THE EFFECT OF ADDING A TASK-BASED SPEAKING TEST ON THE PREDICTIVE POWER OF THE CURRENT ENGLISH FOR ACADEMIC PURPOSES PROFICIENCY TEST AT METU NCC SCHOOL OF FOREIGN LANGUAGES

A THESIS SUBMITTED TO
THE BOARD OF GRADUATE PROGRAMS
OF
MIDDLE EAST TECHNICAL UNIVERSITY, NORTHERN CYPRUS CAMPUS

BY

REZA NEIRIZ NAGHADEHI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE
DEGREE OF MASTER OF ARTS
IN
THE
ENGLISH LANGUAGE TEACHING

JULY 2016

Approval of the Board of Graduate Programs

____

Prof. Dr. Tanju Mehmetoğlu

Chairperson

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Arts.

____

Assist. Prof. Dr. Ali Fuad Selvi

Program Coordinator

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Arts.

____

Assist. Prof. Dr. Mary Ann Walter

Supervisor

Examining Committee Members

Assoc. Prof. Dr. Bilal Kırkıcı ____________________
(METU, Foreign Language Education)

Assist. Prof. Dr. Mary Ann Walter ____________________
(METU NCC, Teaching English as a Foreign Language)

Dr. Elvan Eda Işık Taş ____________________
(METU NCC, School of Foreign Languages)

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Reza Neiriz Naghadehi

Signature:

ABSTRACT

THE EFFECT OF ADDING A TASK-BASED SPEAKING TEST ON THE PREDICTIVE POWER OF THE CURRENT ENGLISH FOR ACADEMIC PURPOSES PROFICIENCY TEST AT METU NCC SCHOOL OF FOREIGN LANGUAGES

Neiriz Naghadehi, Reza
Department of English Language Teaching
Supervisor: Assist. Prof. Dr. Mary Ann Walter

July 2016

I this thesis, a proposed computerized task-based speaking test (PTBST) was designed based on the communicative language ability model proposed by Bachman (1990), and its effect on the predictive power of the academic English proficiency exam (EPE) administered at the Middle East Technical University, Northern Cyprus Campus (METU NCC) as an English proficiency screening criterion was examined.

Two important constructs for academic success were operationalized in PTBST not covered by EPE: the ability to synthesize written an aural stimuli into spoken responses and presenting the resultant synthesis orally in a comprehensible and fluent fashion. Correlation, regression, and factor analyses were conducted to explore the predictive power of PTBST and EPE as well as to examine the construct exclusiveness of these two tests.

The results showed a better correlation of PTBST with the GPAs of non-engineering students compared to METU EPE. PTBST also showed to have a potential to exhibit the same results for engineering students. Moreover, an exploratory factor analysis showed that the PTBST measures a different construct compared to METU EPE, which justifies adding PTBST to it. Also, the results of this study corroborated the correlations reported for EPE and GPA in METU Ankara. In addition, two distinct types of synthesizing were found and showed a different pattern of correlation with engineering and non-engineering students. Finally, it was found that it might be a better practice to report proficiency tests with different weightings of its sections, giving higher weightings to the sections more relevant to engineering vs. non-engineering disciplines.

Key words: *English for academic purposes, language proficiency tests, academic achievement prediction studies, task-based speaking tests*

To My Beloved Wife and Always Supportive Parents

ACKNOWLEGEMENTS

I would like to express my deepest and most sincere gratitude to my supervisor, Assist. Prof. Dr. Mary Ann Walter, without whose guidance, support, and insightful comments and criticism the completion of this thesis would not have been possible.

I would also like to express my thanks and gratefulness to the examining committee members, Dr. Elvan Eda Işık Taş and Assoc. Prof. Dr. Bilal Kırkıcı, for their constructive and insightful suggestions and comments which made a great contribution to this study.

Furthermore, I would like to thank all the participants who agreed to take part in this study and made this research possible.

Last, but not least, my special thanks goes to my lovely wife whose patience and support were the greatest sources of my inspiration, tenacity, and passionate work in this study.

TABLE OF CONTENTS

CHAPTERS

TABLE OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ASP.Net	Active Server Page .Net
CEFR	Common European Framework of Reference
CI	Confidence Interval
Coef.	Coefficient
Corr.	Correlation
EIL	English as International Language
EPE	English Proficiency Exam
ETS	Educational Testing Service
FCE	First Certificate in English
FOR	Frame of Reference
GPA	Grade Point Average
GRE	Graduate Record Examinations
IELTS	International English Language Testing System
IRT	Item Response Theory
METU	Middle East Technical University
METU NCC	Middle East Technical University, Northern Cyprus Campus
MFRM	Many-Facet Rasch Measurement
Mn. Sq.	Mean Square
MS SQL	Microsoft Structured Query Language
NES	Native English Speaker
NNES	Non-Native English Speaker
PTBST	Proposed Task-Based Speaking Test

SFL	School of Foreign Languages
Std.	Standard
TED	Technology, Entertainment, and Design
TLU	Target Language Use
TOEFL BPT	Test of English as a Foreign Language Paper-Based Test
TOEFL iBT	Test of English as a Foreign Language Internet-Based Test
TSE	Test of Spoken English
VB	Visual Basic
VIF	Variance Inflation Factor

CHAPTER 1

INTRODUCTION

De signing and compiling an English language test (or any test for that matter), despite its straightforward goals, has proven to be a controversial and complicated task. Traditionally, the question of what to be tested, or more specifically, what language constructs needed to be tested was answered simply with whatever possible to be measured psychometrically. The traditional paper and pencil medium of language assessment dictated the replacement of proxies for the actual constructs to be tested. For instance, in the Paper-Based Test of English as Foreign Language (TOEFL PBT), test of grammar knowledge was used as a proxy for the communicative abilities of test-takers. In other words, as Bachman and Palmer's framework puts it (as cited in Jones, 2012), the demonstration of grammatical knowledge through mainly discrete-point items was interpreted as the ability of the test-taker to fluently communicate in English. They posit that language knowledge and strategic competence contribute to the production of the language appropriate to the situation. Taking language knowledge as an essential part of the equation, test items that elicit such knowledge conditioned by the communicative situation present in that item mediated by strategic competence can be a token of test-taker's language ability. In other words, as long as a test-taker knows that a present perfect tense is needed for an item (using strategic competence) and produces the right syntax of it (using language knowledge), enough evidence for the communicative competence of the test-taker is obtained (Douglas, 2000, p. 28).

The test administered at some English-medium universities in Turkey and Northern Cyprus is a paper-and-pencil based one following, perforce, the old traditions of English language assessment, i.e. relying heavily on linguistic knowledge as a proxy for communicative abilities. This choice stems from practical considerations. To illustrate, due to impossibility of sampling oral production (which is a closer approximation of actual language use) in a paper-and-pencil test, the performance of the test-taker in the language section, which mainly consist of items eliciting grammar and vocabulary knowledge, is considered to be a token of that test-taker's communication proficiency in the target language. Passing such gate-keeper tests provides the students with a passport to enter their faculties and pursue their education at their respective academic programs which follow English-medium curricula. Yet, the concerns regarding the ineffectiveness of English-medium education and the choice of many universities to offer their courses in Turkish, rather than English (Kirkgoz, 2007), can be attributed to the fact that these tests are not powerful enough to filter out the students who have caused such concerns due to their inadequate ability to succeed in an English-medium tertiary program. One reason for the mentioned inability of the proficiency tests might be due to the fact that they assess *grammatical knowledge* rather than *grammatical ability*. As Purpura (2004) puts it, the latter refers to the ability of a language learner to use the grammatical knowledge for communicative purposes while the former taps just into the learners' declarative knowledge. As the former does not necessitate the latter (Weir, 1990), resorting to the declarative knowledge or grammar knowledge through discrete-point test items as an indication of procedural knowledge or grammar ability is unwarranted. Moreover, the wash-back effect of these tests can lead the curriculum of the English preparatory courses offered at these universities toward the teaching of grammatical knowledge instead of grammatical ability.

Another concern with some English proficiency tests pertains to their treatment of the listening construct through multiple-choice items. Freedle and Kostin (1999), in a study of multiple-choice listening items, argue that there is only a modest amount of evidence to support the construct validity of these items. In other words, it is not the listening passage which accounts for the item difficulty in 67% of the times. Therefore, it is not safe to assume that a certain score, e.g. 60, can become the baseline to separate failing and passing test-takers. With this rate of construct validity, a score of 60 in listening section can mean the ability of the test-taker to comprehend the listening passage at any level between 20 and 60 percent. Yet, listening comprehension is not the sole concern in academia. Ferris and Tagg (1996) report that most university instructors are concerned with the lack of evidence in their students' listening comprehension ability judging by the paucity of oral responses and reactions by their students during lectures and seminars. This might be attributed to the cultural concerns as the authors point out, but listening comprehension does not happen in a vacuum and is not valued in and of itself in academia unless it contributes to reactions and involvement of students in some form, either in oral or written, which brings the whole matter to the next issue, namely speaking.

Lack or inappropriateness of speaking assessment in the current testing system of English preparatory courses in Turkish and Northern Cypriot universities is another issue which merits attention (Hughes, 1988). Inappropriateness of these speaking tests stems from the issues of construct validity. Since tests takers are required to speak mostly about academic material they have already read or written once they enter their faculties, testing speaking based on the elicitation of candidates' ideas on general and every-day matters cannot fully represent speaking tasks which test-takers are expected to perform at their faculties (Brown, Iwashita, & McNamara, 2005). In other words, synthesizing, which is an integral part of many oral communication occurring in university classes, is ignored for the most part in these tests. While these two constructs together are vital academic skills necessary to the delivery of talks, presentation, and asking questions, the proficiency test, as a case in point, administered at METU NCC assumes that the demonstration of grammatical knowledge, reading comprehension skills, and written form of response to communicative situations as well as some short oral answers to questions about every-day matters can provide the necessary evidence of the test-taker's ability in oral communication for academic purposes. One reason for this line of action might stem from the practicality issues. Test of speaking for academic purposes is conducted by ETS in TOEFL iBT. The assessed construct for this test is operationalized as the ability of the test-taker to synthesize the written and spoken academic or social material into a spoken summary, along with the ability to voice personal thoughts (Alderson, 2009; Butler, Eignor, Jones, McNamara, & Suomi, 2000). This test assesses test-takers' synthesizing abilities along with their reading, listening, and writing skills in addition to speaking in an attempt to capture all the required constructs. This is a more appropriate approach than traditional tests of language which separate skills because the academic language capabilities are not isolated from other skills and do not occur in a vacuum. Therefore, this test has managed to measure the speaking ability of test-takers in a more realistic and communicative manner.

At METU NCC SFL, Speaking assessment is conducted through video projects, classroom presentations, and formal speaking exams. The first two are carried out throughout the year, and are formative assessments of speaking. In both of them, a list of topics is given to students. In video projects, students are asked to record a short video of themselves talking about their topic of choice for the video project. However, for presentations, students are given feedback on the content and layout of their presentation before they deliver it to their classmates, and then are judged based on the quality of their presentation. In both cases (i.e. video projects and presentations), the teacher uses a set rubric to grade their speaking, and there is no standardization and rater training involved. The formal speaking exam, however, is an achievement test. It is in the form of an interview with two examinees and two examiners. Each of the two examinees first talk to one of the examiners while the other examiner grades the speaking. Then, the two examinee speak to each other while one examiner facilitates the communication and the other grades the performance of the two examinees. There is no standardization involved in this form of assessment either, and only the two examiners, who are also the examinee's teachers, grade the performance using set rubrics and in a consensus-based way. The grade of this test is added to other grades from other achievement tests to decide if a student can take the exit exam which is a proficiency test. Apart from many limitations concerning interviews as speaking tests, which will be discussed under *Why a Computerized Speaking Test*, this test is not part of the exit exam which is an English proficiency test, neither are the videos and presentations. Moreover, there is a lack of integrating in these speaking tests through input, and are mainly about the personal ideas of test-takers on everyday topics, which, according to Brown, Iwashita, and McNamara (2005) do not represent the construct of academic speaking adequately if there is no input and synthesizing of that input involved.

This study aims at looking at the degree of change brought about by adding a task-based speaking section in the predictive power of the current proficiency test administered at METU NCC in terms of the academic success of the test-takers at Middle East Technical University Northern Cyprus Campus (METU NCC), which is an English-medium university offering undergraduate and graduate courses through English. The study hypothesizes that the proposed section will increase the predictive power of the current test regarding student success in their first semester of studies at their faculties indicated by their GPAs, holding that the added section will cover two important constructs (i.e. synthesizing and speaking ability) with close approximation of the skills and tasks required of university students — constructs that are not covered by EPE. To test this hypothesis, the predictive power and construct exclusiveness of PTBST in comparison to different parts of EPE were analyzed through correlation, regression, and factor analyses. For the correlation and regression analyses, GPAs and weighed GPAs of the participants were used as the indicator of academic success in the first semester of their studies in their respective disciplines. It must be noted that the weighted GPA is the GPA of students after taking out the English 101. Below, the reason for the use of this type of weighted GPA will be explained thoroughly, but, in general, this type of GPA is important to obtain the predictive power of both PTBST and EPE regarding the success of the participants in courses that do not rely purely on English and are more characteristic of their subject areas. Finally, due to the difference in the degree to which engineering and non-engineering disciplines rely on English proficiency and linguistic medium of education, the predictive power of PTBST and EPE were studied separately for these two groups of disciplines.

CHAPTER 2

REVIEW OF LITERATURE

2.1 The Role of Task-Based Speaking in Proficiency Tests

A language proficiency test assesses the “global competence in a language” with no specific course as its content (Brown, 2004). In other words, no specific course could be considered as its content source. Another quality of proficiency tests is their norm-referenced and summative nature (Brown, 2004). That is to say, this kind of language test defers to a certain set of norms in a language (e.g. British or American English), and the reported scores for this kind of test are equalized and reported in a single numerical score. Therefore, they are not inherently designed to provide detailed diagnostic feedback and are meant to label a candidate with the degree of competence in the language assessed in the test.

A proficiency test needs to accommodate communicative competence, which in fact pertains to the language use ability rather than language knowledge (Taylor & Angelis, 2008). Hence, task-based language testing has been implemented to accommodate this feature of the language. This method of testing proficiency is goal-oriented and content-focused, have real outcomes, and reflect real-life language use (Shehadeh, 2012). For example in TOEFL iBT, in speaking question number 4, a candidate has to read a short text defining an academic concept and listen to a professor illustrating that concept through examples. This speaking test item expects test-takers to demonstrate their ability in synthesizing academic content into a spoken response, so it focuses on an academic content area, and reflects the real-life needs of university students who are expected to carry out similar tasks throughout their studies. Ellis (2000) believes that performance on task-based assessment is the closest approximation to that in the real world. Also, Shehadeh (2012) posits that task-based assessment is formative, performance-based, direct, and authentic. In other words, the very assessment lends itself to promoting further development in test-takers. It also focuses on the performance and language use rather than language knowledge, without asking test-takers to demonstrate their grammar or vocabulary knowledge in a vacuum.

Moreover, it does not entail making inferences regarding test-takers' language ability (as traditional language tests do). For instance, the paper-based version of TOEFL had grammar items consisting of sentences with four underlined phrases or words. Test-takers had to choose the one out of those four parts of the sentences which was grammatically wrong (Lim, Kurtin, & Wellman, 2001). Therefore, the score reported for this part of the test was based on the inference that if a candidate knows what is grammatically right (or wrong for that matter) is also able to use the language successfully. Criticism regarding this approach is pointed out by Larsen-Freeman (as cited in Graham, 1987, p. 514). She believes that test items must give a chance to test-takers to demonstrate what they know and can do with the language rather than being punished for their mistakes which is taken to be an indication of what they don't know, and hence a measure of their proficiency in the language. Finally, authenticity of a proficiency test can be substantiated by eliciting and expecting performance relevant to the context a test-taker would be operating in linguistically (Shehadeh, 2012; Butler, Eignor, Jones, McNamara, & Suomi, 2000). Hence, a test which draws heavily, for instance, on fill-the-gap question items has little compatibility with what a university student will be doing during their studies when comprehension, synthesizing, and production of the language with a focus on content and outcome make up the majority of the linguistic functions.

One example of the inclusion of task-based assessment in an English proficiency test is the speaking part in TOEFL iBT, which is founded based on an extended version of communicative competence and communicative language ability (Alderson, 2009; Chapelle, Grabe, & Berns, 1997; Brooks & Swain, 2014). However, like many other forms of assessment, this test has its own share of criticism. The monologic nature of speaking tasks imposed by the tests administered through computers is found to affect the validity of the test, which in turn fosters reservations in extending the validity of the test to real-life communicative situations (Butler, Eignor, Jones, McNamara, & Suomi, 2000). Brook and Swain (2014) found that the grammatical, discursive, and lexical features of the speaking performance on the speaking section of the TOEFL iBT are different from those in the communication happening inside and outside classrooms within academic perimeter. Yet, Michel, Kuiken, and Vedder (2007) report that monologue speaking tasks contain linguistically more complex features than dialogues. In a sense, inclusion of such tasks in a proficiency test can elicit a linguistically more complex sample of test-takers' oral production providing a better measure of maximum ability of the test-taker in the skill and task in question and yielding scores which cover a broader spectrum of ability levels. Therefore, a considerable cutback in terms of time and resources can be achieved by directly aiming at the highest level of linguistic ability of a candidate rather than tiring them with a long test which in turn introduces the extraneous element of fatigue into the performance and the assessment.

Another important aspect of the inclusion of task-based speaking section in a proficiency test is its washback effect. Wall and Horak (2011), in a study of the washback effect of TOEFL iBT, report that although the course book seemed to be the ultimate mediator of language teaching focus, one of the three participating teachers made a dramatic transition from a teacher-centered, grammar-based instruction to a more student-centered, communicative style as a result of the washback effect after the introduction of speaking test into TOEFL iBT. In the same vein, Byrnes (2002) argues that the inclusion of tasked-based assessment in their foreign language program proved to be a valuable contribution to the development of specific curriculum, provided further insights into the expectations regarding the performance of students on the part of the faculty, and resulted in more effort by the students to develop their production in the target language. Also, van den Branden, Depauw, and Gysen (2002) argue that a careful development of a task-based test based on the real needs of the learners can empower any educational system to accomplish more student learning. Shehadeh (2012), also asserts that the reason high priority is given to speaking and task-based assessment in high-stake tests is that it mitigates the inclination in teachers to teach for tests. Teaching for the test is not an uncommon and unjustified practice when the task-based nature is absent in a language proficiency tests. For instance, Cohn and Upton (2006) report that the verbal protocols elicited from the candidates taking TOEFL iBT reading section covered only three of the 28 reading strategies that test designers hoped to invoke in the candidates, while the use of 20 out of 28 test management strategies by the test-takers attests to the fact that this section of the test measures test management construct rather than reading strategies. This is an unfortunate washback effect on English preparatory courses which pushes the nature of teaching away from communication with language use as its goal and converts it into a test strategy crash course. Apart from having a more positive washback effect, inclusion of a task-based speaking component could also remedy the problem of noninclusive construct in proficiency exams testing the reading and listening proficiency of test-takers through discrete-point or multiple-choice questions.

2.2 English Proficiency as a Predictive Factor of Academic Success

One of the most common uses of English proficiency tests pertains to admission decisions of undergraduate and graduate programs. Administrators and stake-holders use the scores of these tests as one of several criteria to decide who has the ability to succeed in an English-medium university. Yet, the question is whether these tests and their reported scores have enough predictive power to make these tests worthy gate-keepers. Several studies have been carried out to investigate the predictive power of these tests; however, there are some issues that must be taken into consideration before using these tests scores as a token of a student's ability to succeed in an English-medium university.

Academic success is affected by many more factors than simply language proficiency (Graham, 1987). Oliver, Vanderford, and Grote (2012) ascertain that there are not enough controlling techniques and ways to capture non-linguistic factors contributing to academic success or failure. Yet, research in this area has used the first semester GPA and its correlation with proficiency test scores to establish the predictive power of language proficiency tests regarding academic success (Graham, 1987; Enginarlar, 2006; Enginarlar, 2007; Enginarlar, 2009; Enginarlar, 2012). Moreover, according to Oliver, Vanderfrod, and Grote (2012), internationally recognized commercial proficiency tests like IELTS and TOEFL seem to be more strongly correlated with academic success. However, in the case of a lower correlation between first semester GPA and commercial tests like TOEFL iBT, disregarding the test as a whole solely due to this low correlation can be counter-productive and might warrant a closer examination. As a case in point, Cho and Bridgeman (2012) showed that although the correlation between TOEFL iBT scores and first-semester GPA of 2,594 participants was low, it was a meaningful relationship. In other words, using expectancy graphs, the researchers found that participants who fell in the top quartile of GPA had received a high TOEFL iBT grades, while only a small percentage of those in the bottom quartile of GPA had received a high TOEFL iBT score. Hence, looking at the correlation between the scores obtained from a proficiency test and first semester GPA seems to be a valid method of analysis at the moment. But, what if we could pinpoint those who fail academically only because of a lack of sufficient English proficiency?

Imagine that a department would allow all the applicants, regardless of their language proficiency, enter their departments. Then, the GPAs of these students in the first semester would show a wider range. In other words, some would fail and some would pass the semester. Ergo, with this wider range of GPAs, one could easily probe if the failure was due to English proficiency or not, while the opposite is not easy to prove statistically due to the ceiling effect. That is to say, if there were 50 participants in a study including those allowed to enter their respective faculties regardless of their low scores on a language proficiency test, and 20 of them failed their first semester, a comparison of proficiency scores would more readily show if the participants who have failed in their first semester and had received a low score are statistically different from those who have passed their first semester and had received a higher score in their proficiency test. A statistical test of group difference, e.g. t-test, would be more apt than a correlation test in such a scenario. If the results show that those who failed had consistently received a low score in an English proficiency test, the significance of proficiency scores in a participant's academic success could be established. This is ratified by Graham (1987). As another support for this argument, Oliver, Vanderford, and Grote (2012) found that there is a higher correlation between IELTS or TOEFL scores and academic success compared to other methods of evidencing English language proficiency, like in-house proficiency tests or completion of English courses. Because the population in their study included those with limited language proficiency, using Pearson's Product Moment Correlation Coefficient analyses, they could easily show the correlation between academic success and language proficiency evidenced by TOEFL or IELTS. Reporting two other studies, Graham (1987) also provides further evidence for this argument. She points to two studies, one of which reports a higher correlation between the proficiency test scores and first-semester GPAs compared to the other one. The mean score of the proficiency test reported for the participants in the former study is lower than that in the latter. Hence, the author concludes that admission of students with lower English proficiency scores helped to pinpoint those who failed their semester mainly because of low English proficiency. She continues that the latter study showed a lower correlation between English proficiency scores and first-semester GPAs only because all the participants had already achieved a high level of proficiency and it could not have played a strong role in the academic success or failure of the participants.

