grades of these 111 students were used to calculate correlation between traditional and alternative assessment used in English preparation program. Then 10 students were interviewed by the researcher. Three of the students were in elementary level while four of them were in intermediate and three of them were in upper level. All the students are Turkish Republic citizens and their mother tongue is Turkish. All of these participants are students in English language school of a private university in Ankara. Convenience sampling method was used to choose the subjects of the study. Thus, students who were accessible to the researcher were asked to take part in the study and invited to an interview. A consent form was signed by students to state that participation was voluntary and they could withdraw whenever they liked.

4.2. Instruments and Data Analysis

A mixed research method was used in this study to answer the research questions. As a quantitative research method Pearson correlation coefficient calculation was made by using the final grades and portfolio grades of participants. To collect qualitative data, a group interview with students who were exposed to both of these methods was conducted. With the help of mixed method, and so by combining quantitative and qualitative research and data, the researcher obtained in breadth and depth of understanding and corroboration, and reduced the weaknesses inherent to using each approach by itself.

So as to calculate the Pearson correlation coefficient, final grades and portfolio grades of the students in each level were used and the value r which shows the correlation was calculated. As Rodgers and Nicewander (1988) point out, Pearson correlation coefficient is utilized to measure the strength of a linear association between two variables. It is also known as the product moment correlation coefficient and showed in a sample by r . It can take a value ranging from -1 to 0 or $+1$. If r value is -1 , it means there is a strong negative correlation which tells us that while one variable increases, the other one decreases. If r value is 0 , it means there is no correlation between these two variables at all. Finally, if r value is $+1$, it tells us that there is a strong relation between these two variables. This

calculation method was used in this study to see if a correlation exists between traditional and alternative assessment methods used within the same context.

In order to present a complete report of this study, it is essential to give information about the nature of the portfolio tasks and the final. Portfolio included speaking and writing tasks which were completed by students in every two weeks. In every two weeks, students completed a unit and learned new vocabulary items, read passages and listened to texts related to the theme of that specific unit. Next, they wrote a text related to the topic. They addressed a problem, responded to a question, made predictions or planned new systems of testing. They wrote their first draft at school by getting help from each other and the internet. Then they got feedback from their instructors and they wrote the second draft. Their instructors gave feedback on their second draft as well, but writing the third draft or not was up to them. When it comes to their speaking tasks, students were expected to do individual or group presentations and to give feedback to their friends' presentations. In these presentations, students presented a research they did, a memory of them, or something they designed to solve a problem in the world. Students took notes while their peers were presenting and they gave constructive feedback to each other. Instructors gave feedback and graded the presentations and writing texts of the students. For the presentations, instructors considered fluency, accuracy in grammar and pronunciation, use of vocabulary and content. To grade the writing texts, teachers considered content, logical flow of ideas, use of vocabulary and grammar. Students got half of their portfolio grades from their writing tasks and the other half from the speaking tasks.

The final exam that was used in this study included reading, writing and listening sections. Reading and listening sections consisted of multiple choice questions. Knowledge of grammar and vocabulary was implicitly tested in the questions. That is, grammar or vocabulary questions were not directly asked, but students needed to know particular grammar topics and vocabulary items to understand the text or questions. In the writing part of the exam, students wrote a text selecting one of the two topics presented in the exam. Their writing was evaluated based on a set of criteria similar to the portfolio's, but the rubric of the exam was more detailed and included sections such as content, logical flow of ideas, use of vocabulary, and grammar.

Students got grades out of 15 from portfolio and out of 30 from the exam. Then, by comparing these two numbers, Pearson correlation coefficient was calculated. However, to dig deeper and to see students' ideas and preferences on these two methods, an in depth interview was planned. Thus, a group interview was conducted with 10 students and they were asked: "Which one of these types had a more positive affect on your learning: tests or portfolio? Please explain why?" The answers of the students were recorded and transcribed. First, the number of the students who preferred portfolio or tests was calculated. Then in order to analyze this structured interview, the constant comparative method was implemented. The constant comparative method was defined by Maykut and Morehouse (1994) in the following way:

A method of analyzing qualitative data which combines inductive category coding with a simultaneous comparison of all units of meaning obtained. As each unit of meaning is selected for analysis, it is compared to all other units of meaning and subsequently grouped (categorizing and coded) with similar units of meaning. If there are no similar units of meaning, a new category is formed. In this process, there is room for continuous refinement; initial categories are changed, merged, or omitted; new categories are generated; and new relationships can be discovered (p. 134).