Another factor which must be taken into consideration is the fact that GPA is not a good indicator of academic success. One reason for this is that factors other than English proficiency, or any ability of a student for that matter, can render GPA non-representative of student abilities. One such factor is teacher sympathy, which can distort the correlation between English proficiency and academic success unreliable (Heil & Aleamoni, 1974). However, factors like this are not possible to control for and are sure to contribute to lower correlation between an English proficiency test scores and first semester GPAs. Nonetheless, there are other distorting factors that can be controlled for. For instance, Graham (1987) believes that GPA should be replaced by the actual amount of academic work which is completed by a participant. Another reason why GPAs must be used with caution in predictive power studies is that, GPA is an average of total grades obtained for a combination of courses which weigh linguistic and non-linguistic medium of learning differently. For instance, a chemistry final exam score would depend on English proficiency up to a certain level when formulas and mathematical calculations are not required. Burgess and Greis (1970) concluded that, when taking the courses which do not need much English (like art or mathematics) out of GPA, the resultant weighted GPA correlated more strongly with the TOEFL scores of the participants. This means that in a study of correlation, which examines the role English proficiency plays in academic success, care must be taken to control for the variables which can distort the sought-after correlation.

While courses with low dependency on linguistic factors like mathematics can distort the correlation between an English proficiency test and first semester GPA, the presence of courses which solely depend on English proficiency and nothing else also have the potential to distort this correlation. To illustrate, if a considerable part of GPA is accounted for by English 101 course (4 units out of 13), the high correlation between the English proficiency test and GPA might be due to the effect of the English 101 course. While the former case is reported in the literature to yield higher correlations between English proficiency test scores and GPAs (Burgess & Greis, 1970), the latter has not been proved as far as this study’s review of literature goes. Therefore, studying the correlation of an English proficiency test with the first GPA after taking English 101 out of it can yield a better picture of the predictive power of the proficiency test. The resultant picture can be informative in that it can show how successful the test construct is in capturing some non-linguistic factors in combination with linguistic ones, which are closely related to the language and academic studies, e.g. synthesizing information.

Closely related to the difference in correlation of English proficiency tests with GPAs based on the degree to which courses contributing to those GPAs rely on English or not is the comparison of the predictive power of English proficiency test for engineering and non-engineering disciplines. In other words, engineering and non-engineering disciplines have courses which generally differ in terms of their reliance on English. Therefore, it might be expected to find a higher correlation between language proficiency and non-engineering GPAs, in comparison to engineering GPAs, as the former rely more on linguistic factors. As a case in point, Wait and Gressel (2009) found that TOEFL scores have a better correlation with the GPA of non-engineering students than those of engineering. However, this does not mean that English proficiency is completely irrelevant to non-Engineering disciplines. In fact, Vinke and Jochemes (1993) found that a certain band of English proficiency measured by the TOEFL PBT predicts the academic success of one engineering discipline, i.e. sanitary engineering, but any score below or above this band loses its predictive power. Moreover, Ayers and Quattlebaum (1992), as cited in Cho and Bridgeman (2012), found that the scores of the quantitative section of GRE (Graduate Record Examinations, created and administered by Educational Testing Service, USA), had a more predictive power of engineering academic success measured by GPA than its verbal sections or the TOEFL scores. Cho and Bridgeman (2012) also report that TOEFL iBT scores correlate better with the GPAs of business, humanities and arts, and social sciences than those of sciences and engineering. With these results reported in the literature, it is worth looking at the nature of correlation between EPE and PTBST with the GPA of engineering and non-engineering students to see if the same pattern also exists in the context of METU NCC.

Furthermore, it is worth looking at the predictive power of the sub-sections of proficiency tests regarding academic success as these test different constructs and can help us get a clearer picture of the predictive power of proficiency tests for different disciplines before discarding them based on their lack of or low predictive power for certain disciplines as whole. For instance, according to Wait and Gressel (2009) literature students, as a subset of non-engineering cohort, spend more time writing in academic life than their engineering counterparts for whom laboratory reports, solving problems, and dealing with quantitative and numerical values are more common types of academic tasks. Therefore, they believe that the writing section of a proficiency test would have less predictive power regarding the academic success of engineering disciplines. Also, Al-Musawi and Al-Ansari (1999) found that the cloze test section of FCE was one of the two test sections that could predict the academic success of English major students measured by their GPA. These results show that investigating the predictive power of proficiency test sub-sections can yield clearer pictures of the predictive power of proficiency tests for different disciplines and can help test developers and tests users to adopt a more discipline-specific approach to test development and test use. This is further corroborated by Bachman (1991) who believes that both task content and task method of a test task are major players in deciding the resultant performance of test-takers. Therefore, with the results obtained from analyzing different sub-sections of EPE and PTBST tasks and the pattern of their predictive power for different disciplines, tests with more predictive power can be designed to reflect the common task types and task contents of engineering and non-engineering disciplines and, hence, to exhibit a higher predictive power for both groups of disciplines.

2.3 Studies on the Current METU EPE

Enginarlar (2012) has conducted an inter-componential correlation analysis to see how different sub-sections of the current proficiency test correlate with one another in METU Ankara. According to this report, note-taking and writing sections show a lower correlation with the other sections of the test, and these two sections are reported to demonstrate a weak relationship with the total score. He also reports the correlation of each component or section of the test with the total score through subtracting the score of the component from the total score and calculating a Pearson Correlation. In another report, Enginarlar (2009) provides the correlation analysis data conducted on the English proficiency exam’s total scores, first-semester GPA, and English 101 (an English for Academic Purposes course offered at SFL). He reports a high correlation of 0.47 for all the faculties in general. The same index is reported to be 0.444 in Enginarlar (2007) and between 0.3 and 0.4 in Englinarlar (2006). Enginarlar (2012) summarizes these reports and indicates that the correlation between EPE and first semester GPA ranged between .45 and .55. All these reports show an acceptable level of predictive validity of the current English proficiency test administered by METU

SFL. Apart from the similarities in procedures and the identical English proficiency tests, there might be differing conditions in Ankara campus and METU NCC due to student population admission criteria, which warrants similar analyses of the current EPE. However, conducting correlation tests is simply not enough. As a case in point, it does not show how much of variance in the first semester performance (GPA) EPE actually accounts for. So, correlation analyses can be complemented by other tests to gain further insight into the nature of relationship between METU EPE and first semester GPA.

Due to the similarity of the context, looking at other studies of predictive power of English proficiency tests conducted in Turkey and Northern Cyprus outside METU is also of value. As far as the literature review of the present study goes, two such studies have been found, but more might have been conducted which are either not published or were not accessible to this researcher. Yapar (2003) is one such study. He found that using item information scores through item-response theory analyses, which take into account both the item difficulty and test-taker ability to give a better picture of the actual abilities of test-takers, yield better estimates of correlation and predictive validity for proficiency tests than raw scores. Of course, this is only possible if the information related to all items of a test for all the participants is available. Otherwise, conducting such analyses is impractical. In another study, Aydın (2012) found that higher English proficiency, along with better communicative capabilities, positive self-concept of English, and higher levels of stress contributed to better performance in the English proficiency test they took, but this study does not report the predictive power of these factors regarding the academic success after these students enter their studies in their disciplines.

2.4 Research Questions

Taking into consideration some important factors offered in the literature regarding English proficiency tests, their predictive power, and communicative and task-based tests, the present study seeks to answer the following questions. These questions will guide the analyses and the subsequent discussion on the results, and will help examine the predictive power of METU EPE and its sub-sections along with that of PTBST and its tasks.

1) How powerful is the current proficiency test administered at METU NCC SFL in predicting the success of the students who obtain a score above the cut-off point to enter their departments at METU NCC?

   (a) Is there any section of the test which has more predictive power regarding the success of the test-takers at their respective department?

2) How does the predictive power of PTBST compare to that of METU EPE and its sub-sections?

   a) Do PTBST tasks measure a different construct than those sections already in the current proficiency test?

3) Do PTBST and its tasks along with EPE and its sub-sections show a different predictive power pattern for engineering and non-engineering students?

CHAPTER 3

PROPOSED TASK-BASED SPEAKING TEST

For this study, a computerized task-based speaking test has been developed following the latest findings reported in the literature to ensure the validity of the test. In this section, various aspects of this component will be discussed. First, the reason why a computerized speaking test is chosen rather than a traditional face-to-face interview will be probed. Then, the design and the underlying construct of this test will be delineated. Finally, the factors and criteria of the proposed task-based speaking test (PTBST) will be explained.

3.1 Description of the Proposed Task-Based Speaking Test Tasks

Before starting the next sections, I will give a description of the tasks in PTBST. A complete description of the test will be given in the methodology section; however, being familiar with PTBST and its tasks is necessary since references will be made to them in the explanation of the test construct and rubrics development.

This test is composed of three integrated tasks. in other words, all the tasks integrate more than one language skill. In the first task, the test-taker watches and listens to a short lecture by a professor at the first session of a psychology course. Before the task itself, a complete instruction to the task is shown in the written form to the test-taker and a voice-over reads through it. Then, a professor talks about the course syllabus and requirements of psychology 101 course. Immediately after that, the test-taker is given two minutes to prepare to call his friend who was absent at that session and leave a voice-mail summarizing what the professor said. The test-taker is encouraged to take notes while listening and use them while preparing and answering the question. Immediately after the two-minute preparation time is over, the test-taker is prompted to start speaking to the microphone and leave the message. Both during preparation and speaking, a count-down timer shows how much time is left so that the test-taker can make necessary adjustments to accommodate the time limitation. In this task, listening (lecture video), writing (note-taking), and speaking are all integrated. Brooks and Swain (2014) report that there are grammatical, discoursal, and lexical differences between in-class and out of class speaking. Therefore, this activity is included to elicit speaking from this aspect (out-of-class) of academic life.

Immediately after the speaking time is over, the instructions to the second task appears, and the test-taker can read and listen at the same time to the voice-over which reads out the instructions. Then, a reading passage about three different advertisement techniques appear with a count-down timer. The test-taker is given five minutes to read the passage. Immediately after the reading time is over, a video player window appears on the screen, and a professor gives examples for each of the advertisement techniques to clarify the concept. Meanwhile, the reading passage is still available on the screen, and the test-taker can refer to it at any time needed. This resembles a lecture situation in which the professor provides further explanation about the concepts in a textbook or a handout when students can always refer to their book or handout if necessary. Again, the test-taker is encouraged to take notes while reading and listening and use them during the preparation and speaking. Immediately after the lecture finishes, a new screen with the prompt appears and asks the test-taker to explain the three different advertisement techniques she or he has just been presented with using the explanation in the text and the examples provided by the professor. This tasks integrates reading (text about advertisement), listening (lecture), writing (note-taking), and speaking skills. It also elicits a synthesis of information from two different sources, namely the reading passage and the lecture. Like the first task, the test-taker has two minutes to prepare his or her response and two minutes to speak. Bachman (1991) believes that both task content and task method of a test task are major players in deciding the resultant performance of test-takers to ensure the relevance of the test to the target language use (TLU) or criterion situation where the abilities measured in the test are required for a successful operation of the test-taker outside the test situation. Although description of advertisement techniques as the task content might be more suitable for business majors and somehow relevant to the other non-engineering disciplines, the task method which involves a direct reporting of the information from input might be more characteristic of engineering fields. One example of this is the laboratory reports that engineering students write (Wait & Gressel, 2009). Therefore, pattern of correlation of this task with the GPAs of engineering and non-engineering students can give interesting and informative picture regarding the importance of task type and task content.

After the speaking time for the second task is over, the test-taker is shown the instructions to the third task. In this task, the test-taker is asked to choose one of the three advertisement techniques mentioned in the second task which convinced them to buy a product in the past. The test-taker is asked to give the account of the experience while explaining how the technique convinced him or her to buy the product. As in the previous two tasks, the test-taker is given two minutes to prepare his or her response and two minutes to speak. This task also integrates speaking with reading, listening, and writing and encourages synthesizing personal account with a concept presented in the form of college reading and classroom lecture. In terms of both task content and method, the third task seem to be more relevant to non-engineering students as the input is the same as the second task (and it was already discussed that the content of the second task can be more relevant to non-engineering disciplines), and the method requires a combination of direct reporting and an addition of personal opinion. This is because discussing personal ideas is more common in non-engineering than engineering disciplines. Therefore, this task seems to be more relevant to non-engineering disciplines, and this will be attended in the analyses and interpretations of the results. In addition to task content and method, task dependency is one of the criteria Bachman (1991) considers important for communicative language testing, and task three, which has this feature, further adds to the communicative nature of PTBST.

Table 1 – Test Procedure Flow-Chart

Section	**Reading Time**	**Listening Time**	**Planning Time**	**Response Time**
Speaking Task 1 – Syllabus Voice-Mail	Not Applicable	3-4 Minutes	Two Minutes	Two Minutes
Speaking Task 2 – Academic Reading and Listening Synthesizing	5 Minutes	4-6 Minutes	Two Minutes	Two Minutes
Speaking Task 3 – Expanding on Academic Reading and Listening	Not Applicable	Not Applicable	Two Minutes	Two Minutes
Total Time	Approximately 30 Minutes			

Further details of the test design will be given in the methodology section. As mentioned before, the description of the tasks here is provided to give a reference point for the discussion in the subsequent sections.

3.2 Why a Computerized Speaking Test?

One important consideration in test development is the task method (Bachman, 1991). When it comes to speaking, it makes sense to many people that interviews, in which an examiner asks questions to test-takers, should be the only method in contrast to speaking tests delivered through computers in which only the test-takers' oral response is recorded and graded. However, one major criticism of interviews stems from the effect the examiner has over the examinee and interview process (Luoma, 2004; Bachman, 1988; van Lier, 1989; Lazaraton, 1992). Also, since most of the discourse during an interview is co-constructed by the interviewer and interviewee, grading the interviewee's speech fails to account for the interviewer's contribution and, hence, is unfair (Brown, 2003; McNamara, 1997). To control for the examiner's effect, interview pairs can be used in which two examinees communicate with each other with the interviewer monitoring, and at times, facilitating the communication. However, in this method, the participants' personality, communicative styles, and linguistic levels are bound to affect their partners' performance (Luoma, 2004, p. 37). Since the research on the nature of this type of influence has been inconclusive and at times contradictory (Luoma, 2004, p. 37), it might make sense to put this method aside altogether and use the computerized method of administration.

Computerized tests of spoken language also have their fair share of criticism. One criticism pertains to the absence of conversational interaction. For instance, Coulthard (1985, p. 60) talks about controlling the next turn by a speaker involved in a conversation by explicitly alluding to him or her and using an *adjacency pair* like asking a question to yield the floor. Yet, this trait of conversation seems impossible to capture in a computerized test. However, despite the possible criticism directed at computerized speaking tests on the basis of the absence of an interlocutor, the prompt provided by the computer in a spoken form can offset this drawback. Hence, similar characteristics of conversational interaction can be achieved through directing a question orally to the test-taker to which she or he is asked to answer orally. Alternatively, a test-taker can be given a chance to ask clarification questions at certain times of the input and create a simulation of interaction aspect of conversation. This is even more characteristic of turn-taking in academic settings when an authority figure, normally a professor or teaching assistant, nominates students to answer questions or share ideas orally during the classroom. Therefore, this method can offer workarounds to compensate for the absence of conversational characteristics in addition to being effective in eliminating the extraneous variable of interviewers or test partners. Therefore, contrary to the common belief that speaking tests should be conducted by interviewers, they do not necessarily yield better results compared to computerized ones, especially when it comes to speaking in academic settings.

However, computerized speaking tests offer advantages which traditional interviews cannot. One advantage concerns the ease of introducing another aspect of interaction in spoken language which is almost absent in interviews, i.e. the interaction between the speaker and the context (Hymes, 1972b). When it comes to academic life, the context can be duplicated in oral interviews; however, this would entail having more than one examiner to conduct both the role-playing and the interview. For instance, one examiner can deliver a short lecture about a certain topic and another examiner would ask questions to the test-takers about the lecture. Yet, this would introduce the element of fluctuation in the performance of the role-playing examiner, which ca in turn affect the reliability of grades. After all, the input would not have the same quality, and this is bound to affect test-takers' performance, the test reliability, and internal validity. However, providing input as videos can not only ensure a consistent input but also a higher quality and more realistic input.

Videos were chosen as the audio-visual input channels for the input of PTBST tasks. There were two main reasons for this choice. First of all, since PTBST is a task-based test focusing on both linguistic and task characteristics of TLU situation, the simulation of the TLU input mode is of high importance (Bachman, 2007; Harding, 2014; McNamara, 1996), and as Ginther (2002) says: "…item stimuli including visual accompaniments to the audio text are considered better representations of actual communicative situations, so the inclusion of visuals may enhance the measurement of the test-taker's listening comprehension." In other words, students in classrooms and out-of-class situations see and use visual input in addition to the aural input, and this should be duplicated in task-based tests. Second, research has shown that having visual input along with the aural one, can enhance the comprehension of students (Wagner, 2013; Ginther, 2002), and create positive attitude in test-takers as Cubilo and Winke (2013) found that test-takers preferred videos over audios as a form of listening input in tests. Videos in PTBST not only features the speaker and makes paralinguistic input provided by the speaker accessible to test-takers, but also, according to Ginther (2002), it provides content and context visuals. Content visuals are images, graphs, and tables which are commonly used in classrooms as complementary sources of information, and the context visuals are those that show where the communication is happening. The former is provided in the form of close captions and images related to the topic in PTBST input videos, while the latter is evident in the video in which a professor is giving short talks in a classroom situation.

Another advantage of computerized speaking tests is the ease of filing and managing the speech samples. Speaking is an ephemeral phenomenon. When it comes to testing, this becomes a crucially deterring factor for test designers, but more importantly, for raters. In order to achieve a reliable scoring, having constant access to the speech data is indispensable. Having speech data accessible all the time enables raters to have a chance to revisit the speech samples whenever they feel it necessary to, say, settle an internal conflict or uncertainty about the decision they have made. This might be one reason why METU NCC proficiency test has conceded to include writing section (a form of performance test and hence inherently possessing the deterring factors involved in assessing performance like coming up with valid and reliable criteria, tasks and constructs), but not speaking.

There is another advantage of speaking tests delivered by computers which pertains to research and development. Luoma (2004) emphasizes the importance of developing the test by drawing upon the data gathered on its previous administrations. Obviously, having access to the response samples and grading records makes this considerably easier in contrast to traditional interview type tests in which the data is as reliable as the memory of interviewers who most often act as raters simultaneously. With the cognitive overload the interviewers in these types of tests are subjected to as a result of attending to the test administration and scoring at the same time, one can hardly rely on the data coming from memories which are subject to deprecation under normal circumstances, let alone when they are dealing with two complex cognitive activities at the same time, namely administering and rating a speaking test. However, this should not be the only reason to go through the trouble of creating a computer software program to administer a test of speaking. After all, interviews can be easily recorded by portable recording devices or even smart phones available in almost everyone’s pocket. Yet, this does not address the practicality issues of test administration, which is the overwhelming task of interviewing hundreds of students in several days. Apart from the impracticality of this approach, fatigue from the exertion of constant concentration by the raters throughout several days is bound to introduce factors which jeopardize the reliability of the test. To address this issue, computerized tests of speaking can be of high value.

Having a software program which both delivers the speaking test and collects and stores speech samples in a database not only makes the test administration a lot faster, but also ensures the consistency in the delivery of the test. In other words, all the participants face the exact same test administration process and interface, listen to the same prompts, and answer within strictly controlled response time which is not susceptible to fluctuations due to human error or sympathy. In other words, it is hard to believe in an objective approach by the interviewers to timing when they might feel sorry for a candidate and allow extra response time, or vice versa. Moreover, due to the cognitive overload in the interviewers, as mentioned before, there is always a possibility of forgetting to stop the candidate at the exact same time, or conversely, of resisting the urge to encourage a candidate to finish earlier to get the much needed break. Besides, fatigue is bound to affect the tone and pace of test administration by the interviewers while computers are not susceptible to such factors.

Finally, the sheer thought of having to interview hundreds of students might make administrators think twice. After all, it might not seem economical to the decision makers to allocate huge amounts of expensive human resources to the administration of a speaking test, especially if it is to happen more than once throughout a year. However, having the possibility of administering a speaking test to hundreds of students at the same time sitting behind computers can solve the issue of practicality, and, as mentioned before, it can enhance the reliability and consistency of the test administration. One concern arising from such an administration might be the interfering of the voice of the test-takers while answering the questions with other students' concentration. This is not an issue since there are special headsets with noise-cancellation feature which are designed for this purpose and do not cost much higher than the normal headsets. This way, with completely sealed ears, test-takers will not get distracted by other test-takers' voice. Moreover, nowadays, the microphones in almost all headsets are unidirectional. In other words, these microphones are designed to capture sound only from a source directly positioned at a close proximity of it.

Therefore, only the sound of the test-taker will be captured by the microphone filtering out the sound coming from other sources, including other test-takers.

3.3 Designing PTBST

Designing a test is a systematic procedure with crucial steps to follow carefully. Each of these steps contributes to one crucial aspect of the test and ensures achieving reliable and valid results. Although these steps are almost similar in all types of the test, the quality and nature of the steps differ according to the test type. Luoma (2004, p. 28) outlines four steps of designing a test of speaking: (a) defining the nature of speaking which is going to be assessed, (b) designing tasks and rubrics to test this type of speaking, (c) informing the test-takers about the test, and (d) ensuring the alignment of testing and scoring with the test construct. In this section, each of these steps and the nature of applications of these steps to the PTBST will be delineated. There will not be a separate section allocated for the last step, namely ensuring the alignment of testing and scoring with the test construct. The reason is that this step will be embedded in the sections allocated to the first three steps. At the first step, during explaining the test construct and its operationalization, the relevant explanations and connections with the different parts of the test and their alignment with each other will be fully explained. Similarly, the relevant considerations of the test tasks and their relevance to the test construct will be discussed in the second section. Also, in the section allocated for the test rubrics, a full explanation of the alignment of each and every part of the test construct to different aspects of the test will be provided. This is to ensure an easier read and a more convenient comparison of the each step to the test construct for the reader.

3.3.1 Speaking nature and construct operationalization. The first step toward designing a test of performance is performing a "job analysis" (McNamara, 1996, p. 16). This entails careful profiling of the setting in which the test-taker will be operating, in this case undergraduate university classes and campus life. This profiling is especially important in a test like this whose goal is to assess the communicative ability of the test-taker in the criterion situation (academic studies, in this case). Luoma (2004) proposes Hymes' (1972b) framework, known with the acronym "SPEAKING", as a reference point to carry out this profiling. In this framework, S stands for the situation. In this test, the classroom situation and course related conversation between friends are chosen. The former is operationalized in tasks two and three in which either the professor is giving a lecture or a student is giving a presentation. The latter is operationalized in task one in which the student leaves a voice-mail summarizing the course requirements and syllabus from a session for a classmate who has missed that session.

The second letter in the acronym stands for participants. In the former situations which are characteristic of the criterion setting (the setting in which the test-taker will be operating), students and the lecturers are the typical participants. The student (test-taker) takes the role of the speaker in all of the questions when responding to the tasks, while the lecturers are the providers of input in the form of videos with reasonable simulation of the classroom settings.

The third letter stands for "ends" which are defined as the communicative outcomes. In the criterion settings, summarizing and synthesizing information, and reporting or persuasion are the representative communicative outcomes. Reporting is the common outcome in engineering disciplines (Wait & Gressel, 2009), yet persuading seems to be more common in non-engineering disciplines. These are operationalized in questions one, two, and three, respectively. In each of the three tasks of PTBST, the test-taker is presented with information in spoken or written form. This information is either related to the facts that students need to know to complete a course or a task (e.g. course syllabus, how to drop/add courses, and how to prepare a specific assignment) or to the course content which students are expected to synthesize into academic work mostly in the form of discussions, reports, or presentations.

"Act sequence" stands for the fourth letter in the acronym which is the sequence of discourse. This is operationalized as the academic style input, followed by formal or informal output based on the input in the test task. The style or tone is what Hymes (1972b) calls *key*. It is tested in question one as the informal, friendly tone and in questions two and three as formal tone which is a characteristic of giving presentations, classroom discussions, and reports.

Instrumentalities, the next letter in the acronym, is the language production channel, which in the case of this test is oral. Of course, writing is part of the test in the form of note-taking as the lectures are set to be relatively long to enforce this practice which is also characteristic of the criterion setting.

The next letter in the acronym stands for norms which pertains to interactional or interpretational norms. Under the former category, asking questions and asking for clarifications are operationalized in the form of opportunities for the test-takers to ask questions (see the methods part). To address and operationalize the interpretational norms, opportunities for expressing views are given in the response to the third question in which the test-taker gives the account of an experience in line with one of the advertisement techniques and has the chance to comment on the experience itself as well as how this experience gets connected to the advertisement techniques. This interpretational norm seem to be more common in non-engineering disciplines in which expanding on the presented facts and opinions by personal interpretations and ideas is common, while engineering disciplines tend to rely more on facts and reporting them directly.

Finally, genres are what the last letter in the acronym stands for. Lectures (the input of questions one and two), instruction (the answer to the first question), storytelling and presentation (the answer to the last question) are the genres which are most common in the criterion setting and are embedded into the construct of all the questions.

Another point of view worthy of taking into consideration in the test construct development is speech events. Although Hyme's SPEAKING framework provides a solid guideline to approach the speaking test construct, it does not consider speech events, which constitute an important criterion to distinguish the nature of speaking in TLU situation. Speech events are important predictors of speech acts, which are proposed by Hymes (as cited in Coulthard, 1985) and have an important part in Hyme's *communicative competence* model. Accordingly, PTBST accounts for different speech events typical of university context. Although the construct coverage by the material is not exhaustive, it is representative and can probably be considered an improvement over METU English proficiency test. There are three questions in PTBST, each of which account for a different speech event and, hence, a different set of speech acts. For instance, in the first question, the *speech event* concerns informing a friend of the obligations and course regulations. The second question is about synthesizing and directly reporting the important topics presented through a complementary set of text and professor speech (also typical of university context and speech events and acts probably more suitable for engineering students). The third question pertains to the speech event of accounting an experience. The latter is an attempt to operationalize the synthesizing aspect of speaking in academic context and using the synthesized content to support a personal point of view which might be more relevant to non-engineering disciplines. To illustrate, in the third task of PTBST, the test-taker uses the information provided by the reading passage and the professor's lecture about the advertisement techniques to support an account of a personal experience of getting affected by one of those techniques. In a similar fashion, students are often asked to use the information they have acquired through reading the course material and professor's explanations to support a certain point of view on the subject matter mostly in non-engineering disciplines. Therefore, it can be argued that the third question has the potential to capture this form of speech event.

3.3.2 A communicative language competence model in PTBST. A model of communicative language ability for testing proposed by Bachman (1990) is chosen as the underlying construct of PTBST in this study. Therefore, in this section, first this model will be explained, and then how this model was operationalized in PTBST will be delineated.

From Bachman's (1990) point of view, developing a test entails a clear definition of the abilities that we intend to measure along with the tools that measure and quantify (or qualify) those abilities. For the former, he provides a slightly modified definition of communicative language ability drawing upon those put forward by Hymes (1972b) and Canale and Swain (1980). The binding backbone of this approach is seeing communicative language ability as consisting of language competence and language use which were explained in the introduction section. However, his framework adds the dimension of interaction between these two constructs in the context of language testing, i.e. the language knowledge or competence and language use. He considers the earlier models of language proficiency measurement unsuccessful as they distinguish language skills from language components and fail to capture the interaction of these two. Even more relevant to the present study is the failure of these early models to capture the context, which is an important factor in task-based tests. But, in his model which follows the works of scholars like Halliday (as cited in Bachman, 1990), he accounts for the importance of context beyond the sentence level. Of importance to this new dimension are discourse and sociolinguistic competences, with the binding factor of strategic competence. While discourse competence concerns the arrangement of utterances to achieve a desired effect in communication, sociolinguistic competence pertains to the rules which govern the choice of certain discourse patterns. Moreover, these dimensions are not independent from the context in which the communication takes place, and using the strategic competence, a language user makes the necessary connection between the context and the language.

The framework Bachman (1990) proposes is composed of three components: language competence, strategic competence, and psychophysiological mechanism. It is worth clarifying that by strategic competence, Bachman refers to something similar to language use. In other words, in his framework, strategic competence is the ability of a person to use their language knowledge or competence, along with the sociocultural and real-world knowledge, to communicate a message with respect to the needs of context including the interlocutor. The final element of his model, namely psychophysiological mechanism, pertains to the "neurological' and "psychological" processes involved in communication.

In his categorization of language competence, Bachman (1990) introduces two main categories: organizational competence and pragmatic competence. The former is further divided into grammatical competence and textual competence while the latter is composed of illocutionary and sociolinguistic competences as two main subcategories.

He defines the grammatical competence as a collection of knowledge of vocabulary, morphology, syntax, and phonology or graphology. The textual competence is further divided into cohesion and rhetorical organization or coherence. Illocutionary competence is composed of ideational, manipulative, heuristic, and imaginative functions. Finally, sociolinguistic competence is composed of sensitivity to dialects or variety, to register, and naturalness along with cultural references and figures of speech. Below is a diagram adapted from Bachman (1990, p. 87) which shows the proposed structure of linguistic or language competence hierarchically.

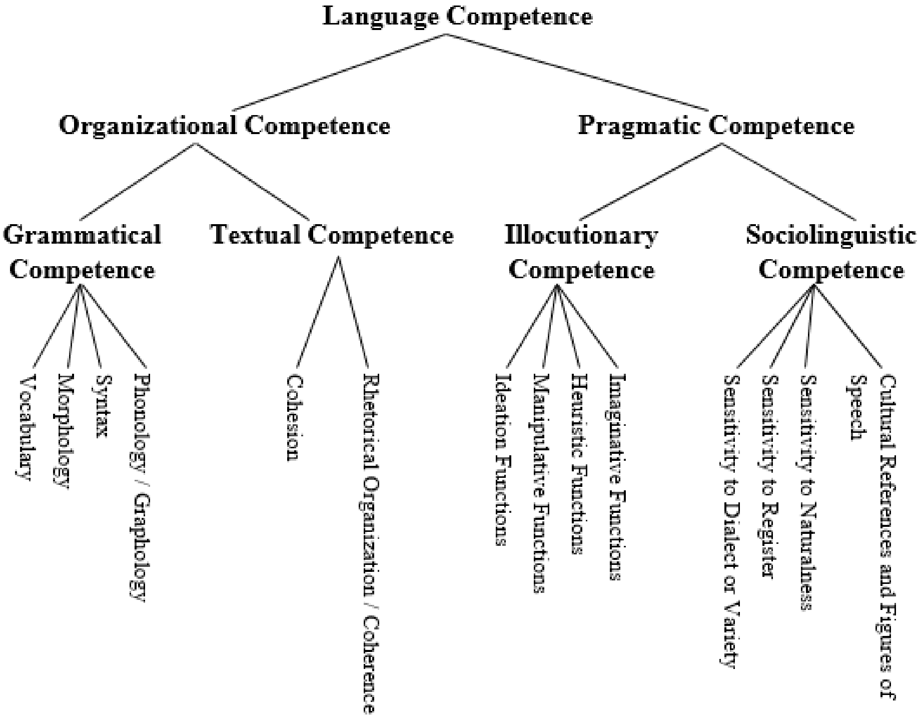


Figure 1 – Components of Language Competence (Bachman, 1990, p. 87).

3.3.2.1 Grammatical competence. Testing for the grammatical competence is the most familiar of all the subsets of language competence. This is because it deals with the building blocks of language, namely vocabulary, syntax, morphology, and phonology or graphology. In the traditional model of testing proposed by scholars like Lado (as cited in Bachman, 1990), these competences were considered separate from the language skills. In other words, each of the grammatical competence components were elicited separately with test items that entailed a minimum amount of integration of language skills, i.e. reading, listening, speaking, and writing. However, in PTBST, these competencies are tested and elicited in combination with each other in the context of more than one skill integrated into one another. For instance, in task 1 of PTBST, the test-taker watches and listens to a lecture by a professor (listening skill), reads the prompt (reading skill), takes notes while listening and watching, (writing skill), and supplies the response in oral form (speaking skill). The same is true for the second and third tasks. Therefore, this test captures the four components of grammatical competence without extracting it and treating it in a different way from language skills.

3.3.2.2 Textual competence. One advantage of PTBST over traditional tests of grammar is its coverage of the textual competence. In order to elicit this competence, test items and elicited responses must move beyond the sentence level and elicit long enough stretches of discourse from test-takers. Although it is true that writing test in METU NCC is capable of eliciting this competence as it elicits a longer stretch of discourse, it does so with some degree of integrating different language skills. This is not sufficient since it is done in only one section of EPE (note-taking) which makes up only 25% of the total writing section. Yet, in PTBST, the test-taker is exposed to the content through more than one skill and is required to produce a longer stretch of discourse in oral form throughout all three tasks of the test. Moreover, since the response is spoken, it entails a different set of discourse conventions which is not the same as those in writing. The discourse competence section will cover this aspect of the test more.

3.3.2.3 Illocutionary competence. According to Bachman (1990, p. 89) pragmatics concerns the nature of relationship between the produced language and the intended function the speaker wishes to achieve by taking context into consideration. Thus, this element captures one aspect of interaction between the context and the language. The first major subcategory of pragmatic competence in Bachman's model is *illocutionary competence*. Illocutionary competence draws upon the illocutionary act in pragmatics which is defined as the purpose of the utterance (Yule, 1996, p. 48). This aspect is divided into ideational, manipulative, heuristic, and imaginative functions. Below, I will give a brief definition for each function as presented by Bachman (1990) and explain how each speaking task in PTBST captures these functions.

Ideational function is defined as the expression of our ideas and feelings about the world around us (Bachman, 1990, pp. 92-93). Task 3 of PTBST captures the ideational function as it elicits the test-takers experiences, ideas, and emotions regarding a past experience by taking into account a contextual factor, which is the advertisement technique presented earlier in the second task. This functions seems to be more common in non-engineering than engineering disciplines as the latter tends to focus more on scientific and mathematical facts than personal ideas. Yet, it is not uncommon to present personal opinions and ideas when it comes to literature, arts, economics, history, psychology, and sociology.

Using language to change the behavior of the people around us is called the manipulative function (Bachman, 1990, p. 93). The first task of PTBST captures the manipulative aspect since its intended purpose is to have a classmate take note of the course syllabus and comply with it. It also captures the instrumental function by using the language to get the classmate to do something, in this case taking note of the syllabus and following it. Two subcategories of manipulative function is also operationalized in PTBST, namely regulatory and interactional functions. Task one of PTBST covers the regulatory function as the test-taker talks about the rules that the classmate should follow as a student of the course. Interactional function is present in this task as it pertains to the maintenance of the relationship between two friends. In other words, if maintaining the relationship did not matter, the test-taker would not call the friend and apprise him/her of the information the negligence of which could have undesirable consequences for the friend. In addition, the response to this task requires phatic language use, like greeting the friend on the phone. This latter aspect will be discussed in the section regarding rubrics.

Another major pragmatic function captured in this test is the heuristic one (Bachman, 1990, p. 93). According to him, heuristic function pertains to the use of language for dissemination of knowledge which is characteristic of teaching and learning situations. This function also entails teaching, learning, problem solving, and conscious memorization. This function is captured in two different forms, i.e. informal and formal, separately in the first and the second tasks. In the first task, the test-taker shares information with a friend, and the nature of the relationship between the test-taker and the friend necessitates an informal form of heuristic function. However, in the second task, in which the test-taker is asked to summarize the main points of the advertisement techniques presented in a reading passage and through a video, the dissemination of the knowledge in a formal manner is elicited. This is achieved through the connotation of the word summary and the test-taker sees themselves as a student in class who is asked to provide a summary of the previously taught material. While the first task necessitates memorization, the latter entails both learning of the concept and then teaching it to imaginary audience. The last task pertains to the problem-solving aspect of this function in which the test-taker has to try to find the reason of one purchase experience in the past using the information about advertisement techniques provided in the second task of PTBST.

The forth subcategory is imaginative function which is defined as the use of language for aesthetic, humorous, or fictional purposes (Bachman, 1990, p. 94). It is worth mentioning that the fourth major subcategory of pragmatics is not operationalized directly in this test. However, it can be argued that in the third task of the test, some test-takers might not have any experience in the past relating to any of the advertisement techniques mentioned in the second task. Consequently, they might invent a story or experience to fulfill the task requirement. This is a common practice in language classes in the form of role-plays or encouragement of the teacher to invent content in communicative tasks when none is available in students' repertoire of experiences. Therefore, it can be argued that this aspect of the pragmatics in Bahman's framework (1990) can be elicited from the test-taker if no personal experience actually exists.

3.3.2.4 Sociolinguistic competence. The second major subcategory of pragmatic competence is *sociolinguistic competence*. Bachman (1990, pp. 95-97) divides this competence into four aspects: sensitivity to dialects and varieties, sensitivity to register, sensitivity to naturalness, and ability to identify and understand cultural references and figures of speech. Below each aspect will be briefly defined and their operationalization in PTBST will be examined.

Sensitivity to dialects and varieties (Bachman, 1990, p. 95) is irrelevant in the context of METU NCC as normally a standard variety of English, either North American or British English is used. Of course, in campus life, being exposed to certain dialects of English is always a possibility; however, the standard varieties are preferred and used in academic contexts. Thus, PTBST does not operationalize this aspect. Yet, it certainly has the potential to do so if the TLU situation makes it necessary. Also, it is worth mentioning that PTBST by no means forces test-takers to use any certain dialect or variety of the language in their responses. It is only the input that is presented in the standard British or North American standard varieties.

Probably, the most important aspect of sociolinguistic competence relevant to the context of METU NCC is the sensitivity to register. According to Halliday, McIntosh, and Strevens (as cited in Bachman 1990, p. 95) there are three aspects which affect the choice of register: discourse domain or filed, discourse mode, and discourse style. According to them, discourse field is the same as the subject matter or the context. For instance, discussing pros and cons of a proposal regarding using solar panels calls for certain choice of register (subject matter) as does the academic discourse (context). PTBST tasks operationalize one context, i.e. academic life inside and outside classrooms, and has the potential of operationalizing a vast variety of subject matters. In this study, course syllabus, economics, and advertising have been chosen as the subject matter. Differences in the mode of discourse also affects the register (Bachman, 1990, p. 95). In the case of PTBST, the spoken mode of discourse has been operationalized. However, since the input is in both spoken and written forms, this test explicitly operationalizes the modality aspect of discourse and register by combining the written and spoken modes of the discourse while eliciting a spoken response in an attempt to capture the ability of the test-taker to distinguish between these two modes and their specific registers. For instance, in the second task of PTBST, the test-taker is expected to read a passage, listen to a lecture, and combine these two forms of input in an appropriate form to produce output. While the content is the same, the test-taker is expected to make the required modifications to the register and discourse to produce an appropriate form of response in terms of register and modality of discourse. Finally, sensitivity to discourse style is operationalized in the form of eliciting formal and informal speech throughout three tasks of PTBST. To illustrate, the test-takers are expected to use an informal style with relevant register in task one while talking to a friend on the phone. However, they are expected to implement a formal style of discourse while summarizing the content in task two and synthesizing the content from task two with a personal experience.

The third aspect of sociolinguistic competence in Bachman's model (1990) is the sensitivity to naturalness. He defines this aspect as the ability to recognize speech as *nativelike*. This aspect is irrelevant to the context of METU NCC as the test-takers and prospective students will be dealing with mostly non-native speakers, and for them, the linguistically well-formedness of utterances is the only important factor in allowing them to communicate successfully in an English-medium academic setting. This view is further accentuated by the relatively new movement in the field of English language teaching which sees native-speakerism as an irrelevant factor (Selvi, 2011).

Understanding the cultural references and figures of speech is the final aspect of the sociolinguistic competence in Bachman's model. According to Bachman, a considerable part of this aspect is lexicalized and is considered part of the lexical knowledge of the language user. For this reason, the input in the first and second tasks of PTBST contain some figures of speech and cultural references which are already lexicalized. However, Bachman refers to some cultural references and figures of speech which are not lexicalized yet as part of the culture. This group of references and figures of speech are irrelevant to the context of English as an International language (McKay, 2003), and it is this variety which is common in the context of METU NCC.

3.3.2.5 Strategic competence. The most important factor in the operationalization and the design of PTBST is strategic competence. This factor is interwoven throughout the fabric and the construct of the test, and it is this factor that adds the "task-based" feature to this test. However, the notion of strategic competence is different from the famous definition provided by Canale and Swain (1980). Bachman (1990) in his model has provided a more comprehensive definition of this construct. While strategic competence pertains to compensating for shortcomings in the language knowledge or performance problems according to Canale and Swain (1980), Bachman (1990) argues that this competence extends to cover not only the mechanisms a language user employs to make up for shortcomings, but also all the mechanisms involved in communicative language use. Bachman's model divides the strategic competence into three phases: assessment, planning, and execution (Bachman, 1990). Below, each phase will be briefly explained and their operationalization in the light of task-based nature of PTBST will be discussed.

3.3.2.5.1 Assessment. This phase consists of four components: identifying the information in the context, determining the knowledge and language competencies available to us, realizing the knowledge and language abilities our interlocutor has, and evaluation of the communicative goal achievement.

Each of these four steps are operationalized in all three tasks of PTBST. To illustrate, in task one, before the lecture begins, the test-taker should identify all the information about lectures, and having identified the context, must go ahead and search his or her schemata for both knowledge of the lectures and course syllabi at universities. Afterwards, the test-takers must identify the important information based on the knowledge they have about lectures and course syllabi. Then, as the interlocutor is a friend, the test-taker should take into consideration the knowledge of the course syllabus and linguistic competence his friend has to limit his communication to only the knowledge that the friend lacks about the syllabus and offer an understandable description of the course syllabus. For instance, a test-taker may not start with the definition of what a course syllabus is, since in his assessment of his and his friend's knowledge, he may realize that it is already a part of the friend's schemata. All these contribute to the final stage of the assessment phase, in which the test-taker decides whether his message is understandable and the communicative goal is achieved. This cannot be operationalized in a conventional sense, since there is no interaction between the test-taker and the friend. However, taking themselves as a departure point and using their evaluation of the friend's knowledge, they try to come up with a message which is most likely to achieve the communicative goal. By the same token, in the second task, the test-takers use the prompt as the criterion to decide what knowledge is needed to communicate (in this case the definition of three different advertisement methods and their relevant example). Then, evaluating their own knowledge and linguistic competence, the test-takers assess the possibility of communicating the message with all the knowledge and competencies available to them. Finally, in the third task, the prompt lets the test-takers know that they are required to give the account of a relevant experience related to one of the advertisement techniques. The test-takers evaluate which parts of their experience are relevant and how they can communicate it successfully taking into consideration the linguistic and schemata knowledge available to them. The assessment stage for the last task might results in different realizations for engineering and non-engineering students. Refereeing to their schemata, engineering students might be more familiar with direct reporting of the input, like they do in their laboratory reports and, hence, have difficulty at this stage of strategic competence, in which they might be unable to find a similar representation of reporting with added personal ideas in their schemata. Yet, this might be easier for non-engineering students who might be more familiar with adding personal ideas to their report of a synthesized input. While, the opposite can be true for the second task which involves only direct reporting.

As mentioned earlier, one shortcoming of this test is the fact that the assessment stage is not readily available to the test-takers. In other words, the interlocutor is not present to provide the necessary feedback on whether he has comprehended the message. Nevertheless, the test-takers draw heavily on the third stage (evaluation of the interlocutor's linguistic and world knowledge) to ensure a successful communication and alleviate the need for the last stage.

3.3.2.5.2 Planning. At this phase, the language user draws upon his language competence in each and every one of its components in the model provided earlier to achieve a plan which successfully communicates the message (Bachman, 1990). Bachman considers this process the most important part of strategic competence. In fact, according to him, this is the phase which is almost synonymous to strategic competence. In operationalized terms, this is the phase which lies at the heart of a task-based communicative speaking test. The reason is that this is the stage where language users convert the language knowledge into language use taking into consideration the contextual factors.

Again, all three tasks of PTBST capture this aspect of strategic competence in a more explicit way. Unlike interviews, PTBST provides the content of the message, and it is clear from the outset which parts of this content must be included in the test-takers' responses to successfully communicate the desired message. For instance, in task one, the professor talks for about 4 minutes. However, the test-takers are not expected to reiterate all the information in the input, nor do they have enough time to produce the same message since memorization factor is ruled out by time limitations imposed by the test administration. Hence, the test-takers must devise a plan which contains the necessary parts of the input from the lecture about the course syllabus and accommodates the linguistic shortcomings. Such a plan, carried out successfully, will result in a message which communicates the necessary information successfully to a friend. The same is true about the second and the third tasks. Selecting and synthesizing the important parts of the reading passage and the lecture about advertisement techniques entail a careful planning of the content with accommodation of the linguistic shortcomings. By the same token, the last question entails enough strategic competence to choose the relevant aspects of advertisement techniques and bind them to the personal experience of a purchase in the past. Once more, the planning stage for the second task might be more familiar for engineering students for the reasons mentioned earlier, while this stage can be more convenient for non-engineering students in the last task.

3.3.2.5.3 Execution. This phase is the actual observable communicative behavior which uses "psychophysiological" recourse in either receptive or productive form through auditory or visual channels (Bachman, 1990, p. 103). The task-based nature of the test has made it possible to implement both visual and auditory channels through inputs in the form of videos and reading passages and output in verbal form. Also, PTBST combines both the receptive and productive modes to achieve a speech sample which entails the combination of both these forms to achieve successful communication of the message.