The answers were coded and listed in a table and they were categorized under different categories. Every answer was taken as an item and they were coded according to their theme. A coding table was created with their frequencies and numbers that occur in the transcriptions. That gave the researcher a better idea on what was reported during the interviews.

5. The Results and Discussion

In this research, correlation between two testing methods namely alternative and traditional testing have been studied by calculating Pearson correlation covariance between them and it has been concluded that there is a very strong correlation between these methods in all levels. The results of the calculation can be seen in the Table 1:

Table 1. The results of Pearson Correlation Coefficient (r) Calculation in all levels

Levels	Average of their Grades		Pearson Correlation Coefficient (r)	Preference between Traditional or Alternative Testing	
	Portfolio	Final	r	Traditional	Alternative
Elementary	12.90	21.10	0,9306	66.6%	33.3%
Intermediate	13.00	19,90	0,9415	33.3%	66.3%
Upper	13,60	20,40	0,9423	00.0%	100.0%

The results of the calculation show that the grades that students received from traditional assessment methods and alternative assessment methods greatly correlate. Although pedagogically they are different and they are created to assess different things that were mentioned earlier, their implications are in the same direction. On the other hand, when we analyze the results carefully, we can see that there are slight differences among the levels. To elaborate, there is a linear increase in the correlation as the level is increased. Upper level which is the top level at this English language school has the most positive correlation while the elementary level had the lowest. Based on these results, it can be inferred that students have gained more skills to obtain, analyze and synthesize knowledge and they have started to achieve better in both of these assessment methods. While students are at elementary level, they are also beginners of language learning process. They may have difficulty in either in alternative testing methods or in traditional testing methods according to their educational and cultural background. They may not be able to master the skills or memorize the necessary vocabulary to get higher scores in exams. In other words, they may not make progress as fast as their peers and learn in their own pace. In another scenario, as students need to set off a new journey with alternative testing methods (namely portfolio in the context of the study) by exploring their skills, strengths, weaknesses and strategies to analyze and synthesize knowledge, it might be a little more difficult for them at first and they might go through a period of adjustment. This is especially true for Turkish students as they are almost never exposed to alternative testing methods in K12 level and it seems brand new to them at university. Because of the possible reasons above, some students might have performed differently in either of these methods particularly in elementary level.

There is also another issue that requires attention in Table 1. Although there are slight differences across levels, all the correlation results are too high. Considering that a number that signals positive correlation is a value between 0 and 1, obtaining a result that is 0,9306, 0,9415 and 0,9423 from elementary to upper respectively is a little bit too high. It almost means that they can be easily used in the place of each other. In the light of the research provided in the introduction part, these results

might be interpreted in two possible ways. First of all, this correlation that is mathematically considered too high might be an indicator of the similarities between these two methods. They might be actually testing the similar skills with the help of diverse methods. Secondly, as traditional assessment methods are what both teachers and students are well adjusted in their teaching and learning experiences, they might be actually shaping the tasks to fit them into traditional and so safe methods of testing. To elaborate, the writing component of the portfolio can be likened to the writing part of the exam while speaking component of the portfolio can be associated with the speaking exams they students take two times in a semester as a part of overall assessment calendar. At this point, the importance of constructing test items becomes even more obvious. Osterlind (2002) notifies that only carefully crafted items on a test decrease the unwanted error variance, or errors of measurement, and so enhances a test score's reliability. This is also valid for alternative assessment methods which are often designed as 'zeal to be 'curriculum-relevant' or 'authentic' or 'realistic.' The designs of such tasks are usually done seemingly without diligent thought to the interpretations that might be obtained from them. Thus, such methods may not give us the real picture of students' progress in a specific field. When we look at the results provided above, we can easily see that the portfolio grade average of the students is too high. It is mainly above 12,5 out of 15. This takes us again to the issues regarding the reliability of the test scores obtained from such scores and to the discussion on the construction of alternative assessment task items. If we only create them to provide students a variety of testing, it is not likely to use them as reliable sources for decision making mechanisms.