Bachman (1990), in the answer to the question whether strategic competence can be measured says that this component is a general ability interwoven throughout all the communication, and it is the successful implementation of these strategies which enables the test-takers to perform successfully in performance tests. He also believes that performance tests are more apt to capture strategic competence since the evaluation process values successful communication over the use of certain language structures with accuracy. So, with a careful preparation of rubrics, this competence is more readily available to assessment. This aspect will be discussed in the rubrics section under task accomplishment.

3.3.3 Task design and the content. The second aspect of communicative language testing model proposed by Bachman (1990) is the task-based design of the test. In this approach, the characteristics of the real-life tasks which test-takers will encounter and operate within are analyzed and replicated in test tasks. This approach has two benefits outlined by Harding (2014). The first benefit is that it resolves the issue of interaction authenticity proposed by Morrow (as cited in Harding, 2014) by putting the authenticity emphasis on the task and creating interaction between test-takers and tasks which is similar to the interaction likely to happen in real-life situations. This is very important since it is a tedious and almost impossible task to define the characteristics of authentic interaction and operationalize them. Moreover, empirical validation of interaction authenticity is not possible (Harding, 2014; Brunfaut, 2014). The second benefit is that obtaining empirical validity evidence for task-based testing is easier and possible. Since, in the case of PTBST, the real-life tasks will be those test-takers encounter in their university classes and their success in that context can be evidenced by their semester grades, obtaining empirical validity evidence is possible.

One central factor guiding the design of the tasks in PTBST was eliciting synthesizing information from the test-takers. For this reason, in order to successfully answer the questions, test-takers are presented with information in written or spoken forms to be used in a synthesized form in the spoken response in all three tasks of PTBST. Luoma (2004, p. 29) asserts that the most important consideration in the task development is making sure that the tasks elicit the type of performance which yields appropriate scores in line with the decisions that the test-users will make. This is even more important when the expected types of performance are different for different test-takers. To illustrate, a certain type of task performance might be more appropriate for engineering students, and another for non-engineering ones. Therefore, if the ability of the test-taker in synthesizing academic texts in the target language is part of the construct and is of interest to the test users, the tasks must elicit a performance which can be scored with synthesizing information as an important part of it, albeit with slightly different characteristics of synthesizing to accommodate the difference in the academic disciplines of the test-takers. In order to achieve this, Luoma (2004, p. 30) proposes the criterion context as a point of departure. That is why, as mentioned in the test-construct section, Hymes' (1972a) framework of SPEAKING is used to establish the appropriate context and criteria for the task design of PTBST.

Another aspect incorporated into PTBST tasks is the informational characteristics of the tasks or, simply put, task content. Luoma (2004, p. 32) suggests that a focus on the informational characteristic of the response in a speaking test can better guide the content area of the test and test tasks. Brown and Yule (1983) and Bygate (1987, p. 27) posit that different types of "informational talk" pertain to different sets of abilities, and hence, having ability in one type of informational talk does not necessarily guarantee the ability in the other types. Using this point of view, three tasks in this test pertain to different types of informational talks and, hence, can be asserted to be testing different constructs with the potential to yield information about what informational talk type is appropriate for which discipline of study. Task one of the test captures instructive type of informational talk in which the test-taker instructs a classmate, on behalf of the lecturer, about what she or he is required to do to successfully complete the course. By the same token, the second task elicits a description and explanation of three techniques of advertisement. As to the last task, narration of an experience of buying an expensive product along with a justification of the choice using one of the advertisement techniques mentioned in the second task comprise another type of informational talk.

Types of informational talk are also categorized under the name of *Macrofunctions* by Common European Framework of Reference for Languages (Luoma, 2004, p.33) which is a similar approach in task design under a different name. In this approach, the focus is on language functions and forms the basic principle of task design for the Test of Spoken English (TSE) developed and administered by the Educational Testing Service (ETS) (Luoma, 2004, p.34). Therefore, in the present test, task one is designed around summarizing function, while the second and third tasks revolve around synthesizing with direct reporting and with an added dimension of personal opinion as their distinctive features of synthesizing, respectively. It is also worth mentioning that asking for information or clarification are captured in the questions that the test-takers can ask during the lectures. One pronounced advantage of this approach is the facility with which the test and task designed can be aligned with the criterion situation using the language functions as reference points.

Summarizing, synthesizing, supporting an opinion with synthesized information, and asking for information or clarification can be said to characterize most of the language function in an academic setting, with some being more relevant to engineering and some others to non-engineering disciplines, which are captured by all three tasks in PTBST collectively.

3.3.3.1 Task design and the dichotomy of pedagogic versus real life. Another dichotomy to take into consideration in task design (as well as development of rubrics) is the pedagogic or real-life nature of the test (Luoma, 2004, p. 40). While the former mostly focuses on the language and pertains to language instruction, the latter subscribes to the nature of communication in non-test situation (criterion situation). To further fine-tune the characterization of the task nature, McNamara (1996) distinguishes between strong and weak performance testing. In the strong version, the criterion tasks and situations are simulated with high fidelity, and the focus is on the accomplishment of the task. For example, a student might be asked to convince a graduate student selection committee of a research design with the presence of real committee members. One big problem with this approach is the feasibility concerns. After all, having a selection committee perform the assessment is not feasible when there are hundreds of test-takers involved. The weak performance testing, however, focuses more on the language with a reasonable degree of simulation of the real life. It is the latter which the task design in the present test is based on. This approach not only increases the feasibility of test administration, but also allows for more focus on the language. However, this does not mean that the real-life situation is completely set aside. As mentioned in the earlier section, the content of the tasks simulate that of the academic life and, to some degree, provides simulation of the real-life language use and frees the test from focusing purely on the language.

3.3.3.2 Integrated nature of tasks in PTBST. Speaking does not occur in a vacuum, and when it comes to academic settings, it is normally followed by input provided by lectures, texts, discussions, etc. This nature of academic speaking can only be captured in integrated task types. This type of design might give rise to concerns about the interference of the comprehension ability in making valid judgement on the test-takers' speaking ability. Brown, Iwashita, and McNamara (2005) explain that even if a test-taker’s performance on an integrated speaking task is lower than stand-alone ones, probably because of the cognitive load the input exerts on the test-takers in this type of tasks, integrated task types are well-justified since it is in the nature of integrated speaking tasks to capture the effect of comprehension on production. After all, this is the nature of speaking in the criterion situation. However, these authors suggest that carefully developed rating scales are of high importance in order to achieve high rating reliability which is reported to be low when it comes to integrated tasks. This is corroborated by Bachman (2004, p. 19) who attributes the reliability of a test to both operationalization (task and test development) of the test construct as well as defining test rubrics in line with the test construct. This is actually done and is explained in the rubrics section.

3.3.3.3 Task-based nature of PTBST. PTBST is a computerized test which tests two constructs at the same time: (a) the ability of the test-takers to synthesize the input in written and spoken form and (b) the degree to which the test-takers are able to communicate orally the synthesized material through direct reporting or along with their personal ideas. The former construct is to ensure the inclusion of one of important task-based testing factors listed by Shehadeh (2012), namely authenticity. Language use and communication does not occur in a vacuum, and in the case of academic English, one of the major capabilities of a student/test-taker is comprehending the input, analyzing its content, and using it to produce the response specific to that context and situation. This construct is not specific to academic context. As a case in point, when someone asks for advice about a problem, like inability to pass a course, the interlocutor first must understand the situation and then, using his personal experience and repertoire of knowledge, analyze it to come up with a proper response. The second construct tested in PTBST is to ensure performance-based, direct, and performance-based nature of the test items which are three other characteristics of a task-based test item mentioned by Shahedeh (2012). The fact that the oral production of the test-taker is assessed by PTBST puts the linguistic performance of the test-takers in the spotlight, rather than focusing on their mistakes (Larsen-Freeman, as cited in Graham, 1987, p. 514). Moreover, there is no need to make inferences about the test-taker's ability in language use, as it is the case with discrete-point grammar or vocabulary items which elicit declarative rather than procedural knowledge. Finally, a focus on production is formative in that it contributes to a direct attention to oral communication, and its washback effect is more likely to reorient the focus of courses toward communication and language use rather than language knowledge.

3.3.4 Developing rubrics for PTBST. Luoma (2004, p. 59) considers assessment scales, or rubrics, as a manifestation of test construct subject to the perception of the developer of what the test construct is. Bachman (2004, p. 19) also considers the rubrics, or the operational definition of the test construct, as the important link between the test construct and the scores test-takers receive from that test. However, developing scales is not easy according to Bachman (2004). The difficulty in defining identifiable learning evidence is one factor contributing to this difficulty. This is further confounded by the need to prepare the scales in as short statements as possible. This brevity adds to the possibility of misunderstanding of the scales by raters. To address the first problem, an analytic rubric is created for PTBST with five categories, namely linguistic competence, delivery, discourse competence, task accomplishment, and sociolinguistic competence, and to mitigate the problem of rater confusion, a limited set of criteria are included in the assessment scales with only five levels from 0 to 4.

One important factor of assessment criteria to be taken into consideration is the choice between holistic and analytical approaches. Holistic assessment scales are used in tests like TOEFL iBT speaking and METU EPE writing. The reason why this kind of rubrics is so common is that it does not exert much cognitive load on the raters and is more comprehensive in describing different characteristics of performance at each level (Luoma, 2004, p. 62). Nonetheless, they cannot capture the weak and strong points of individual test-takers, and thus cannot provide useful feedback (Luoma, 2004, p. 62). In addition, holistic assessment criteria depend too much on qualifiers subject only to the judgement of the raters, and thus lowering the reliability of the scores (North, 2012). Moreover, Barkaoui (2010), in a study of difference in the process of rating using analytic and holistic scoring, found that raters tended to be more consistent with themselves when using analytic scoring. Also, using analytic rubrics, raters focused more on the assessment criteria paying closer attention to different criteria and exercising more judgement and self-monitoring strategies. More importantly, Barkaoui (2010) found that analytic rubrics enabled novice raters to focus more on overall language ability rather than individual linguistic features. Taking into consideration PTBST construct and in an attempt to align it with the judgments of raters, a choice of analytic rubrics is justified. In other words, using analytic rubrics, raters would have to make fewer interpretations (as holistic rubrics calls for according to Barkaoui) and would not jeopardize the score reliability. Moreover, more inter-rater consistency will be achieved, all the assessment criteria will be attended to by the raters more carefully (as this is a criterion-referenced test), and linguistic criterion of the rubrics will be assessed in an overall fashion with a focus on the task and content criteria (after all, language use is what PTBST is focused on). That is why an analytical approach to assessment criteria (rubrics) is adopted for PTBST.

Another important factor concerning rubric design is the quantification approach. Bachman (2004, pp. 16-17) classifies the assessment, or "quantifying the observations" into quality judgement and score counting. The former refers to assigning different levels of grades in a continuum along with a range of performance descriptors, widely known as rubrics. The latter is simply assigning values of zero or one based on correctness of responses to discrete point questions. Bachman recommends the first type of assessment for items which elicit extended production of language by test-takers in either oral or written form. Since the responses to the tasks in the PTBST are all of this type, a quality judgment type of assessment, or rubrics, will be used.

Another aspect to take into consideration regarding rubrics is the precision factor, or how precisely the rubrics can provide a measurement of the trait under assessment. Apparently, the higher the number of the tasks and rubric scales, the more precise the measurement; however, this is not always the case, and the precision factor can only be established with appropriate description of tasks and relevant scales and statistical analysis of their relationships (Bachman, 2004, p. 29). So, although the PTBST is composed of three tasks and 5 levels of assessment scale (from 0 to 4 with intervals of 1 point), it does not necessarily mean that this test yields less precise measurements of traits under assessment.

Norm-referenced versus criterion-referenced approach to rubrics is another important consideration in developing assessment criteria. Bachman (2004, p. 30) defines norm-referenced tests as those in which the basis of the decision regarding the performance of the test-takers is the performance of the norm or reference group. In an operationalized terms, a test of speaking to decide whether the test-takers are fit to enter their departments at METU NCC would have a group of successful students at the department as its point of reference. In practical terms, it is not easy to choose such a group to sit an exam and form the basis of a norm-referenced test. This is because defining the characteristics of such group in linguistic terms is not possible since the success of these students does not solely depend on linguistic criteria, and factors like study skills, compensation strategies (like recording lectures), and extra study times can compensate for linguistic deficiencies (McNamara, 1996, p. 42). Hence, a criterion-referenced test is more practical in operationalization terms.

A criterion-referenced test, on the other hand, evaluates the performance of a test-taker in reference to pre-defined criteria (Bachman 2004, p. 31). These criteria can be defined through job analysis — a process through which the determining characteristics of successful performance in non-test situations are investigated and drawn up — to make the test criteria more meaningful and valid (McNamara, 1996, p. 16). One advantage of this approach is the consistency between the test construct and rubrics (assessment criteria). Having determined the important characteristics of successful performance in non-test (criteria) situation, a test can be designed based on these characteristics, and the rubrics can be complied with clearer correspondence with those characteristics and, hence, test construct. As a case in point, the ability of synthesizing the written and spoken input to form an argument is one of the features which marks the success of a university student in linguistic terms. This can be operationalized in the test and the relevant rubrics can be put in place to measure such trait in test-takers with no dependence on the performance of a norm group. In addition, with this approach, salient differences in success criteria based on the academic discipline, e.g. engineering vs. non-engineering, can be captured and operationalized in both the test tasks and the test rubrics.

Taking into consideration the arguments mentioned in this section, the rubrics will be analytical and will be quantified based on quality judgements and will have a criterion-referenced approach. These criteria are explained under each element of the rubrics designed for the PTBST.

3.3.4.1 Linguistic competence. This aspect of the rubrics corresponds to the grammatical competence in the test construct and Bachman's communicative language ability model (Bachman, 1990). The proper use of vocabulary, syntax, and morphology is the focus of attention in this subcategory of rubrics. The second aspect of linguistic competence in the mentioned model, namely textual competence, is treated under discourse competence having coherence and cohesion as its focus of attention. Also, the phonology, which is the fourth aspect of grammatical competence in the model, is treated under delivery.

As Bachman (1990) proposes, here the inclusion of more advanced vocabulary and syntax with high accuracy is not the only criteria to take into consideration, as such an approach ignores the strategic competence of the test construct model altogether. Hence, the effective use of the syntax, vocabulary, morphology, and phonology knowledge by the test-takers to achieve the communicative goal is the major criterion in the linguistic competence component of the rubric.

3.3.4.2 Delivery. Low tolerance for pause and silence is one of the characteristics of speech according to Coulthard (1985, p. 63). This feature of spoken discourse is operationalized as fluency under the delivery criterion in the rubrics. Long, unnatural pauses, which affect the flow of speech, will be penalized because of this important feature of spoken discourse. Moreover, a comprehensible pronunciation along with a natural intonation comprises another part of delivery. This corresponds to the phonology aspect of grammatical competence in the test construct model. It is worth mentioning that intonation is operationalized from a different angle under discourse competence section since it is used to signal organization of the speech when other markers are absent. This will be further explained in the discourse competence section.

3.3.4.3 Discourse. The two aspects of the textual competence component of linguistic competence of the test construct model is treated under this heading in the rubrics. These two aspects are cohesion and coherence (rhetorical organization). Since the responses elicited by all three tasks of PTBST are longer stretches of discourse, they need proper organizational (textual) features to give a proper organization to the responses. Moreover, since the responses are in the spoken form, these cohesive and coherence devices are different from those used in writing. Below are some other important features of spoken language discourse features which are specifically used in the rubrics.

Coulthard (1985, p. 64) mentions pre-structured long turns as a technique to keep the floor for speakers who intend to have it for a longer time. Although the raters are familiar with the main ideas of the talk and are not in any position to interrupt the test-takers (as the speech samples are recorded), they are reminded that they should keep an eye out for these discourse markers of organization of speech. This is important from two points of view. First, in academic discourse, the use of organizational markers (cohesive devices) is important to ensure an effective delivery of the message. Also, it is important for university students to learn to keep the floor for a long enough time to deliver their messages completely. So, these markers become even more important since they serve two purposes at the same time: organization and guarantee of long enough turn to deliver the message completely. These markers can be categorized under meta-interactive acts (Coulthard, 1985, p. 126), which in turn are subcategorized as *markers*, *metastatements*, and *loops*. Of these, the first and second are important in delineating the discourse organization of the response. Marker are phrases that indicate the start of the move (e.g. moreover). Also, metastatements, like “the second problem is that”, signals the focus on a subtopic or topic. These are relevant to responses to all three tasks of PTBST and evaluating the responses to them since these are central to any discourse typical of classroom situations. The presence of these markers to organize the test-takers’ responses serve as the token of better performance and contribute to assigning higher grades under discourse competence criterion in the grading process.

One of the important characteristics of conversation analysis is the concept of *newsworthiness* (Coulthard, 1985, p. 72). According to this characteristic, one must apprise an audience (limited or unlimited) of what the speaker deems to bear importance and consequences for the receiver of the news. In fact this is the same principle which was drawn upon to create the famous slogan for TED Talks series, “ideas worth spreading.” This concept is operationalized in the first question. Such characteristic ensures willingness on the part of the listener to listen to a long stretch of conversation turn without interruption. Also, telling the news falls under the category of story-telling, which also needs securing a long turn by the speaker and avoiding interruption by the listener. *Story-preface* is a technique mentioned by Coulthard (1985, p. 82) to secure such a turn. Therefore, *story-preface* is an important criterion in analyzing the discourse facet of the response in the first question. The speaker must appraise the friend (either directly or indirectly) of the *newsworthiness* of the message she or he is going to leave on the answering machine perhaps with a question like, “Do you want to know what you are supposed to do in psychology 101 course?” as a form of *story-preface*. Of course, this is not all there is to establish the evidence of discourse, but it is an important criterion to keep an eye out for.

The points mentioned before pertain to the initiation and the sustaining of the orally delivered message. Drawing on the principles of the conversation analysis, one other criterion the raters will be looking at is the markers of the termination of the message (Q1), summary (Q2), and description of an experience with references to ideas from a previous task (Q3). Coulthard (1985, pp. 91-92) lists *closing pairs* (goodbye, that’s it, etc.), “…a proverbial or aphoristic summary or comment on the topic which the other party can agree with” (e.g. well it’s going to be a tough semester, isn’t it?), producing fillers with falling intonation (e.g. well), explicit indication of termination by uttering a reason (e.g. well, I gotta go now), making arrangements (e.g. let’s meet for a drink), re-emphasizing previously mentioned points (e.g. so exams seem to be very important for this professor), or the reason for the call (e.g. I just thought I should let you know about what you are supposed to do for this course). The possibility of all of these utterances indicate the appropriacy of the task to elicit complicated, yet necessary, utterances to sustain a solid discourse flow, which is normally difficult, if not impossible, to test and, hence, ignored in exams like METU EPE.

Referring to the studies on discourse and conversation analysis, Coulthard (1985, p. 96) reports that attention is given to intonation only when the other features of speech cannot account for the differences. The same principle can be applied to the intonation criterion of the rating rubrics in a way that when there are not any clear discourse markers to show the organization of the response (e.g. when one main idea finishes and the other starts), intonation patterns will be accounted for to establish organization. He also quotes O’Connor and Arnold (1959, as cited in Coulthard, 1985, p. 98) that intonation patterns are clear indication of the speaker’s attitude pertaining to the situation. This is further supported by Brazil (1985, as cited in Coulthard, 1985, p. 100) who attributes the choice of intonation patterns not to the grammatical units, but to the situation and the decision of the speaker on what part of the utterance should have a specific intonation pattern. By the same token, in the introduction to the concept of discourse, Coulthard (1985, p. 124) explains that speakers use pitch as a tool to arrange the layout of a message and low pitch termination intonation pattern to indicate the completeness of a message. Therefore, a clear rising intonation after a significantly falling one after the end of a long utterance signals the speaker’s attitude toward the new utterance as the marker of a new topic or sub-topic. Hence, paying attention to the intonation is a legitimate way of looking into the speaker’s thinking process without having to stop the speaker or interfere with his or her natural cognitive process as it is the case with think-aloud studies.

3.3.4.4 Task accomplishment. This criterion is operationalized as the test-takers' ability in getting the message across (Luoma, 2004, p. 23). Therefore, successful communication of the main ideas and supporting details in questions 1 and

2 along with the understandable account of an experience with clear connection with one of the advertisement techniques can make up the *Task Accomplishment* criteria. This criterion of the rubrics corresponds directly with the strategic competence defined by Bachman's model (1990). As mentioned in the section related to the operationalization of this model, this competence at the assessment and planning stages takes into account the content and the context along with the linguistic and world-knowledge resources available to the speaker and the interlocutor to successfully communicate the message. Since, the content of the message is the same for all the test-takers, a fair and consistent judgment of the performance of the test-takers from a strategic competence point of view becomes feasible. In other words, while all the conditions (context, content, interlocutor, and other factors affecting the communication) are held constant for all the test-taker, the only source of variance in the performance of the test-takers will be the difference in their strategic competence along with grammatical and pragmatic competences. It is this criterion which captures the ability of the test-takers in combining all the linguistic and non-linguistic resources available to them through their strategic competence to successfully communicate the intended message.

Now, it might be argued that task accomplishment involves some level of memorization, which in turn can affect test validity as speaking ability must be the construct of interest. Crossley, Clevinger, and Kim (2014), investigated the effect of the input text (spoken) properties on the ability of the test-taker to recall the information and better integrate them into their spoken response. They found that the frequency of the words, their rate of occurrence in the input text, and their occurrence in clauses linked with a positive conjunction (e.g. and) can predict whether the test-taker will be able to recall and integrate them into their response. These authors argue that tests of speaking involving input and synthesizing information also measure the ability to recall the information, and that this does not affect the construct of the test negatively since recalling is a major ability in synthesizing information in academic context, and it should be tested as well. Moreover, since these characteristics of the text can predict the performance of the test-takers and the judgements of the raters, it allows for a better and more accurate control over item difficulty by test developers. Therefore, the task accomplishment criterion of the rubrics can be affected by the memorization ability of test-takers, but this does not affect test validity negatively as memorization is a necessary part of synthesizing in academic TLU situation, and the task accomplishment criterion can cover this aspect as well.

3.3.4.5 Sociolinguistic competence. This criterion of the rubrics concerns two aspects of the sociolinguistic competence under the pragmatic competence component of the communicative language ability model (Bachman, 1990): sensitivity to register and style. For the reasons mentioned in the operationalization of the model section, the sensitivity to dialects and variety and sensitivity to naturalness are not included in the test construct and, consequently, in the rubrics. As a reminder, the use of English as an International Language (EIL) is the characteristic of the language in METU NCC context and, hence, dialects, varieties, and native-likeness are irrelevant. However, sensitivity to register and style are both relevant and are included in the rubrics.

For the first task of PTBST, the test-takers are required to use the register related to the course syllabus while using an informal style since the intended interlocutor is a friend. However, in the second and third tasks, while a knowledge of academic register by the test-takers in the context of economics are required, a formal style is elicited in these tasks as they resemble presentations in academic classes.

3.3.4.6 Questions asked. As mentioned earlier, to address the shortcoming of the computerized test regarding interaction, test-takers are provided with three chances to ask questions about the input lectures and, then, are allowed to rewind the lecture to listen to the part that they have a question about. In each lecture (tasks one and two), after each main idea is presented by the speaker in the videos, the test-takers are prompted to ask clarification questions about that main idea if they want to. If yes, they are prompted to ask the question to the microphone. After that, they can rewind the video within the boundaries where that main idea is presented. Since there are three main ideas in both videos, the test-takers can do this three times (once for each main idea) in each of the videos in tasks one and two. In the rubrics, a separate section is included to address and evaluate these questions. It is worth mentioning that asking questions is not mandatory, just like it is in real-world, and the evaluation is relevant only when a question is asked by the test-taker. However, if the test-taker clicks on "yes" indicating that she or he wants to ask a question, but does not do so, a score of zero will be given to that question. The rubrics for questions asked takes into consideration all the aspects of the rubrics together, i.e. linguistic competence, delivery, discourse competence, task accomplishment, and sociolinguistic competence.

Test-takers are required to use appropriate syntax, morphology, and vocabulary to form a question (linguistic competence). They also need to deliver the question with proper phonology and intonation (delivery). They have to use necessary cohesive and coherence devices to create understandable relevance to the part of the lecture the question is about (discourse). Their question must be relevant and accomplish the communicative goal of expressing the need for certain information (task accomplishment). Finally, since the questions are asked to the professor, they must have a formal style and contain proper register related to the topic and the context (sociolinguistic competence).

3.3.4.7 Summary. All the relevant aspects of the underlying construct which draws on Bachman's (1990) model of communicative language ability are included in five criteria in the rubrics. The descriptors of different levels of performance are provided in Appendix A, PTBST rubrics.

3.3.5 Informing the test-takers about the test. The next step in designing a test according to Luoma (2004, p. 28) is informing the test-takers about the test. The informing process is not limited to the test itself and tasks in it. It also covers informing examinees about the rubrics so that they can make conscious decisions about what performance they should aim for to attain their desired grades (Luoma, 2004, pp. 61-62). Informing the test-takers about the tasks themselves is accomplished through the complete instructions provided at the beginning of the test and before each tasks (see Appendix F). However, once this test is implemented for non-research purposes to make actual admission decisions based on its scores, test-takers must be fully informed of the test prior to taking it. In other words, test tasks must be introduced throughout the English Preparatory Courses offered at METU NCC, and familiarize the test-takers with the tasks through sample questions and instruction of strategies to accomplish the tasks more successfully. This, naturally, will contribute to the washback effect of the test itself and put a higher priority on speaking skills.

As to the rubrics, the participants in this study were provided with explanations about the rubrics and scoring criteria before taking the test. The major reason was to limit the variance in the performance to test-takers' communicative language ability factors and isolate the test from factors like errors of measurement (i.e. variance in performance stemming from factors other those PTBST is intended to measure such as difference in the level of familiarity with test tasks and measurement criteria). This is one of the important considerations that Bachman (2004, p. 93) points out to be taken into consideration to ensure test reliability. Therefore, poor performance of the test-taker will not be attributed to their unawareness of the test rubrics, nor can it be attributed to the lack of familiarity with the test tasks as they were explained both orally before the test to each test-taker, and they were offered a full instruction of each task by PTBST software program.

CHAPTER 4

METHOD

4.1 Participants

Thirty-four first-semester students of METU NCC who had just received a score above the cut-off point in METU EPE participated in this study. The cut-off score of EPE is 70 and 60 for students of English Language Teaching department and other departments, respectively. Those who had obtained half a point below this cut-off score were also allowed to enter their respective departments.

Table 2 – Number of Participants and Their Majors

Programs	Major	Number of Participants	Number of Drop-outs
Economics and Administrative Sciences	Business Administration	1	
	Economics	1	
	Political Science and International Relations	4	
Engineering Programs	Civil Engineering	3	1
	Electrical and Electronics Engineering	6	
	Computer Engineering	1	
	Mechanical Engineering	2	
	Petroleum and Natural Gas Engineering	2	
Education/Humanities Programs	Teaching English as a Foreign Language	3	
	Guidance and Psychological Counseling	3	1
	Psychology	8	1
Total		34	3

Of these, six were from Economics and Administrative Sciences program, and 14 students from each of the Engineering programs and Education/Humanities programs participated in this study. Table 2 shows the number of participants from each major under these three programs.

The recorded responses of the psychology department student who dropped out of the study did not contain any speech. Consequently, it was taken as a sign of this student's dropping out of the study. The other two drop-outs did not send their first semester academic record. One of them did not respond to the contact attempts made by the researcher; however, the other participant informed the researcher of their dropping out of the first semester due to excessive workload and low possibility of success. Therefore, the study was completed with 31 participants.

Prior to taking their METU EPE exam, these participants were students at the English preparatory program of School of Foreign Languages (SFL) and were required to obtain a certain cumulative grade throughout their one- to two-year training at the mentioned school before being eligible to take the proficiency exam. During their preparatory course, they are grouped into mainly three levels of Pre-Intermediate, Intermediate, and Upper-Intermediate groups. The first level roughly corresponds to A2$^+$ to B1 levels of Common European Framework of Reference (CEFR), the second to B1$^+$ to B2, and the third to B2$^+$ to C1. The CEFR levels are approximate equivalents of the three levels mentioned at METU NCC SFL, and are inferred based on the level of Language Leader course books (Lebeau & Rees, 2008) which are assigned to these levels by the school. Of the remaining 31 participants, 10 were from pre-intermediate level, 16 from intermediate, and 5 from upper-intermediate levels. It is worth mentioning that all three participants who dropped out of the study were from pre-intermediate level.

The average age of the participants was 20. Of these 31 participants who remained in the study, six were not from Turkey or Northern Cyprus. So, the admission process for these students is mainly based on high school average grade (METU Cyprus, 2016). However, the remaining 25 Turkish and Northern Cypriot participants went through a different procedure in order to be admitted to METU NCC. Their admission to the university was based on their performance on a university entrance exam (METU Cyprus, 2016). The admission policies of the university mandates choosing the first group based on a relatively competitive high school (METU Cyprus, 2016), and the second group from top quartile of the university entrance exam results, which is evidenced by the minimum scores required (mostly higher than 75%) to be admitted to different departments of METU NCC reported on their official webpage (ODTÜ Kuzey Kıbrıs Kampusu, 2016). This is important to keep in mind since the extraneous variable of students' scholastic ability might be argued to affect the first-semester university Grade Point Average (Graham, 1987), and referring to the correlation of proficiency test scores and GPA as a measure of predictive power of PTBST or EPE might be called into question. Yet, these students are high-achievers which is evident by their choice of university which is among the top 100 universities in the world (METU Ranking, 2016). This, along with the fact that they possess the relatively high admission requirements of METU NCC, can mitigate the effect of the mentioned extraneous variable considerably.

The participation was on a voluntary basis. Due to the excessive workload of the students and the university policies, it was not possible to have access to a high number of participants and perform a random sampling. Consequently, a convenient sampling was adopted. Yet, Table 2 shows that the sample covers both engineering and non-engineering majors, though not equally. The sample also covers participants from the pre-intermediate, intermediate, and upper-intermediate levels, albeit not in equal numbers again. The unequal sampling of participants from different language ability levels can have a minimum negative effect on the research design as this study is interested in the correspondence of language ability to academic success, and those with low language ability are hypothesized to have lower GPAs compared to those who have a higher language ability. However, having equal numbers of engineering and non-engineering participants is more important to the research design as the pattern of difference in predictive power PTBST and EPE may show for these two disciplines are of interest in this study.

4.2 Data and Permissions

Data for this study were gathered from three different sources: METU EPE scores, first semester GPA, and PTBST scores. The first two were obtained directly from the participants and the latter was obtained after administering and grading the PTBST.

Having applied to the METU Ankara Human Subjects Ethics Committee, as it is the required procedure, the necessary permissions to conduct the study were obtained. Also, the participant consent form was approved by the same committee, and according to this consent form, students agreed to participate in the study, provide their METU EPE grades and first semester GPAs, and take the PTBST. The signed forms were filed, and are available if needed.

Each participant gave their METU EPE scores and sub-scores (scores for each section of EPE) after signing the consent form and before taking the PTBST, by logging into their EPE score report page at the presence of the researcher. The consent form ensured the participants of anonymity and confidentiality of all the data obtained directly from them or from the PTBST. The scores were stored in a password protected MS SQL Server database. Then, having taken the PTBST, the participants agreed to send a screenshot of their first semester grade report and GPA through email to the researcher. After the semester grades were reported, the participants were sent reminder emails to send their grades. At this point, three students did not send their grades and were considered as dropouts. The reports were entered into the password protected MS SQL Server database and were stored on a computer.

Finally, the PTBST scores were obtained and finalized and were entered to the password protected MS SQL Server database. The procedures of grading will be explained in the procedures sections.

4.3 Tools and Material

The major data collection tool in this study is PTBST software program. This is an online speaking test delivery software program which is custom-designed by the researcher for this study. The major body of the website is created using MS Visual Studio 2013 with ASP.Net technology. Simply put, this a website designing technology which allows for interactive website content design and delivery with a database for storing user data and information. The database which stores the question contents, texts, and user information was designed using MS SQL Server 2012. Also, to manage the video content, voice recording, and user interface (the webpages with which the test-taker interacted), JavaScript technology was implemented. However, the video was delivered through MS Windows Media Player which is installed as default on any windows operating system. Finally, a desktop web browser application was designed using VB6 to increase the security of the test. To illustrate, this custom-designed web browser, once run, locks the screen and shows only the test content and allows interacting only with it. This is to further increase test security and block the test-takers' access to any other resource on the computer such as dictionaries, encyclopedias, or websites. All these technologies and tools combined served as the designing tools of the online software program which delivered and stored the speaking samples of the PTBST in a secure environment.

The program was designed to deliver each step and task of the test in the correct order and allocated timing. Having entered a combination of username and password chosen by the test-takers at the time of online registration, the test-takers are taken into the test environment where the content, instructions, tasks, and prompts are delivered to them. In the case of any interruption in test delivery due to power cuts, computer failure, or any other similar reasons, the software program is designed to resume the test from where it was left after logging into the system again once the problem is resolved.

The recorded files of the test-takers are uploaded to the server which hosts the PTBST software and stores it with a timestamp and the credentials of the test-taker. These files are easily retrievable since they are stored in an orderly fashion on a secure host.

Prior to the administration of the test, the software program showed a digital version of the consent form, and the test-takers were allowed to proceed only after agreeing to this digital version of the consent form. The date, time, and details of the agreement were all stored in the password protected MS SQL Server database. As another precautionary step, immediately after agreeing to the digital consent form, the voice recording is tested by a practice question: "Describe your university." The test-takers were asked to speak for 15 seconds while the software program analyzed the voice volume and quality. If desirable sound quality was obtained, the test-takers were allowed to continue with the test. Otherwise, the test stopped and did not continue until a good recording quality was obtained.

As mentioned before, the participants could replay parts of the lecture with limited control. In other words, the test-takers could tweak the progress bar of the video three times during the lecture. This was to add an element of interactivity to the test administration, and maximize the reliance of performance on mostly speaking ability. In other words, as asking questions to a lecturer is a common characteristic of academic classes, this feature was operationalized in PTBST in the form of allowing the test-takers to ask a question and record it and to tweak the progress bar only once for each question in return as a simulation of getting an answer to the question. This was done three times as there were three main ideas in each lecture, and participants could ask questions at the end of each main idea and before starting to listen to the next main idea. After each main idea is presented by the speaker in the video, a message box appears asking the test-takers if they have any questions with two buttons: "yes" and "no". If the test-takers click on yes, they are prompted to ask the question by speaking into the microphone for 15 seconds. Then, they can rewind the video within the boundaries of the section containing that main idea. If the test-takers click on yes, and ask no question, as explained in the rubrics part, they are given a score of zero for the respective part in the rubrics. However, if the test-takers click on "no," the video resumes and the lecture continues. This is to ensure that test administration does not take forever. This feature of the test caters for the shortcoming regarding the interactivity of speech in computerized speaking tests.

The first task contains a video of a professor talking about the course syllabus and requirements of psychology 101 course at the first session of the course. The script of the video was adopted from a real online course syllabus PDF document (Kermond, 2012). The second task contains a four-paragraph text about three advertising techniques along with a video featuring a professor giving examples to illustrate those three techniques. The content for the second task was adapted from an online article (Wiebe, 2013) about advertising techniques which makes expensive products seem cheap. The adaptation of the contents in both tasks was to make the input suitable for the test construct and design. The third task has no content, and the test-takers were asked to use the contents of the second task to make connection between a purchase experience and one of the techniques mentioned there.

The researcher acts as a professor in both videos. The reason for this decision was that there was no one the researcher could ask to go through the time-consuming and painful process of shooting a video on campus, as all the possible candidates had a very busy schedule. Likewise, online and pre-made videos could not have been used in the research because of copyright issues, quality considerations, and compatibility of the content of the video to the construct of the test. One concern raising from this matter is that the contact the researcher had with the participants could have affected their performance. However, since there was a minimal contact between the researcher and the participants, and the level of contact was the same for all the participants, the results could not have been biased. Further analysis under *Reliability Analysis of PTBST* attested to the absence of such a bias. The videos were recorded by a Sony HD camera, and was edited by Sony Vegas video editing program. During the post-production editing of the video, captions were added to both videos to provide more paralinguistic clues to the test-takers and aid them in better understanding of the lectures. In addition to captions, the second video contained some images which were also added during the post-production editing.

4.4 Procedures

4.4.1 Designing and piloting the test.

Having carefully outlined the test blueprint following the construct mentioned in the test construct section, the researcher started designing the test software program and creating the content of the test using the sources and tools mentioned in the previous section. Then, a domain and host service was purchased from isimtescil.net — a website hosting company in Turkey — and the website was published to the hosting service. Then, an email was sent to all teaching staff at METU NCC SFL, asking them to volunteer to sit a piloting session. Three instructors agreed to help, one of which was a native speaker of English. These three volunteers sat a session at different times and gave their feedback on different aspect of the test, including the time allocated for the reading passage in task two. Initially, a two-to-three-minute reading time had been allocated for the reading in the second task. However, this was increased to five minutes in order to prevent introducing the effect of reading competence factor into a speaking test. All three volunteers found a five-minute reading time adequate. Moreover, they agreed that the instructions given before each task were comprehensive, necessary, and clear.

Another important aspect of the timing which was scheduled to be decided upon during the piloting was the preparation and response time for each of the tasks. According to Shehadeh (2012), planning and response time are of the key implementation procedures which can affect the performance of the test-takers. Therefore, the participants in the piloting were specifically asked to give their feedback on the allocated time for preparation and response in each task. They also agreed that the allocated time for preparation was sufficient. Also, their responses to the questions were analyzed to see if the expected answer could be delivered during the allocated two minutes while talking with a normal pace. All the piloting participants successfully gave a complete spoken response to the tasks using a normal pace in roughly two minutes.

Another reason for piloting the test was to see if any problem arose from the software program during the test in order to fix them before the actual administration of the test. Two administrations of the test went smoothly while the third one was interrupted. The researcher analyzed the problem and realized that the interruption was due to an internet connection loss. When the connection was reestablished, the test software resumed the session from where it had been left off. This provided a chance to pilot for this aspect of the test. In the future, in order to have an even smoother administration of the test, a change will be made to the software program in order to allow the test administration to continue even if an interruption in the internet connection occurs. For this purpose, the software will have to be modified to store the information locally on the computer, and upload any necessary files to the server as soon as the connection is reestablished.

4.4.2 Sampling. In order to find participants for this study, a message was posted on a Facebook group of which almost all the students at METU NCC are a member. Those who responded, were asked to send information about their last EPE date, their English language level when they were at METU NCC SFL, and their majors. Having acquired this information and ensuring that there were volunteers from all the levels of METU NCC SFL and both human sciences and engineering departments, the researcher decided that the population is representative enough of the target population. It is a fact that the number of participants is fairly low; however, since the circumstances did not allow for any further sampling and due to time constraints, the researcher had no choice but to terminate the sampling and start with data collection.

4.4.3 Test administration. Since the software program is online and needs a minimum amount of preparation of the hosting computer, the participants' personal laptops were also used whenever there was more than one participant at the same time because the researcher had only one laptop and could host only one participant at a time without having extra computers. In the case of more than one participant taking the test, appropriate headsets which covered the ears completely were used to prevent the distraction caused by other participants' voice. Also, the headsets had unidirectional microphones which captured only the voice of the test taker and filtered out the sound coming from other sources.

The location of the test administration was kept constant, and since the test administration was identical, thanks to the software program, other than the variability of the date, all the participants took the test under same conditions. So, as a validity argument, variation in the performance of the participants could not have resulted from not having a simultaneous administration of the test. Besides, since they chose to come at a time convenient for them, it can be argued that they were at the top of their performance capacity, which is another advantage regarding the test validity since factors like fatigue can affect performance in the test, and the variance in the performance could be attributed to factors other than the ability/knowledge which the test is designed to measure (Bachman, 2004, p. 287). However, care was taken to keep the identity of participants confidential from each other so that they would not share the question contents with other participants. The statistical proof of this can be found under *Reliability Analysis of PTBST* where a Many-Facet Rasch Analysis shows only two cases of bias out of 99 interactions between participants and tasks. This is favorable, and minimal cases of bias between a participant and task attests to the lack of possible communication of test content among the participants.

Having signed the consent form, each participant was asked to formally provide their EPE scores and dates (as mentioned before). When the conditions satisfying the participation criteria (passing the EPE recently and being the first semester student at their respective faculty) was established, the participants agreed to send their first-semester grade report by signing the consent form. Then, the researcher explained to the participants what they are expected to do and explaining the procedure (see *informing the test-taker about the test* section for the reason for this step of the procedure). After that, the test-takers asked to sign into the test software and complete the test. Having finished the test, they were told that their grades would be sent to them as soon as they were ready.

Perhaps discussing the participant attitude toward PTBST is of value at this point, as it is a computerized test and is not a type of test the participants in this study were familiar with. Initially, the participants expressed their reservations regarding taking a computerized test. The major reason was that they did not know how they were expected to take the test, especially because it was a speaking test and did not fit in their perception of a speaking test, in which they expected a face-to-face communication with an examiner. However, they were all told that all the instructions were explained elaborately in the testing software itself, and that they would not have any problem with taking the test having paid careful attention to those instructions. Also, they were told that everything was automated and they needed a minimal amount of interaction with the interface to manage the administration of the test. Despite all these explanations, they seemed not to have been convinced. However, once the administration of the test began, they showed little anxiety or confusion regarding the test interface and its procedures. In fact, almost all of the participants exhibited facility in interacting with the test, and the confidence was evident in the way they took notes, responded to questions, and focused on the input. Moreover, the only complaint they had after finishing the test was that it was a little difficult. This might arise from the novelty of this kind of speaking test for them. Alternatively, it can stem from simply the task types which entailed synthesizing the input which were relatively long. In short, apart from only one participant who did not answer the questions at all, all the other participants responded to all the tasks, and some of them attempted to use the *Ask Questions* feature, which can be taken as the evidence of their comfort with the test method. Furthermore, no participant said that a speaking test must not be delivered through a computer, and even some had a more positive attitude toward a computerized administration of the speaking test.

4.4.4 Rater training and rating. After preparing the rubrics based on the test construct (see the developing rubrics for PTBST), an email was sent to all the teaching staff at METU NCC SFL. Of these, five instructors volunteered, two of which were native speakers of English, to participate in rater training session and ratings. A date was agreed upon and a three-hour rater training session, with two five-minute breaks at the end of each hour, was conducted. Initially, the test itself was introduced to the raters. They were shown the instructions, prompts, and the reading/listening contents of the three tasks. Then, the rubrics were given to the five volunteers and were asked to read through them. The first hour terminated with a question-and-answer segment to clarify the rubrics. The raters were told that and shown that the rubrics contained five different criteria for each task with five levels of performance from 0 to 4. They were also told that in order to maintain consistency, only one of five whole numbers should be assigned as a score with no decimal points.

After the short break, the speech samples of three test-takers were given to the raters and were asked to rate each task alone. After each rating, the scores given by each rater was elicited. If there was no more than a one-point difference between the scores, the rating of the same task of another participant was started. However, in case of a discrepancy of more than one point, the rater with the aberrant score was asked to explain the reason for their decision. Through these explanations the reason for the discrepancy, which was always a misunderstanding of the descriptors in the rubrics, were elicited and further explanation was offered to reach a consensus. With no discrepancy in two subsequent ratings of the same task, the rating of the next task started. This procedure was followed throughout the last two hours of the rater training session. There were only two cases of discrepancies of more than one point, one of which concerned discourse competence criterion and the other sociolinguistic competence. However, through the same method of eliciting reasons and providing further explanation of the descriptors of the criteria in each case, the discrepancy was resolved and did not arise in the subsequent two ratings of the same task.

It is worth mentioning that two out of three speech samples used in training were used in the subsequent analyses in this thesis. The other one was one of the drop-outs. It might be argued that involving these two participants in the training could have biased the results. However, the bias statistics of Many-Face Rasch Analysis under *Reliability Analysis of PTBST* shows that out of 66 cases of rating (two ratings for each task) only three were biased, which proves that such a bias could not have existed since nine tasks (three for each participant) were graded in the training session. To further confirm the lack of such a bias, the two participants which were included in the analyses were taken out, and all the analyses were run again. The results showed that taking out the grades of these participants increased the correlation statistics of PTBST with total and weighted GPA. Therefore, including these participants in rater training clearly did not introduce any bias into the research results.

The procedure mentioned above was followed to rate the responses to all three tasks plus the recorded questions of the participants about the lectures in task two and three (see *questions asked* under *developing rubrics for PTBST*). At the end of the rater training session, the raters were given the recorded responses of 12 participants with arbitrary numbers assigned to each file to protect the anonymity of the participants and prevent possible prejudice by the raters in case they recognized a participant (all the participants were METU NCC SFL students before, and there was a chance that they were the rater's student). These arbitrary numbers were assigned automatically by PTBST software and were back-traceable to the participants. Therefore, the scores given by the raters could be easily entered in the right place in the database. Each rater rated half of the same participants as two other raters. In other words, each rater were assigned to rate 12 participants' responses, and each participant was assigned to two different raters. Figure 2 shows the distribution of the participants among the raters more clearly.