When it comes to the preferences, again we can see a linear increase in the preference rates of the students as we go through the upper level. As mentioned in the introduction, alternative assessment has a considerable potential to help students to be autonomous learners and take the responsibility in their learning journey. They get the opportunity to work as partners with their teachers. Looking at the results related to their preferences, it is possible to deduce that students become more independent learners as their level go up, and they start to prefer this kind of an assessment more. This is also because they are likely to have gained more skills to reach any kind of knowledge in the target language easier and quicker than elementary level students. Also, when they study at the upper level, alternative assessment tools are more familiar for them. As they have chance to use their creativity and personalize the topics in such kind of an assessment system, they are more likely to enjoy alternative assessment methods. To explore the implications of the correlation and students' preferences, the researcher of this study conducted interviews with students from each level and their answers were coded. The topics reflected in the interviews by the students and their frequency can be seen in Table 2 below:

Table 2. Results of the interview conducted with students

Benefits of the Portfolio	f	Benefits of the exams	f
Keeping a portfolio helped us to study regularly	13	They motivated us to learn more vocabulary	8
It was much more fun	8	We were tested on what we learned in class	6
Tasks were similar to the ones in real life	8	We did not know the content of the exam, so we had to study a lot to pass	6
Teachers provided detailed feedback that is specific to everyone	7	They test each skill	2
It was not as stressful as exams	7		
It enabled me to understand my mistakes better	5		
We learned real-life English thanks to the tasks	3		
I corrected my mistakes in each task	3		

As seen in Table 2, students indicated many advantages of portfolio for them. As they do tasks often than the exams, they stated that keeping a portfolio promoted regular study among them. This can be regarded as the greatest benefit of the portfolio as regular study is one of the best ways to acquire a new language. On the other hand, students also reported that exams encouraged them to study and learn new vocabulary as their content was unknown to them. Therefore, both testing methods can be said to foster regular study habits.

On the other hand, it is also clear that students mentioned many more benefits of portfolio than tests. Feedback issue that they raised is vital to highlight. Students reported that teachers gave feedback that was specific to every student and his mistakes. This helped them to improve the accuracy of their writing and speaking particularly. This might also prove the positive washback of alternative testing methods on learning. As students feel that they are making progress, they become more and more enthusiastic to learn. Besides, one student reported that actually they had to learn and make progress.

In a multiple choice test, although they do not know the answer you can always circle an option randomly and give an answer. However, this is not possible in the portfolio tasks. It is just like the real life. You do not have any chance to circle something, you have to know and produce. (Student 1)

As students had to do research and analyze and summarize the information that they found, it has been a great opportunity for them to keep studying and practicing. Besides, obviously they believe that they were exposed to the tasks that simulate the ones in real life.

On the other hand, it would be too assertive to completely ignore the benefits of traditional testing. Students also reported that exams kept them motivated to enhance their skills and keep up with the curriculum.

Exams were very effective tools to force us to follow the program. We had no idea about the content of the exams, so we tried our best to learn whatever we can during the semester. I memorized vocabulary items on a regular basis because there was going to a vocabulary pop quiz and we did not know when. Therefore, I believe exams are indispensable for us. (Student 6)

As seen in the quotation from the interview, the student uses the word “indispensable” for traditional testing. This is probably because this type of testing has been a crucial part of their educational experiences since primary school and they are seen as the one and only indicator of the success. However, although this type of testing provides comfort to do students in terms of being tested in the way they are used to, it fails to meet the necessities of a quality assessment method suggested by Herman, Aschbacher, and Winters (1992). To illustrate, they suggest that “ performance-based activities do not constitute assessment per se” and “an integrated and active view of student learning requires the assessment of holistic and complex performance” (p.7). However, traditional testing is solely performance-based and gives us little idea about holistic or complex performance. This takes us back to the discussion on the purpose of testing. Assessment design depends on assessment purpose. While grading and monitoring student progress can be a good reason to test, they are completely different from diagnosis and improvement for more progress. As explained in the introduction part, traditional assessment methods are designed to test knowledge itself while alternative testing methods help students to develop skills on how to get access to knowledge and to recreate it within their context. Therefore, teachers need to make more decisions about which type of method to use in their particular context. This is even more challenging for language teachers who have more decisions to make on this issue (Brown & Hudson, 1998). It is best to note that traditional assessment methods can give quick and objective results on the evaluation of content knowledge. However, the outcome they test is highly irrelevant in real world and might have a negative impact on the instruction. Because testing inevitably effects instruction, teachers are likely to spare a great deal of time for exam practice both due to student pressure and many other factors such as financial concerns of private schools, decision made based on exam scores and the need they feel for testing what they teach. Yet, too much emphasis on exam skills might cause neglect in complex thinking, analyzing, synthesizing and problem solving skills which are much needed in real world (Baker, 1989).