After the ratings were finished, the grades were entered into the database. Each criterion of each task performed by each participant received two scores from two raters. The two ratings were entered next to each other and the discrepancies were explored. The discrepancies of more than one point were flagged for adjudication. For adjudication, another volunteer was chosen. The training of this rater was more rigorous. The first part of training was similar to that of the other raters. However, in the second section, all the tasks which had received the exact same score from two raters were chosen, and the adjudicating rater was asked to rate all of them. A Many-Facet Rasch Analysis had been run before, and based on the statistics, the mentioned raters had the closest logit values to zero. In other words, these raters were the most objective ones. Thus, using their agreed upon grades as a reference point was an attempt to cater for one of the most important aspects of rater training in *Frame of Reference* (FOR) framework (Roch, Woehr, Mishra, & Kieszczynska, 2012). According to this framework, in rater training, a reference point which represents the characteristic features of performance at different levels is used to train a rater. These features were obtained from the ratings of the mentioned raters.

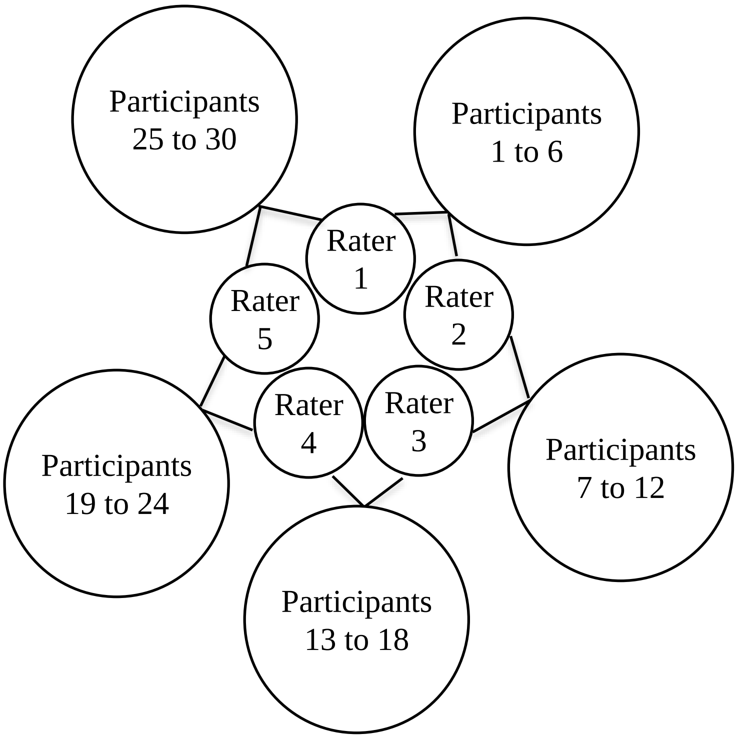


Figure 2 – Participant-Rater Assignment Plan
This chart illustrates how the participants were assigned to each rater.

In the case of any discrepancy of even one point, the rater was asked to explain his/her reasons, and if a misunderstanding of the rubric descriptors was the source of discrepancy, further explanation was offered to ensure a clear understanding of the descriptors in the rubrics. In case the reason was the rater’s strictness, which means that the rater tended to deduct more points than the rubrics outlined, he was asked to adjust his strictness according to the aforementioned reference judgements, and he was asked to use the difference between his and the other raters' judgements as a further guiding point. Again, this is an important part of rater training within FOR framework as Roch, Woehr, Mishra, and Kieszczynska (2012) put it:

> …the importance of raters sharing a common conceptualization of performance categories, that is, a shared FOR, is not only relevant to performance appraisal but for all human resource functions that rely on raters. Researchers have shown that FOR training is directly applicable to a variety of evaluative contexts including assessment [centers].

After a consistent rating of at least three times of each criteria in the rubrics and the tasks, the adjudicating rater was asked to assign a score to the criterion for which the initial two raters had given scores with more than a one-point difference. Then, the score given by the adjudicating rater was compared to those given by the two previous raters and the closest score was chosen, and then averaged out to get the final grade for the criterion. In case the grade given by the adjudicating rater was between the two previously given grades, the former was chosen as the adjudicated score.

After the adjudication was over, the scores which had no discrepancies were entered to the final database as is. Those with only one-point discrepancies were averaged out, and the average score was entered under the relevant criterion. Finally, the finalized scores after the adjudication process were entered for those with more than one-point discrepancy.

4.4.5 GPAs and semester grade reports. The grades of PTBST were not reported to the participants immediately. This decision was made mainly for expediency. In other words, immediate reporting of the grade would have distracted the participants as they were taking their final exams, and it was a better idea to wait until their final exams were over. Moreover, having given them the results might have discouraged them from sharing their GPAs and semester grade reports since they might have seen no reason for sharing them after getting their grades. Therefore, as soon as the semester grades were reported by the participants' respective faculties, they were sent an email reminding that it was time they sent their GPAs and semester grade reports. They were also told that their PTBST grades were ready and they could receive them after they sent a screenshot of their semester grade reports along with their GPAs to the researcher through email or any other available messaging means. Except for two participants, all of them sent the screenshots and received their PTBST scores in return. PTBST grades were converted to a 100-point-base score by simply multiplying their raw score into 25. This decision was made to report the scores in a more familiar fashion as they are used to receiving scores in a 100-point scale when it comes to English language tests — a habit probably formed through their English preparatory program at METU NCC SFL. In addition to those two participants, as mentioned in the participants' section, there was one who did not speak at all, and, hence, his performance in PTBST was not possible to grade. Despite this, an email was sent to the participant asking for their GPA and semester grades. As expected, this participant did not respond. Therefore, at this last stage of data collection, 31 out of 34 participants provided enough data for analysis, and the other three were considered as drop-outs.

Having received the semester grade reports and GPAs, the data were entered into the database. At this point, each of the 31 participants had three sets of scores: PTBST total and task scores with scores for each criterion in the rubrics, METU EPE total scores along with sub-scores for each section of the test, and first semester GPA accompanied by semester grade report containing information for each course such as the grade and number of credits. With all the needed data at hand, the process of statistical analyses was started.

CHAPTER 5

ANALYSIS

5.1 Reliability Analysis of PTBST

One of the most important aspects of any assessment ıs its validity which is the accuracy of the scores reflecting the tested abilities of a test-taker, and one decisive contributor to this accuracy is the rater when it comes to tests of performance like speaking and writing (Morgan, Zhu, Johnson, & Hodge, 2014). Hence, the reliability of the judgements made by raters contributes substantially to test validity (Bachman, 2004). In other words, in order for a test of performance to be reliable and valid, it must measure the ability of interest and exclude the measurement errors, which in this case are those that arise from factors like subjectivity in performance judgement (Bachman, 2004; Morgan, Zhu, Johnson, & Hodge, 2014). One approach to test for this measurement error is looking into the consistency of raters through reliability estimate like Pearson product-moment and Spearman rank-order correlation coefficient, which Bachman (2004) calls classical test theory estimates of inter-rater reliability. One problem with using these estimates is that they do not take into account the interactions between the task difficulty, rater severity, and test-takers (Bachman, Lynch, & Mason, 1995; Bachman, 2004). Moreover, Morgan, Zhu, Johnson, and Hodge (2014) report that Pearson product-moment and Spearman rank-order reliability coefficients which are commonly used to estimate inter-rater reliability can show as much as a-third of real estimates of inter-rater reliability. Therefore, using another reliability estimates can be more accurate.

One of the methods to empirically study the reliability of performance and task-based tests is the use of Many-Facet Rasch Measurement (MFRM). The measurement reveals the differences in task difficulty and rater severity, as well as the interactions between these factors. Using this model, the inconsistencies of raters in their judgment can also be revealed.

MFRM is an extension of Item-Response Theory (IRT). According to this theory, the performance of a test-taker is the interactive function of test item difficulty, test-taker ability, chance, and the discriminatory power of items in separating weak and strong test-takers (Embretson & Reise, 2000). However, the basic form of this model is used only for dichotomous items with only right or wrong answers. In the case of performance and task-based tests, in which partial grading is possible, this model is not applicable. Therefore, an extension of this model called MFRM is developed. This model can work with tests with partial grading (Bachman, Lynch, & Mason, 1995).

It is worth mentioning that MFRM operates based on a probabilistic approach. (Bachman, Lynch, & Mason, 1995; Embretson & Reise, 2000). In other words, using linear scales called logits, the results of this analysis show test-taker ability, rater severity, task difficulty, or any other facet of measurement on logit scales. Moreover, this model analyzes the relationships of each two facets at each step, and combining all these estimates, produces a final probability model which shows the positioning of each of these aspects along with the scales.

For the analysis of reliability of PTBST, the grades given to each task by each rater was entered into a three-facet model using FACETS program, version 3.71.4. The test-taker variable was appointed to the first facet, while rater and task score variables were assigned to the second and third facets. The analysis was run and the relevant results are reported here. Table 3 shows rater severity ordered from the harshest to the most lenient. The first column shows values for measure logits. Any value beyond ±2 is considered too much deviation from the model (Linacre & Wright, 1999) and indicates the need for a retraining or replacement of the rater (Bachman, 2004). The results show that all the raters' measure logits fall within ±2. Moreover, rater one with a logit value of 0.58 is the harshest of the raters, while rater four with a logit value of -1.45 is the most lenient of all. However, except for that of rater four, the differences between the values are small and shows a good consistency among the raters. Since rater four had a logit value between ±2, her ratings were kept in the study, taking into consideration that pairing her with two other raters with harshness value close to zero can balance out the leniency.

Table 3 – Rater Severity Results of MFRM

Measure Logit	Model Error	Infit MnSq	ZStd	Rater
0.58	0.20	1.12	0.7	1
0.47	0.20	.60	-3.0	3
0.38	0.23	1.28	1.4	2
0.02	0.22	1.11	0.6	5
-1.45	0.23	.90	-0.5	4

Mean Square values between 0.5 and 1.5 are productive for measurement. Standardized values between -1.9 and 1.9 have reasonable productivity. Values below -2 are too predictive and are constrained by other dimensions (Linacre & Wright, 1999).

Zhang and Elder (2010) tried to find out if non-native English speakers (NNES) differ from native English speakers (NES) when they judge the speaking performance of test candidates. They report previous studies to have come up with mixed findings. Some attribute harshness to either non-native or native speaker raters. Some others characterize either of the groups with consistency in their judgments. Still, the remaining group finds no differences. This inconclusive observations were echoed in this study as well. As mentioned before, two of the five raters were NES. While one of NES raters is the most lenient one, the other is the third harshest rater according to the results in table 3. Nevertheless, both NES and NNES exhibited a good and acceptable rate of consistency in their judgments.

Table 4 shows the results for the relative task difficulty. All the values for the three tasks are close to each other and have a moderate difficulty level. The results show that the third task is the most difficult of all, and the second the easiest. The results suggest that the order of the questions must be changed to follow the general guideline of testing which mandates ordering the items from the easiest to the most difficult (Bachman, 2004). However, since the content of the second question is directly used in the third question, a better action would be adding to the difficulty level of the second task.

Table 4 – Relative Task Difficulty Results of MFRM

Measure Logit	Model Error	Infit MnSq Std	Task
0.25	0.02	5.70	3
-0.02	0.02	4.21	1
-0.23	0.03	4.70	2

Finally, table 5 shows the interaction of each of the two facets in PTBST, i.e. test-taker by rater, test-taker by task, and rater by task. The results contain two numbers. The second number is the instances of bias reported by the analysis, and the one on the right shows the number of significant bias instances (Bachman, Lynch, &

Mason, 1995; Linacre & Wright, 1999). The highest ration of significant bias is among the interaction between test-taker by rater. Two of these occurrences is due to one test-taker not answering a task, and hence getting a zero from both raters. Excluding this case from the analyses, two of the reported significant bias measures in person-by-rater and person-by-task will be eliminated, leaving the results with only one biased interaction between the test-taker and rater. The results show almost no biased interaction among the three facets of the measurement. This, along with the rest of the results in this section, show a high consistency of raters and unbiased interaction of three facets. In other words, PTBST has yielded reliable results.

Table 5 – Interaction Analysis Results of MFRM

Interactions	Person x Rater	Person x Task	Rater x Task
Biased Ratings	3/66	2/99	0/15

5.2 Correlation Analyses of METU EPE and GPA

To answer the first question and its sub-question of the study regarding the predictive power of EPE and its sub-sections, correlation analyses of METU EPE and first semester GPAs of the participants were calculated. As mentioned in the participants' section, of 34 participants, 31 gave their first semester grade reports along with their GPAs. Also, METU EPE scores of these participants along with their sub-scores for listening, reading, note-taking, writing, cloze test, and Dialogue and Situation were obtained directly from the participants (they opened their online EPE grade report page, and the scores were taken from this page). Table 6 shows the descriptive statistics of GPAs and the aforementioned scores.

A Shapiro-Wilk test of normality of distribution was conducted in SPSS (v. 22) and the results showed that the note-taking scores were not normally distributed. This was also visually confirmed by inspecting the Q-Q plot, which showed a non-sporadic pattern, hence indicating a non-normal distribution (Larson-Hall, 2010). Since, this does not satisfy the normality of distribution assumption of Pearson Product-Moment correlation (Bachman, 2004, p. 87), a test of Spearman Rank Order correlation was used to study the correlation between the note-taking section and GPAs. For the other variables, a Pearson Product-Moment test was conducted.

Table 6 –GPA and EPE Statistics of all Participants

Descriptive Statistics of GPAs, EPE Scores, and EPE Sub-Scores of All 31 Participants (Engineering and Non-Engineering)

Variable	Mean	STD	Shapiro-Wilk
GPA	2.20	1.09	.078
EPE	72.69	9.31	.086
EPE Listening	22.65	4.16	.456
EPE Reading	21.58	3.86	.277
EPE Note-Taking	3.97	0.78	.030
EPE Writing	10.89	1.32	.060
EPE Cloze Test	5.9	1.43	.632
Dialogue and Situation	7.71	1.48	.169

The results showed a significance correlation between the EPE scores and GPA (r=.461, p<0.01). The power analysis, conducted in RStudio v 0.99.879 (RStudio Team, 2015) through pwr package v 1.1-3 (Champely, 2015), showed a value of .53. According to Larson-Hall (2010, p. 105), this satisfies the minimum value of 0.5; however, the optimum value should be 0.8 (Larson-Hall, 2010, p. 105), which means that with a power size of 0.8, there is an 80 percent chance of finding this correlation, and to obtain such a value at a significance level of 0.01, 50 participants are needed, which is 19 more than the number of current participants. However, there was not a statistically significant correlation between the note-taking, listening, and writing sub-score of EPE and GPA. There were also significant correlations between GPA and reading (*r*=.471, p<0.01, power = 0.56; 48 participants needed for a power of 0.8), GPA and cloze test (*r*=.428, p<0.05, power = 0.69; 40 participants needed for a power of 0.8), and GPA and Dialogue and Situation (*r*=.466, p<0.05, power = 0.77; 34 participants needed for a power of 0.8).

Another set of correlation analyses were run after removing English 101 grades from first semester GPA. This was done to counter-balance the possible bias of English 101 grades as a result of construct/content similarities of English 101 tests and EPE. This variable will be called Weighted GPA (*M* = 2.03 and *STD* = 1.21). The results showed a significant correlation between EPE and weighted GPA (*r* = .394, p<.05, power = 0.61; 48 participants needed for a power of 0.8). The other significant correlations were between the weighted GPA and reading (*r* = .436, p<.05, power = 0.71; 39 participants needed for a power of 0.8), cloze test (*r* = .386, p<.05, power = 0.59; 50 participants needed for a power or 0.8), and dialogue and situation (*r* = .432, p<.05, power = 0.70; 39 participants needed for a power or 0.8). Once more, the listening, note-taking, and writing sections did not have a statistically significant correlation with the weighted GPA. All the sub-sections of EPE along with EPE itself had a lower correlation with weighted GPA than with GPA.

These results are similar to what Enginarlar (2012) reports for EPE results from METU Ankara. He summarizes all reports he had written up to that point and indicates that the correlation between EPE and first semester GPA ranged between .45 and .55. The results in this study echo the same findings. However, he does not report any correlations between the subsections of EPE and GPA. The results here also confirm what Burgess and Greis' (1970) report that, with the presence of English 101 grades in GPA, English proficiency tests show a higher correlation with GPA.

5.3 Standard Multiple Regression Analyses Between EPE and GPA

Correlation statistics are used to explore the pattern of relationship between variables and does not show any causality (Larson-Hall, 2010, p. 148). In other words, in order to study the predictive power of an independent variable, and more importantly, to investigate the amount of variance in the dependent variable predicted by each one of independent variables, if there is more than one, a regression analysis should be used. For instance, in order to investigate the predictive power of TOEFL iBT scores regarding the academic success of the participants, Cho and Bridgeman (2012) conducted a stepwise multiple regression. However, Larson-Hall (2010) does not recommend this type of multiple regression when the unique contribution of the predictive variables is of interest. She recommends a standard multiple regression which is more powerful and reveals the unique predictive power of each variable while treating them all at the same level. Therefore, since a part of this study investigates the predictive power of the EPE and its sub-sections as independent variables in predicting the variance in GPA and weighted GPA (GPA with English 101 factored out) as the dependent variables, two regression analyses were conducted. Both are standard multiple regressions with one having the total GPA as its dependent variable and the other the weighted GPA.

Prior to performing the standard multiple regressions, the correlation scatterplot matrices were visually inspected and the assumption of linearity of correlation were satisfied. Moreover, the P-P plot of residuals showed a linear pattern which indicated the satisfaction of normality assumption for both standard regressions. Also, Larson-Hall (2010, 196) recommends examining the values for standard residuals to see if there are any values above 3.0 or below -3.0. Both regression analyses also satisfied this criterion. Finally, a look at Cook's distance showed that there were no outliers in the model, with all the values between 0 and 1 (Larson-Hall, 2010, 196). Table 7 shows the correlation matrix of EPE section scores for all the participants (engineering and non-engineering) and their total GPA. It is worth mentioning that the correlation matrix reported in the regression analysis output is one-tailed (Larson-Hall, 2010), so some correlations reported to be statistically insignificant in *Correlation Analyses for METU EPE and GPA* section might be statistically significant here. However, since both the upper and lower sides of the correlations are of importance in this study, the two-tailed correlation values are the valid ones.

Table 7 – All Group GPA and EPE Correlation Matrix

Correlation Matrix of EPE Sections and Total GPA for All Participants (Engineering and Non-Engineering)

	GPA	Listening	Reading	NoteTaking	Writing	Cloze
Listening	.313* .043 31					
Reading	.471* .004 31	.721* .000 31				
NoteTaking	.297 .052 31	.216 .122 31	.023 .451 31			
Writing	-.278 .065 31	.090 .316 31	.013 .472 31	.294 .054 31		
Cloze	.428* .008 31	.673* .000 31	.552* .001 31	.293 .055 31	-.081 .333 31	
DialogandSituation	.466* .004 31	.219 .119 31	.194 .148 31	.387* .016 31	.060 .375 31	.230 .106 31

Values with * are significant at p<.05. The sections of EPE are written in the same order administered in the actual test. The significance values are one-tailed.

There were correlations between GPA and four sections of EPE (listening, reading, cloze, and dialogue and situation), with reading section having the highest correlation (*r* = .471). There were also correlations among the explanatory variables, and the highest was between reading and listening sections (*r* = .721) followed by that between cloze test and listening sections (*r* = .673). It must be noted that theses correlation statistics are of one-tailed type, and this type of correlation is common when reporting the results of regression analyses (Larson-Hall, 2010). These results are similar to those reported by Enginarlar (2012). However, there are two differences between the findings of this study and those of Enginarlar's (2012). First, in his report, the writing section has a moderate correlation with all the sections of EPE, while in this study it is not so, yet he reports that the lowest inter-componential correlation in EPE sections belong to that of the writing and reading sections, which is mirrored in this study as well. He also reports that writing is one of the two sections which have the lowest correlation with the total EPE scores. Therefore, the difference in his findings and the ones in this study is related to only the magnitude of the correlation coefficients, which are smaller in this study. Second, the note-taking section also has a moderate correlation with all the sections of EPE in Enginarlar's (2012); however, in this study it is not the case. In other words, there is almost no correlation between the note-taking and the reading sections. Enginarlar (2012) reports that the note-taking section is the other section of the two in EPE with the lowest correlation with the total EPE score, and perhaps this is why this section, like the writing, exhibits such a low correlation with the other sections of EPE. Again, the real difference between the results of this study and those in Enginarlar's (2012) relates to the magnitude of the correlations, and the pattern is still the same.

The regression model was statistically significant ($F_{6,\ 24}$ = 4.790, p = 0.002), with a total R^2 = .545. In other words, the model explained 54.5 percent of variance in Total GPA. Also, the two significant factors with unique contribution to regression were reading (B = .145, sr^2 = .341, CI = .024, .267, p = .021) and writing (B = -.296, sr^2 = -.329, CI = -.552, -.40, p = .025). In other words, reading explained 34.1 percent and the writing section 32.9 percent of variation in total GPA. However, these two sections are at opposite direction, and they seem to be canceling each other out, giving a murky picture of the relationship between EPE and GPA. Table 8 shows the coefficient statistics of all the explanatory variables in the model. It is worth mentioning that the Reading sections has a high correlation with listening and cloze test sections. Therefore, an issue of collinearity can render the unique contribution of these mentioned factors statistically insignificant. Results of a standard regression shows the unique contribution of each factor to the model, and when there is a significant amount of overlap among explanatory variables, their contribution can be obscured in the model. In general, this analysis did not provide a desirably clear picture of the predictive power of EPE sub-sections regarding the GPA. This is because while the writing section does not have a high correlation with the GPA, it is a significant predictor factor in the standard multiple-regression model.

Table 8 – Coefficient Statistics of EPE sections and GPA of all Participants

Coefficients table of Standard Multiple-Regression of EPE test sections and Total GPA for All Participants (Engineering and Non-Engineering)

	Unstandardized Coefficients		Std. Coef.		95.0% Confidence Interval for B		Corr.	Collinearity Statistics	
	B	Std. Error	Beta	Sig.	Lower Bound	Upper Bound	sr^2	Tolerance	VIF
(Constant)	-.173	1.621		.916	-3.517	3.172			
Listening	-.061	.061	-.234	.322	-.187	.064	-.139	.355	2.815
Reading	**.145**	**.059**	**.514**	**.021**	**.024**	**.267**	**.341**	**.439**	**2.276**
NoteTaking	.404	.234	.290	.098	-.080	.888	.237	.669	1.494
Writing	**-.296**	**.124**	**-.358**	**.025**	**-.552**	**-.040**	**-.329**	**.845**	**1.184**
Cloze	.091	.153	.119	.558	-.225	.406	.082	.472	2.120
Dialog and Situation	.221	.113	.299	.062	-.012	.455	.269	.812	1.231

5.4 Standard Multiple Regression Analyses Between EPE and weighted GPA

Now, the regression model statistics between EPE sections and weighted GPA will be explored. Table 9 is the correlation matrix of this model. As mentioned before, the reported correlation statistics in table 9, which is reported in regression statistics, are of one-tailed type. However, the two-tailed correlation values are the accepted statistics for in this study. Therefore, the reported two-tailed correlation statistics in *Correlation Analyses of METU EPE and GPA* have the priority, and the statistics reported in table 9 are just for the sake of complete reporting of regression statistics.

There were also correlations between weighted GPA and EPE sections, with reading sections having the highest correlation (r = .436), again, followed by dialogue and situation section (r = .432). EPE sections also correlated among each other with reading and listening having the highest correlation (*r* = .721) followed by that between listening and cloze test (*r* = .673).

Table 9 – All Groups Weighted GPA and EPE Correlation Matrix

Correlation Matrix of EPE Sections and Weighted GPA for All Participants (Engineering and Non-Engineering)

	WGPA	Listening	Reading	Note Taking	Writing	Cloze
Listening	.294 .076 31					
Reading	.436* .007 31	.721* .000 31				
NoteTaking	.194 .148 31	.216 .122 31	.023 .451 31			
Writing	-.345* .029 29	.090 .316 31	.013 .472 31	.294 .054 31		
Cloze	.386* .016 31	.673* .000 31	.552* .001 31	.293 .055 31	-.081 .333 31	
Dialogue and Situation	.432* .008 31	.219 .119 31	.194 .148 31	.387* .016 31	.060 .375 31	.230 .106 31

Values with * are significant at p<.05. The sections of EPE are written in the same order administered in the actual test. The significance values are one-tailed.

This regression model was also statistically significant (R^2 = .503, $F_{6, 24}$ = 4.046, p = 0.006). The total variance explained in this model is lower than that in the previous one. That is, subsections of EPE together explained 50.3 percent of variation in weighted GPA. Nonetheless, this is a high value. In this model, the reading and writing sections were the two significant factors contributing to the model (similar to the previous model with GPA as the dependent variable). Table 10 shows the coefficient statistics of the standard multiple regression between EPE sections and weighted GPA. The coefficient statistics of the reading section (B = .151, sr^2 = .320, CI = .011, .292, p = .036) show that it explains 32 percent of the variance in the weighted GPA. Also, the coefficient statistics of the writing section (B = -.363, sr^2 = -.364, CI = -.659, -.067, p = .018) indicate a 36.4 percent of variation in the weighted GPA. Again, like the previous regression analysis, the opposite directions of these two significant predictors do not give us a clear picture of how these factors collectively can predict the variance in the weighted GPA. In other words, it is counter-intuitive to have a section in a proficiency test which negatively affects its predictive power in terms of academic success. As mentioned earlier, proficiency tests and their sub-sections tend to show different pattern of predictive power for engineering and non-engineering disciplines (Al-Musawi & Al-Ansari, 1999; Ayers & Quanttlebaum, 1992, as cited in Cho & Bridgeman, 2012; Vinke & Jochemes, 1993; Wait & Gressel, 2009). So, looking at the pattern of predictive power from this point of view might give a clearer picture about why, for instance, the writing sections has a negative correlation with the total and weighted GPAs. Hence, to further scrutinize this matter, the participants were divided into two engineering and non-engineering groups to see if a clearer picture could be obtained. Before, moving on to analyzing participant scores separately based on subject areas, the correlation of the PTBST with the total and weighted GPA was also conducted and was compared to EPE and its sub-sections scores.

Table 10 – Coefficient Statistics of EPE sections and Weighted GPA of all Participants

Coefficients table of Standard Multiple-Regression of EPE test sections and Weighted GPA for All Participants (Engineering and Non-Engineering)

	Unstandardized Coefficients		Std. Coef.		95.0% Confidence Interval for B		Corr.	Collinearity Statistics	
	B	Std. Error	Beta	Sig.	Lower Bound	Upper Bound	sr^2	Tolerance	VIF
(Constant)	.534	1.876		.778	-3.337	4.405			
Listening	-.070	.070	-.240	.331	-.215	.075	-.143	.355	2.815
Reading	**.151**	**.068**	**.483**	**.036**	**.011**	**.292**	**.320**	**.439**	**2.276**
Note-Taking	.302	.271	.196	.277	-.258	.862	.160	.669	1.494
Writing	**-.363**	**.143**	**-.396**	**.018**	**-.659**	**-.067**	**-.364**	**.845**	**1.184**
Cloze	.101	.177	.120	.572	-.264	.466	.083	.472	2.120
Dialogue and Situation	.255	.131	.311	.063	-.015	.525	.280	.812	1.231

5.5 Correlation Analyses of PTBST and GPA

Having ensured that PTBST scores satisfy the assumptions of Pearson Product-Moment correlation, two analyses of correlation were performed. The first analysis between PTBST and GPA showed a significant correlation (r = .430, p<0.05, power =

0.7; 40 participants needed for a power of 0.8). The weighted GPA also showed a significant correlation with PTBST (r = .364, p<0.05, power = 0.54; 57 participants needed for a power of 0.8). PTBST showed a higher correlation with the total GPA compared to the listening, note-taking, writing, and cloze test. However, it had a lower correlation with GPA than EPE, the reading, and the dialogue and situation sections. The results are the same with the weighted GPA, with the exception of the cloze test which has a higher correlation with the weighted GPA than the PTBST. Table 11 shows the summary of the correlation analyses between total and weighted GPAs on the one hand and PTBST, EPE, and EPE sub-sections on the other for all the participants (engineering and non-engineering).

Table 11 – Correlations of PTBST, EPE, and EPE Sections with all Group Total and Weighted GPAs

Summary of Correlation Analysis Between GPA, PTBST, EPE and EPE Sub-Sections for All Participants (Engineering and Non-Engineering)

GPA	PTBST	EPE	Listening	Reading	Note-taking[1]	Writing	Cloze-Test	D&S
Total	**.430** p = .016 31	**.461** p = .009 31	**.313** p = .086 31	**.471** p = .007 31	**.297** p = .105 31	**-.278** p = .130 31	**.428** p = .016 31	**.466** p = .008 31
Weighted	**.364** p = .044 31	**.394** p = .028 31	**.264** p = .151 31	**.436** p = .014 31	**.194** p = .295 31	**-0.345** p = .057 31	**.386** p = .032 31	**.432** p = .015 31

[1] Except for this, all the other correlations are of Pearson Product-Moment. Correlations reported for the note-taking scores are Spearman rho. Statistically significant correlations are indicated by the shaded cells.

Once more, the results seem to indicate an inconsistency, especially for the productive skills, e.g. note-taking, writing, and dialogue and situation sections. In other words, why would the dialogue and situation section which is a test of pragmatic understanding and elicits a controlled response have a better correlation with both total and weighted GPAs of all the participants than the writing or note-taking sections which are more characteristic of skills required of these students in their academic studies? Since there are difference in skills needed for success in different academic disciplines, separate analysis of correlation for engineering and non-engineering disciplines may provide a clearer picture of the predictive power of the mentioned productive skills. This is shown in the literature through the different pattern of predictive power tests like TOEFL, FCE, and GRE show for academic success in different disciplines, e.g. engineering vs. non-engineering (Al-Musawi & Al-Ansari, 1999; Ayers & Quanttlebaum, 1992, as cited in Cho & Bridgeman, 2012; Vinke &

Jochemes, 1993; Wait & Gressel, 2009). This, along with the unclear picture obtained from the standard multiple-regression analyses (the significant, but negative, coefficient of the writing section), prompted the researcher to analyze engineering and non-engineering participants separately and compare the results with each other. The details are in the next section.

5.6 Correlation Analyses for Non-Engineering Students

As mentioned earlier, in order to get a better picture of how each section of EPE predicts the academic performance of the first-semester students, a correlation between the first semester total and weighted GPA of the non-engineering students and their scores in PTBST, EPE, and EPE sub-sections were analyzed. The new set of data satisfied the assumptions of Pearson Product-Moment correlation test. Table 12 summarizes the correlation statistics for non-engineering students.

Table 12 – Correlations of PTBST, EPE, and EPE Sections with all Non-Engineering Total and Weighted GPAs

Summary of Correlation Analysis Between GPA and PTBST, EPE and EPE Sub-Sections for Non-Engineering Students

GPA	PTBST	EPE	Listening	Reading	Note-taking	Writing	Cloze-Test	D&S
Total	**.506** p = .032 18	**.355** p = .148 18	**.268** p = .283 18	**.357** p = .146 18	**.252** p = .312 18	**-.472** p = .048 18	**.417** p = .085 18	**.486** p = .041 18
Weighted	**.428** p = .076 18	**.311** p = .209 18	**.226** p = .367 18	**.326** p = .187 18	**.200** p = .426 18	**-0.512** p = .030 18	**.380** p = .120 18	**.497** p = .036 18

Statistically significant correlations are indicated by the shaded cells. There are 18 non-engineering students in this study.

The results show that PTBST has the highest correlation with the total GPAs of non-engineering participants, and it is followed by the dialogue and situation and writing sections. The power analysis of the correlation between PTBST and non-engineering GPAs showed a value of 0.66, and 25 participants are needed to achieve a power of 0.8, which is seven more than the number of current participants. Also, the power analysis of the correlation between the dialogue and situation and the GPA showed a value of 0.56, yet to obtain the optimum value of 0.8, 31 participants are needed. While PTBST still has a good correlation with the weighted GPA, albeit statistically non-significant, it is the dialogue and situation section which shows a higher correlation with the weighted GPA, and it is interesting that the dialogue and situation correlates better with the weighted GPA than the total GPA for non-engineering students. A power analysis of the correlation between PTBST and the weighted GPA showed that in order to obtain a statistically significant correlation with a power of 0.8, 40 participants are needed. The same power value for the correlation between dialogue and situation section and the weighted GPA yielded a value of 0.58. Again this was above the cut-off point of 0.5, which is a satisfactory level, but to obtain an optimum power of 0.8, 29 participants are needed. The dialogue and situation section of EPE elicits pragmatic understanding in a written form. In other words, a test-taker must take into consideration the context and co-text provided in the prompts, which are unfinished conversations or communicative situations, in order to use the language in meaningful and appropriate way to supply a relevant response. The fact that both dialogue and situation and PTBST show a good correlation for non-engineering students with the total and weighted GPA, clearly attests to the importance of the language use construct, rather than language knowledge, in academic success in these disciplines. Moreover, comprehension alone might not provide much information about the academic success of non-engineering students as the correlations of the reading and listening sections with both the total and weighted GPAs are low and statistically insignificant. As to the writing, while it has a negative correlation with the total GPA (power = 0.53; 33 participants needed for a power of 0.8), it has an even bigger negative correlation with the weighted GPA ($r = -0.512$, $p < 0.05$ power = 0.61; 27 participants needed for a power of 0.8). This means that productive skills without an element of synthesizing may not comprise the academic English proficiency construct for non-engineering students. Therefore, from these results, it can be assumed that testing skills in a separate manner does not yield much information about the academic success of non-engineering students, and it is the sections of EPE, which combine comprehension and production in a task, that provide the desired information, as does PTBST. One might argue that the note-taking section also combines comprehension and production (listening and writing), and if combining comprehension and production tasks provide good predictive information on academic success of non-engineering students, why is it not so for the note-taking? To further scrutinize this, the correlation analyses between GPAs of non-engineering students and the scores of three tasks of PTBST were conducted. Table 13 summarizes the results.

Table 13 – PTBST Tasks Correlations with Non-Engineering Total and Weighted GPA

Summary of Correlation Analysis Between GPA and Three Tasks of PTBST for Non-Engineering Students

GPA	Task 1	Task 2	Task 3
Total	**.360** p = .142 18	**.163** p = .518 18	**.539** p = .021 18
Weighted	**.313** p = .205 18	**.150** p = .552 18	**.441** p = .067 18

Statistically significant correlations are indicated by the shaded cells. There are 18 non-engineering students in this study.

Task three of PTBST incorporates the personal ideas of the test-taker into material previously presented in task two. In other words, this task does not entail a direct reporting of the comprehended input, as note-taking does. The fact that task three of PTBST shows a high correlation with the GPA of non-engineering students (r = .539, power = 0.67; 24 participants needed for a power of 0.8) shows that language production which results from a mixture of input and personal ideas is what contributes to the success of non-engineering students, not just talking or writing about personal ideas. This further confirms that having speaking tests that only elicit personal ideas (as it is the case with the current formative and summative speaking tests at METU NCC SFL) may not necessarily reflect the nature of language use construct needed to succeed in English-medium universities, at least for non-engineering students. It is worth mentioning that the correlation of the third task of PTBST with the weighted GPA is statistically insignificant, and, in fact, to obtain a statistically significant correlation between the mentioned variables, 38 participants are needed.

The analyses in this section not only provided a clearer picture of how powerful the sub-sections of EPE are in predicting the academic success of non-engineering students judging by their first semester total and weighted GPA, but also showed that, as a whole, EPE exhibits a much lower and statistically insignificant correlation with the total GPA than PTBST. The same analysis was carried out for engineering students to get a similarly clear picture of the predictive power that PTBST, EPE, and EPE sub-sections have regarding the first-semester total and weighted GPAs.

5.7 Correlation Analyses for Engineering Students

The data for engineering participants were also tested to see if they satisfy the assumptions of Pearson Product-Moment correlation analysis. All the data satisfied the assumptions. Table 14 summarizes the correlation values for engineering students.

Table 14 – Correlations of PTBST, EPE, and EPE Sections with all Engineering Total and Weighted GPAs

Summary of Correlation Analysis Between GPA and PTBST, EPE and EPE Sub-Sections for Engineering Students

GPA	PTB ST	EPE	Listening	Reading	Note-taking	Writing	Cloze-Test	D&S
Total	**.376** p = .205 13	**.734** p = .004 13	**.495** p = .085 13	**.724** p = .005 13	**.402** p = .174 13	**.021** p = .944 13	**.471** p = .105 13	**.508** p = .077 13
Weighted	**.323** p = .281 13	**.626** p = .022 13	**.421** p = .152 13	**.700** p = .008 13	**.224** p = .461 13	**-.078** p = .800 13	**.422** p = .151 13	**.402** p = .173 13

Statistically significant correlations are indicated by the shaded cells. There are 13 engineering students in this study.

EPE shows a very high correlation with the total (r = 0.734, $p < 0.01$, power = 0.66; 16 participants needed for a power of 0.8) and weighted GPA of engineering students (r = 0.626, $p < 0.05$, power = 0.67; 17 participants needed for a power of 0.8), while PTBST has a lower correlation with GPAs and seem to be a weak predictor of academic success for engineering students, with a statistically insignificant correlation. However, according to power analysis, 53 participants are needed for a statistically significant correlation ($p < 0.05$; power = 0.8), and probably this lack of significant correlation is due to a low number of participation. The reading section of EPE shows the highest correlation with both total (r = 0.724, $p < 0.01$, power = 0.64; 17 participants needed for a power of 0.8) and weighted GPAs (r = 0.7, $p < 0.01$, power = 0.58, 19 participants needed for a power of 0.8) for engineering students. There is also no section that shows a higher correlation with the weighted GPA than with the total GPA, which is what to be expected as language is not as salient a factor in engineering subjects as non-engineering ones.

One important difference in the correlation patterns of the engineering students compared to their non-engineering counterparts is that comprehension skills play an important role in the predictive power of the proficiency test, but what about the sections that combine comprehension and production skills?

The dialogue and situation has a good correlation with the total GPA, but this correlation drops significantly when the weighted GPA is the dependent variable. As to the note-taking, we witness a similar drop. It seems that productive skills do not play a significant role in capturing the construct needed for success in non-English courses of engineering students. To further investigate this, the correlation statistics of the total and weighted GPAs of the engineering students with their scores in three tasks of PTBST were scrutinized. Table 15 contains the results.

Table 15 – PTBST Tasks Correlations with Engineering Total and Weighted GPA

Summary of Correlation Analysis Between GPA and Three Tasks of PTBST for Engineering Students

GPA	Task 1	Task 2	Task 3
Total	**.369** p = .215 13	**.407** p = .167 13	**.195** p = .524 13
Weighted	**.290** p = .336 13	**.273** p = .368 13	**.237** p = .435 13

There are 13 engineering students in this study. There is no statistically significant correlation in this table.

All the tasks in PTBST test the productive skills of the test-takers, and a careful comparison of the three tasks can yield a good picture of the language use construct underlying academic success of engineering students. While task three was the strongest predictor for non-engineering students, it is the weakest for their engineering counterparts. However, it is the opposite for the second task. Task two had the lowest correlation with the total and weighted GPAs of the non-engineering students, while it has a higher correlation with the weighted GPA, albeit statistically insignificant, than that of the third task. The power analysis showed that to achieve a statistically significant value for the correlation of the second task with the GPA of engineering students with a power of 0.8 ($p < 0.05$), 45 participants are needed. The difference between tasks two and three is in the type of synthesis elicited. In the second task, a direct reporting is required of the test-taker without adding any personal ideas. Yet, the third task introduces the personal idea to the synthesis. As a result, taking into consideration these results and those from table 14, it can be assumed that in general, receptive skills (listening and reading) are the strongest predictors of the academic success of the engineering students. Also, as long as the synthesizing construct is limited to direct reporting, it can be an important factor contributing to the success of the engineering students, which is successfully captured in the first and the second tasks of PTBST. This might be the reason why the note-taking section of EPE shows a higher correlation with the GPAs of engineering students than the writing section since the note-taking section also elicits a direct reporting of input. However, the correlation of the note-taking section with the GPAs of the engineering students is not statistically significant. If this is due to the low number of participants, according to the power analysis, 46 participants would be needed to achieve a statistically significant correlation with the same value reported here with a power of 0.8. Moreover, it is noteworthy that the note-taking section does not entail synthesizing two sources of information for a direct reporting, unlike the second task of PTBST, and larger sample size can show a better picture of a difference, if there is any, between the direct reporting from one source of input and direct reporting through synthesizing two sources of information. Yet, it can be argued that language production items based solely on personal ideas (e.g. the writing sections of EPE and the formative and summative speaking assessments at METU NCC SFL) may not be powerful predictors of academic success for engineering disciplines, as well as non-engineering ones.

So far, all the arguments were based on the underlying construct of different sections of EPE and three tasks of PTBST. However, these claims must be evidenced by empirical data. To this end a factor analysis was conducted.

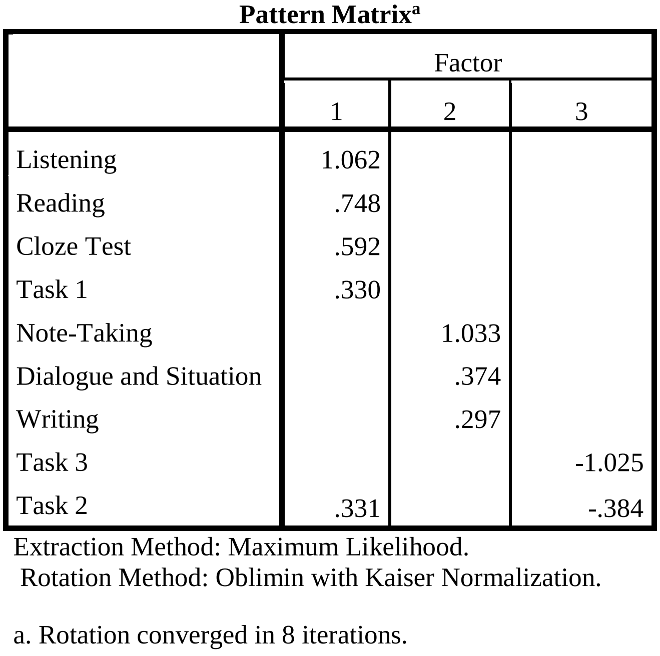
5.8 Factor Analysis

Another important matter to probe in order to justify adding PTBST to EPE is the construct exclusiveness of these two tests. One way of doing this is through theories of language learning (Bachman, 1990; Bachman, 2004; Bachman, 2007). This is a crucial part of developing and supporting validity arguments of a test. However, Bachman (2004, p. 279) argues that statistical analyses like factor analysis are also required if conducting them is possible in order to justify the test construct. To this end, an exploratory factor analysis was conducted to see if sub-scores of EPE load onto factors different from those PTBST tasks do. To this end, an exploratory factor analysis was conducted through maximum likelihood extraction method with an Eigen value of 1. According to Yong and Pearce (2013), this method of extraction is useful if a confirmatory factor analysis is going to be performed. Furthermore, an oblique rotation was chosen for this analysis since the factors can correlate with one another. There are two oblique rotation methods: *Direct Oblimin* and *Promax*. *Direct Oblimin* rotation is the most apt since *Promax* is used for large sample sizes (Yong & Pearce, 2013), which is obviously not the case in this study.

Bartlett's test of sphericity results showed that the model satisfies one of the assumptions (χ^2 = 84.169, p = 0.000). Moreover, Kaiser-Meyer-Olkin measure of sample adequacy statistics was above the cut-off point of 0.5 suggested by Yong and Pearce (2013). However, there were 13 cases of (36%) non-redundant residuals with absolute values greater than .05 which is above the cut-off point of 10 percent recommended by Yong and Pearce (2013), and this one last evidence of the adequacy of the sample was not satisfied. Nevertheless, with the evidence obtained from the Kaiser-Meyer-Olkin measure of adequacy statistics, the analysis was carried on. Table 16 shows the factor loading values in the pattern matrix with a *direct Oblimin* rotation and *maximum likelihood* extraction. Yong and Pearce (2013) recommend hiding factor loading values below .32; however, since no factor loading was shown for the writing section, to have a better picture, this value was reduced to .296.

Table 16 – Factor Analysis Pattern Matrix

Pattern Matrix of factor analysis of EPE sections and PTBST Tasks



Pattern Matrix[a]

	Factor		
	1	2	3
Listening	1.062		
Reading	.748		
Cloze Test	.592		
Task 1	.330		
Note-Taking		1.033	
Dialogue and Situation		.374	
Writing		.297	
Task 3			-1.025
Task 2	.331		-.384

Extraction Method: Maximum Likelihood.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 8 iterations.

Listening, reading, and cloze test sections of EPE along with the first and second tasks of PTBST are the sections with loadings on the first factor. In both reading and listening sections, test-takers are asked to answer multiple-choice and open-ended questions directly related to the content of the listening and reading passages. Cloze tests, on the other hand, ask test-takers to fill gaps with only one word in a reading passage. It is clear that listening and reading involve comprehension from two different channels, and cloze test relies on the reading comprehension skills of test-takers at clausal and inter-clausal levels. In other words, in order to answer cloze test items correctly, test-takers must attend to both grammatical and semantic clues at sentence level, as well as coherence and cohesive clues at inter-clausal levels (Bachman, 1982). The same skills are required for successful comprehension of relatively long reading and listening passages. This might be one reason why the cloze test section of EPE has such a high loading on this factor. Moreover, the first and second tasks of PTBST have loadings on this factor as well. As mentioned earlier, these two tasks rely on listening as well as reading comprehension, and involve a form of direct reporting. However, there is difference between the first and the second tasks of PTBST which is evident from the secondary loading of the second task on the third factor. It can be argued that the first task does not involve any synthesizing, and it is all about direct reporting from a single source, i.e. listening in this case. However, the second task involves direct reporting through synthesizing the input from two sources, i.e. listening and reading. This is probably why the second task has a loading on the third factor. Therefore, it can be said that this factor can represent language comprehension skills, and speaking questions with no synthesizing as part of their construct can probably contribute nothing more to what different sections of EPE already cover.