There is a constant search for reliable, valid and objective assessment methods for language teachers. Traditional testing methods are trusted since they have been providing objective information on student success for a very long period of time. This study fills a gap in research showing that this trusted method highly correlates with the new and alternative one. It also justifies other studies indicating the positive effect of alternative assessment tools on students’ learning. To illustrate, Burnaz (2011) underscores that majority of university students favoured portfolio assessment to traditional assessment as participants pointed out that traditional assessment caused time pressure, it led to memorization and it did not evaluate their English speaking skills effectively. Furthermore, the participants of the study expressed dissatisfaction with traditional assessment system and reported that they became more autonomous thanks to the portfolio implementation. Similarly, Goker (2012) affirms that alternative assessment tools such as reflective journals and portfolios can be employed to encourage reflective, self-directed learning to improve students’ competencies in various areas of the language. Efe (2016) supports this finding and adds that reflective assessment can help educators address both cognitive and affective domains. It can even be beneficial to assist students in achieving better in high stakes standardized tests. These studies support the notion that we as educators can benefit from the parallelism shown in the present study and use alternative methods either as integrated to traditional methods or as a replacement of traditional methods. In either of these cases they will provide us much more detailed information about student progress and learning. With the help of them, students will be able to do tasks that represent meaningful instructional activities, approximate real-world application, and construct a product of their own rather than fill in someone

else's work. They will be exposed to multiculturally sensitive tasks. They will have more input on their strengths and weaknesses in various aspects of their skills. Also, as Huerta-Macias (1995) stated students will have nonintrusive tasks which acts as an extension of classroom activities as portfolio tasks are identified as less stressful by the students. Besides, teachers will take over new roles as facilitators of students' new journey rather than the source of content knowledge and they will do the judgement instead of machines. However, research also shows that there is a need to provide more training to teachers on the alternative assessment and its method during both pre-service and in service training as Hatioğlu (2015) indicates that only 1.1% of the students suggested that alternative ways of testing (e.g., portfolios) should be included in the testing course during the interviews which signals a need for raising awareness on the issue in the teacher training programs.

Consequently, the results obtained from the calculation of Pearson correlation coefficient and interviews conducted with students show that students prefer alternative assessment tools more as their level goes up and correlation between traditional method of testing and alternative method of testing is totally positive which becomes even more obvious through the upper level. When it comes to the benefits of tests and portfolio, encouraging students to keep studying and learning is the common benefit of each methods while portfolios have been found less stressful and more efficient in terms of the feedback provided on them by the teachers.

Last but not least, there is one more implication of this study to add related to the relation between these two methods. As they highly correlate, but one of them was reported to have more advantages than the other, maybe now it is time to integrate more alternative assessment tools to our testing system. Alternative assessment can be used more often as a decision making tool and it can be a bigger part of the testing system as it is clear that it has a better washback effect on teaching and learning.

References

- Alderson, J. C. (1991). Language testing in the 1990s: How far have we got? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing* (Vol. 25, pp. 1-26).
- Alderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115-29.
- Alderson, J.C , Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press
- Alderson, J., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*,35(2), 79-113. doi:10.1017/S0261444802001751
- Anderson, D. S., and Piazza, J. A. "Changing Beliefs: Teaching and Learning Mathematics in Constructivist Preservice Classrooms." *Action in Teacher Education*, 1996, 18 (2), 51-62.
- Anderson, R. S. (1998). Why talk about different ways to grade? The shift from traditional assessment to alternative assessment. *New directions for Teaching and Learning*, 1998(74), 5-16.
- Baker, D. (1989). *Language testing: A critical survey and practical guide*. London: Arnold
- Berlak, H. "The Need for a New Science of Assessment." In H. Berlak and others (eds.), *Toward a New Science of Educational Testing and Assessment*. Albany: State University of New York, 1992.
- Bintz, W. P. "'Staying Connected': Exploring New functions for Assessment." *Contemporary Education*, 1991, 62 (4), 307-312.

- Brooks, J. G. "Instructors and Students: Constructivist Forging New Connections." *Educational Leadership*, 1990, 47 (5), 68-71.
- Brooks, J. G., and Brooks, M. G. *In Search of Understanding: The Case for Constructivist Classrooms*. Alexandria Va.: ASCD, 1993.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, 32(4), 653-675.
- Bruner, J. (1986) *.Actual Minds, Possible Worlds*. Cambridge, Mass.: Harvard University Press,
- Burnaz Y.E. (2011). *Perceptions of EFL Learners towards Portfolios as a Method of Alternative Assessment: A Case Study at a Turkish State University*. (Unpublished master's thesis). METU. Ankara, Turkey
- Carlson, D. (1991). Changing the face of testing in California. *California Curriculum News Report*, 16/3.
- Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19,254-72.
- Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education* 11(1), 38-54.
- Cheng, L. (1999). Changing assessment: washback on teacher perceptions and actions. *Teaching and Teacher Education*, 15, 253-271.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. <http://dx.doi.org/10.1037/h0040957> Retrieved from <https://search.proquest.com/docview/614340907?accountid=165780>
- Danon-Boileau, L., & Barbier, D. (2000). *Play-On: Un logiciel d'entraînement a la lecture* ". Paris: Audivi-Media.
- Davison, C., & Cummins, J. (2007). Introduction: Assessment and evaluation in ELT: Shifting paradigms and practices. In *International handbook of English language teaching* (pp. 415-420). Springer, Boston, MA.
- Efe, H. (2016). Reflective Portfolio Assessment in an EFL Context. *The Anthropologist*, 24(1), 157-163.
- Fosnot, C. W. "Constructivism: A Psychological Theory of Learning." In C. W. Fosnot (ed.), *Constructivism: Theory, Perspectives, and Practices*. New York: Teachers College Press, 1996.
- Göker S. D. (2012). Reflective leadership in EFL. *Theory and Practice in Language Studies*, 2(8): 13- 55.
- Haggstrom, M. (1994). Using a videocamera and task-based activities to make classroom oral testing a more realistic communicative experience. *Foreign Language Annals*, 27(2), 161-75
- Harry Torrance (2007) Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. , *Assessment in Education*, 14:3, 281-294, DOI: 10.1080/09695940701591867
- Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Heron, J. "Assessment Revisited." In D. Boud (ed.), *Developing Student Autonomy in Learning*. (2nd ed.) New York: Nichols, 1988.
- Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal*, 5(1), 8-11.
- Hutchings, P. "Principles of Good Practice for Assessing Student Learning." *Assessment Update*, 1993, 5 (1), 6-7.

- Javaherbakhsh, M. R. (2010). The impact of self-assessment on Iranian EFL learners' writing skill. *English Language Teaching*, 3(2), 213.
- Johnston, P. H. *Constructive Evaluation of Literate Activity*. White Plains, N.Y.: Longman, 1992.
- Lewkowicz, J. (2000). Authenticity in language testing: some outstanding questions. *Language Testing*, 17(1), 43-64.
- Linn, R. L. (1994). Performance assessment: Policy, promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14
- Marino, R. P., Pickering P., and McTighe, J. *Assessing Student Outcomes*. Alexandria, Va.: Association for Supervision and Curriculum Development, 1993.
- Morehouse, R. E., & Maykut, P. (2002). *Beginning qualitative research: A philosophical and practical guide*. Routledge.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-56
- Michaels, W., and Karnes, M. R. *Measuring Educational Achievement*. New York: McGraw-Hill, 1950.
- Moeller, A.J. & Reschke, C. (1993). A second look at grading and classroom performance: report of a research study. *Modern Language Journal*, 77(2), 163
- Osterlind, S. J. (2002). *Constructing test items: multiple-choice, constructed-response, performance and other formats*. Boston, USA: Springer Publishing
- Piaget, J. *The Science of Education and the Psychology of the Child*. New York: Basic Books, 1970.
- Raven, J. "A Model of Competence, Motivation, and Behavior, and a Paradigm for Assessment." In H. Berlak and others (eds.), *Toward a New Science of Educational Testing and Assessment*. Albany: State University of New York, 1992.
- Roderick, J. A. (1991). *Context-responsive approaches to assessing children's language*. Natl Council of Teachers.
- Richards J.C. & W.A. Renandya. (eds.). 2002. *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge: Cambridge University Press
- Rodgers, J.L. and Nicewander, W.A. (1988) Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42, 59-66. <https://doi.org/10.2307/2685263>
- Sessions, R. "Education is a Gift, Not a Commodity." Paper presented at the National Conference of the Community Colleges Humanities Association, Washington, D.C., November 1995.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, 29(7), 4-14.
- Short, K. G., and Burke, C. *Creating Curriculum: Instructors and Students as a Community of Learners*. Portsmouth, N.H.: Heinemann, 1991.
- Vygotsky, L. S. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, Mass.: Harvard University Press, 1978.