Note-taking, dialogue and situation, and writing sections have loadings on the second factor, with note-taking having the highest loading. These three sections in EPE elicit writing production in three different ways. In the note-taking section, test-takers are asked to listen to a five-minute lecture and take notes on the main points. Then, a question is asked related to one of the main ideas of the talk, and test-takers are asked to write a short paragraph to answer the question. In Dialogue and situation section, hypothetical communicative situations are described and test-takers are asked to write a proper response taking into consideration the context and co-text. In the writing section, test-takers are asked to write a one-paragraph argumentative essay about a common issue. Since, all these involve a writing production, this factor can be called writing competence. Enginarlar (2012) reports that these three sections have the lowest correlation with the total EPE compared to the other three sub-sections of EPE, and this is probably why these three sections tend to load on the same factor due to a common feature among them, i.e. low correlation with the total EPE scores. However, in the previous sections, the correlation analyses showed that while note-taking correlates better with the engineering student GPAs, it is the writing section that correlates well with non-engineering student GPAs, albeit negatively. It was hypothesized that writing without an element of synthesis and involving a lot of personal ideas might be appropriate language production test items neither for engineering nor non-engineering students.

The third task of PTBST has a high loading on the third factor along with a secondary loading of the second task of PTBST. As mentioned earlier, both the second and third tasks of PTBST entail a form of information synthesis. In other words, both these tasks draw on two sources of input (one written and one spoken) and ask the test-taker to combine the information from these two sources. However, there must be a difference between the second task and the third one in PTBST; otherwise, the third task would show a similar loading pattern to the second task. The difference is that while the second task elicits synthesizing information, it still relies on direct reporting. Yet, the third task elicits synthesizing information with an added dimension of interweaving personal ideas into the synthesized response. The results in the previous sections showed that while the second task of PTBST may be appropriate for engineering students, the last one can be more apt for non-engineering ones as the construct in the last task is more typical of non-engineering courses.

CHAPTER 6

SUMMARY OF RESULTS

In this section, the summary of the results is presented, and the ideas are organized by the research questions posed earlier in this report.

The first research question asked in this study was how powerful METU EPE is in predicting the academic success of those students who pass this test and, hence, are allowed to start their studies at their faculties. The answer to this question is not different from what Enginarlar (2012) reports regarding the predictive power of METU EPE and first semester GPA. Also, the results of the correlation with the weighted GPA corroborated Burgess and Greis' (1970) findings that English proficiency tests have a lower correlation with weighted GPAs when English 101 grades are taken out of them.

As to question 1-a, which asked whether there was any section of EPE which had a more predictive power regarding the first-semester GPA, reading, cloze test, and dialogue and situation were the sections which predicted GPA better. Listening, note-taking, and writing sections did not exhibit a high predictive power with the GPAs of engineering and non-engineering students combined. Enginarlar (2012) does not report any correlation statistics between the EPE sections and the first-semester GPA. However, he reports that the writing and note-taking sections do note correlate well with the total EPE score, and perhaps this is why these two sections do not exhibit a high predictive power regarding the total and weighted GPAs. Also, the reading, cloze test, and dialogue and situation sections of the EPE showed a good correlation with the weighted GPA. Once more, the listening, note-taking, and writing sections did not correlate well with the weighted GPAs.

To get the unique predictive power of each of the sections of EPE in terms of both total and weighted GPA, two standard multiple-regression analyses were run. In both, the reading and writing sections had significant unique predictive power after factoring out the common predictive indices of all the sections. These results were contradictory because the writing section did not show any statistically significant correlation with total and weighted GPA. Moreover, the negative coefficient of the writing was surprising as writing seems to be a staple part of any academic studies. It was hypothesized that the contradictory images yielded in theses sections may stem from the difference in predictive power pattern that English proficiency tests show for different disciplines (Al-Musawi & Al-Ansari, 1999; Ayers & Quanttlebaum, 1992, as cited in Cho & Bridgeman, 2012; Vinke & Jochemes, 1993; Wait & Gressel, 2009). Therefore, the participants were divided into two groups: engineering and non-engineering. Then, correlation analyses were rerun. This will be discussed after the summary of factor analysis.

The second research question asked how the predictive power of PTBST compared to that of EPE and its sub-sections for the whole group (engineering and non-engineering participants combined). The results showed that PTBST had a less predictive power than EPE in terms of both total and weighted GPAs. Also, PTBST correlated better with the total than the weighted GPA, and this again corroborates Burgess and Greis' (1990) findings that proficiency tests correlate better with GPAs that have English course grades in them. Still, PTBST correlated better than the listening, note-taking, writing, and cloze test sections of EPE with the total GPA of engineering and non-engineering participants combined. It also correlated better than the listening, note-taking, and writing with the weighted GPA. The reading and dialogue and situation sections correlated better with both GPA and weighted GPA than the PTBST, and the cloze test was better than PTBST when it came to the weighted GPA. In general, PTBST showed a slightly lower correlation with the total GPA of all participants (engineering and non-engineering combined) than the total score of EPE (0.031 units of Pearson Product-Moment correlation less), however, the 0.7 value of power for the correlation of PTBST showed that there is a 70 percent chance that such a statistically significant correlation can be found (Larson-Hall, 2010) compared to 53 percent (power = .53) of the correlation found for the total EPE scores. As to the other sections of EPE which had a higher correlation than PTBST with the total GPA (reading and dialogue and situation), it was the dialogue and situation section whose correlation had a more statistical power than that of PTBST (power = 0.77). Yet, when it came to the weighted GPA, all the sub-sections of EPE having a higher correlation than PTBST also had a higher statistical power ranging from 0.59 to 0.71 compared to the 0.54 power statistic of the correlation between PTBST and the weighted GPA. It seems that EPE and some of its sections, i.e. reading, cloze test, and dialogue and situation are better predictors of weighted GPA for all the participants combined. Yet, PTBST is not far behind and still shows a relatively good correlation with weighted GPA, a good correlation with GPAs with a power close to the optimum level of 0.8. The correlations of the engineering and non-engineering participants separately showed that this is not always the case, and PTBST can actually be a better predictor for non-engineering disciplines than EPE, but not much so for the engineering disciplines.

The answer to question 2-a, which asked whether PTBST tasks measure a different construct from EPE, is positive. Using the results of the factor analysis, it can be argued that PTBST tasks measures a different construct than EPE. This statistical test showed that the third task of PTBST loaded on a completely different factor. One of the loadings of the second task also shared this factor, which had no loading from the different sections of METU EPE. The other loading of this task was on the first factor, which was called the comprehension factor as the reading and listening sections of the EPE had high loadings on this factor. Finally, the first task of PTBST had one single loading which was, again, on the first factor. With a higher value of loadings on a third factor which did not have any loadings from EPE sections, it can be concluded that PTBST measures a different construct. Although this is theoretically obvious as PTBST is a test of speaking, empirical evidence is also necessary to support it. Moreover, the factor analysis provided further support to the ideas inferred from the analyses in which engineering and non-engineering students were separated.

The last research question pertains to the difference of predictive power pattern for engineering and non-engineering disciplines between PTBST and its tasks along with EPE and its sub-sections on the one hand and total and weighted GPAs on the other. The correlation statistics showed that PTBST had a higher correlation with the GPA of non-engineering students compared to EPE and its sub-sections. While writing section of EPE had a high, but negative, correlation with both total and weighted non-engineering GPAs, the dialogue and situation had the second highest correlation with non-engineering GPAs after PTBST, and the highest with the weighted engineering GPAs. A further look at the correlation of the PTBST tasks showed a high and statistically significant correlation of task three of PTBST and the total non-engineering GPAs. However, the correlation of this task and the weighted non-engineering GPAs was lower and statistically insignificant. In general, it was concluded that separating comprehension and productive skills and testing them separately may not be powerful enough to capture the language use construct in an

English-medium academic setting for non-engineering students. Therefore, synthesizing, which is operationalized in PTBST, may be needed to better capture the underlying language ability construct needed for success in English-medium settings for non-engineering students. It was further concluded that the synthesis must have at least two sources of input and must involve going beyond a mere direct reporting. In other words, synthesizing two sources of input and personal ideas, and presenting the resultant message in spoken (and probably written form) may be better suited for testing academic English proficiency of non-engineering students.

As to the engineering students, METU EPE and its reading section were the best predictors of academic success, measured by both total and weighted GPAs. This is different from what Wait and Gressel (2009) speculate. They believe that English language skills like reading and writing seem to have less predictive power for engineering disciplines, but the results of this study shows that this might not be the case at least for the reading skills, as the reading section of EPE shows to be a strong predictor of academic success for these disciplines. However, this conclusion cannot be certain due to the low number of participants. Moreover, even if a replication of this study with a high number of participants from engineering discipline shows the same results, this might be due to student study skills and preference variable. In other words, it might be that the participants in the study from engineering disciplines value input from the reading sources more and prefer to learn from them, and this might, in turn, be spurred by the specific curriculum design and instructor preferences prevalent in theses disciplines (Wait & Gressel, 2009).

Looking at the other high correlations (even though statistically insignificant), listening, note-taking, cloze test, and dialogue situation sections had relatively high correlation with GPA of engineering students. As to the weighted GPA, only dialogue and situation and cloze test sections had high, but statistically insignificant, correlations. This suggested that testing skills separately seems to be an appropriate idea in academic English proficiency tests of engineering students. However, this does not mean that synthesizing must be abolished altogether for these students. The correlation statistics of the second task of PTBST showed a good, although statistically insignificant, correlation with the total GPA of engineering students. As mentioned earlier, this task entails synthesizing two sources of input, but elicits only direct reporting. The note-taking section of the EPE also elicits a direct reporting through summarizing, and this section has a relatively good correlation with the total GPA of engineering students. Therefore, it is safe to assume that direct reporting can be the most important construct that must be captured when measuring the productive skills of academic English proficiency of the engineering students, and if synthesizing is involved, it must be limited to direct reporting.

Finally, dialogue and situation and cloze test sections exhibited a consistently good correlation with the GPAs of both engineering and non-engineering students. This is of no surprise as academic communicative ability entails reading and listening comprehension, but more importantly, using the comprehended message to produce meaningful utterances, which is a pronounced construct in the dialogue and situation section of EPE where test-takers must write utterances suitable for the context and co-text delineated by the item prompts (Bachman, 1990; Canale & Swain, 1980). As to the cloze test section, as Bachman (1982) puts it, depending on the pattern of blanking out the words in a cloze test, it can virtually correlate with any kind of test, which is also corroborated by Al-Musawi and Al-Ansari (1999) who found that the cloze test section of FCE was one of the two test sections that could predict the academic success of English major students measured by their GPA. Perhaps this is the reason for the consistently good correlation of this section of EPE with the GPAs of both engineering and non-engineering participants.

CHAPTER 7

DISCUSSION

7.1 Construct Validity

7.1.1 PTBST construct. Validity was traditionally viewed as a combination of content, criterion-related, and construct validities, yet in the recent views of validity, construct lies at the center of test validity and the other two types can be used to support the construct validity (Chapelle, 1999). Content validity refers to the examination of the test content to establish its representativeness of the criterion situation or target language use situation (TLU). Criterion validity, however, refers to the ability of the test to capture the abilities that are crucial for success in TLU. Yet, construct validity refers to "the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores" (Bachman & Palmer, 1996, pp. 19-21). In other words, test scores "are to be interpreted appropriately…with respect to a specific *domain of generalization,"* or set of tasks in a specific target language use domain (Bachman & Palmer, 1996, pp. 19-21).

According to Bachman (2004), test validity is a conceptual argument which is supported by validation procedures mostly comprised of statistical tests. So, to probe the validity of PTBST, first references to the operationalization of the construct will be made to form the conceptual rationale, and then, using the results of statistical analyses, the evidence to support the conceptual arguments will be provided.

As mentioned in the test operationalization section, PTBST attempts to measure the ability of test-takers in using language and synthesizing information through the target language in the context of English-medium academic environment. From the language use point of view, a communicative framework was implemented proposed by Bachman (1990). This approach is further corroborated by Morrow (as cited in Wall & Taylor, 2014) who argues that tests based on communicative competence are better than traditional ones since they show the quality of performance by test-takers rather than the number of correct answers produced by them. Another reason for adopting a communicative approach in designing PTBST is that, according to Harding (2014), this approach has been the major contributor to the test design and construct of tests so far.

Bachman (2007) identifies three aspects of construct which underlie test development: language ability, context, and the interaction between these two. In the model adopted for PTBST in this study, he proposes that language ability and context should be treated separately during both the test design and test result interpretation. He uses an extended model of language ability proposed by Canale and Swain (1980) to delineate the language ability, or trait, aspect of the test construct. As to the context, he suggests that the characteristics of test tasks be similar to those of the target language use (TLU) or criterion situation.

It might be argued that more recent communicative models of language competence were more apt for PTBST, yet it is not the case. Bachman (2007) discusses other models of language testing construct proposed after his model, like the interactional model. However, he argues that either the empirical evidence is not enough to support them or they are in contrast with the theories of language that the draw on. Yet, he does not reject these approaches altogether. According to Bachman (2007) the way one of these construct models, i.e. ability-in-individual-in-context proposed by Chalhoub-Deville and Deville (as cited in Bachman, 2007), defines the relationships between language ability and context is noteworthy. According to this model the language ability of a language user in a context interacts with the context facets and both change and are changed by those facets. Again, one problem with this proposition is that performance from this point of view is individual- and context-specific. As a result, the results of tests built upon such a construct is not generalizable — one of the crucial aspects of a test. Therefore, Bachman's (1990) model of communicative language testing is the most detailed among those concerning the conventional communicative language testing model (Harding, 2014). According to Harding (2014), not only is the delineation of the language ability a more comprehensive one in Bachman's model, but also it provides a clear distinction between task-based approach to testing and interactional approach. While the latter endeavors to replicate real-life interaction in testing tasks, the former argues that the tasks should be designed in a way that replicates characteristics of the real-life tasks, and this will create the interaction when test-takers put their language knowledge to use through those tasks. Therefore, taking into consideration these real-life, criterion-related characteristics (those that occur in real English-medium university life), the three tasks of PTBST were operationalized and designed. Situations inside and outside classroom were replicated by staying loyal, as much as possible, to major characteristics of these settings, e.g. formality level, register, tone, etc. to ensure adhering to construct validity of the test.

One criticism might concern the authenticity of the content of the test. In other words, it might be argued that the tasks do not have completely authentic language and are, at best, adaptations of authentic texts. However, using authentic source material for the purpose of developing the input for the tasks can affect the authenticity of the communication happening in the test, which is one of the important criteria mentioned by Morrow (as cited in Wall & Taylor, 2014). Besides, Wall and Taylor (2014) believe that capturing full authenticity is impossible in testing since test-takers will always know that they are not participating in an authentic communication. Yet, Bachman (1990) defines authenticity in testing as the similarity of the characteristics of the test task to that in the criterion settings — in this case academic language use situations inside and outside classrooms — which is observed and implemented in PTBST.

Different aspects of Bachman's (1990) model was operationalized in three different tasks in PTBST. This test was a task-based one which is an apt form of testing communicative competence or language use abilities since it allows for adding the element of the unpredictability to the communication happing in the test (Wall & Taylor, 2014). Similary, Jacoby and McNamara (1999) assert that a test of performance is by definition task-based. Furthermore, Morrow (as cited in Wall & Taylor, 2014) considers unpredictability a crucial attribute of tests of communicative competence. In other words, test developers can choose from a wide variety of topics and situations and operationalize them into tasks eliciting performance from test-takers to see what they can do with the knowledge of language. This is also relevant to the strategic competence as a salient part of Bachman's (1990) model, which was discussed under test construct operationalization section.

7.1.2 PTBST rubrics. Douglas (as cited in Brunfaut, 2014) argues that the rubrics, or assessment criteria, must be derived from the criterion situation language analysis. Although a very accurate approach, it needs a long-term and intensive research plan to pinpoint the different levels of each criteria necessary for success in each university subject area. Yet, there is another approach for developing rubrics. Brunafaut (2014) reports that rubrics which treat language and task accomplishment criteria separately are common in the field of language testing. Moreover, this author argues what matters is the implementation of the rubrics in line with the test purposes. Therefore, it is justified to see the linguistic knowledge and the ability of the test-taker to synthesize the information in an academic context as two important criteria reflected in their separate descriptions in the rubrics. Moreover, as Bachman (2007) suggests, rubrics must draw upon the same criteria used in the test construct and must be in line with the TLU. Accordingly, the rubrics for PTBST were developed based on the test criteria and communicative language competence model adopted for PTBST, which was at the same time an attempt to align the test and rubrics with the communicative criteria of TLU, namely department classes at METU NCC.

Aside from the choice of approach to the design of PTBST rubrics, there was one particular section of rubrics allocated to the questions asked by the test-takers while listening to the input lectures. During listening to the lectures, the test-takers were given three chances to ask a question to the lecturer and record it after the end of each main idea, hence, having three chances of asking question in each lecture. Asking a question was not mandatory, and the fact that some test-takers chose to use this feature of the test, while other did not, could provide a data for further analyses, like examining the nature of the role of this factor in predicting the academic success or its correlation with higher (or lower) scores in PTBST, to name a few. However, as the name of PTBST suggests, this is a task-based test, and the major focus in both the test tasks and rubrics was on task accomplishment. Bachman (1991) considers test authenticity an important factor for performance tests which are used to measure the degree of success a test-taker could have in a criterion situation. He divides authenticity into situational and interactional categories. While the former refers to the degree to which a task emulates the distinctive characteristics of the TLU, e.g. register, formality of language, or monologue vs. dialogue, the latter pertains to the task accomplishment and the degree to which test tasks elicit the relevant language ability for a successful task accomplishment. Therefore, treating questions asked would be detrimental to the situational authenticity of PTBST, as it is part of a plethora of factors contributing to the situational authenticity of the test. In other words, the *questions asked* are part of the successful task accomplishment which is viewed as the collective effect of all different aspects of the performance, e.g. linguistic competence, sociolinguistic competence, etc. as well as the *questions asked*. As a result, a single rating for all tasks, to which the scores of questions asked and all the other criteria contributed, was taken as the independent variable in the research design. If the present study was to probe into the different traits of test-takers, e.g. their willingness to engage in asking questions during lectures, conducting such an analysis would be more apt. Yet, this was not the focus of the research design, and the scores of questions asked were taken as a contributing factor, along with all the other factors in the rubrics, to measure the level and quality of task accomplishment by the test-takers.

7.2 Statistical Evidence of Validity

According to Harding (2014), language tests with communicative competence at the heart of their constructs are criterion-referenced with authentic tasks and are validated based on the abilities from the TLU. In other words, these tests are criterion-referenced, and construct and predictive validity studies should be used to assess these tests rather than other types of validation. In the previous section, the construct validity arguments were provided, and the issues related to content validity were addressed. Now, the criterion validity will be addressed, which is the same as the second step in test validation proposed by Bachman (2004), i.e. conducting statistical tests.

One of the statistical tests conducted in this study was to establish the reliability of the test. According to Chapelle (1999) reliability is not a separate and perquisite condition of test validation, but rather as one form of validation. One test of reliability conducted in this study was a Many-Facet Rasch Measurement. The results showed an acceptable rate of consistency between the judgments of the raters along with a very low degree of bias. This evidence provides one form of validity support for PTBST.

The other statistical tests in this study provided evidence for criterion-related validity of PTBST. The criterion was the students' language use ability in an English-medium university defined by their first semester GPAs. The correlation and factor analyses all showed results which supported the construct validity of PTBST. According to the definition of the test construct, PTBST was designed to capture the ability of students in using English successfully in their academic studies with respect to their subject areas. Careful analyses of the different tasks of PTBST and their differing correlation statistics with the first-semester GPAs of engineering and non-engineering students provided the empirical evidence that PTBST tasks can be successful in capturing the underlying abilities needed for success in both engineering and non-engineering faculties. In other words PTBST can be considered a valid test, since it can successfully measure what it purports to do (Bachman & Palmer, 1996).

Moreover, the exploratory factor analysis showed that PTBST is not a redundant form of test and can effectively complement the current METU EPE test, as it measures a different construct from what the sub-sections of EPE do.

7.3 New Findings and Future Research

Perhaps the most important finding of this study is at least two distinct forms of synthesizing input. The second and third task of PTBST both have synthesizing at their heart, but these two tasks correlate differently with engineering and non-engineering student GPAs. The results of factor analysis also imply such a difference. The difference is the added dimension of direct reporting (as captured in the second of PTBST) or integrating personal ideas (task three). The results show that while the former correlates better with engineering student GPAs (the correlation is statistically insignificant though), the latter does so with non-engineering student GPAs. This was further corroborated by the better correlation of note-taking section of EPE with engineering GPAs (again with no statistical significance), as it involves only direct reporting and summarizing. Therefore, it can be argued that the second and third tasks of PTBST measure different kinds of synthesizing, and each of these can be more relevant to one of the engineering or non-engineering disciplines and the academic success in these disciplines.

Another important finding of this study is that testing productive skills in an academic proficiency test may have to entail some sort of synthesis. This is due to the fact that the writing section of EPE, which is a productive skill test with no synthesizing involved, had a negative correlation with non-engineering GPAs and a very low and statistically insignificant correlation with engineering GPAs. This finding merits further study since its confirmation would have considerable implications for users of the commercial test scores like TOEFL iBT in which, for instance, one of the writing tasks does not entail any synthesis while the other has it at its heart, especially considering the fact that the scores for these writing tasks are reported separately. In other words, stake-holders and score users can make better screening or selection decisions by taking into consideration the scores that are more predictive of academic success based on the subject area.

This brings up the final important finding of this study, which means that one size does not fit all. In terms of commercial tests like IELTS or TOEFL, the only change needed is informing the stake-holders of the implications of the scores of different writing or speaking tasks. The same can be adopted for in-house academic proficiency tests such as EPE. However, a better decision would be having test-takers attempt only items or sections of the test that have the most predictive power of academic success regarding their subject area. Of course, this would need adding or modifying sections to yield the maximum possible information about test-takers' abilities. Still, having test-takers attempt only certain parts of a test might be a threat to its face-validity, and valuable data which can be obtained from the sections of the test which bear low relevance to certain subject areas can still be useful. For instance, the test scores and data from engineering students on the sections with low relevance to their disciplines can be used as an external variable and can be compared to the same data from non-engineering test-takers on the same sections (which have high relevance to non-engineering disciplines) as a quality check to ensure construct relevance and maximum predictive power of the test.

Another solution to this problem is using different cut-off points for different disciplines. For instance, Wait and Gressel (2009) believe that a different cut-off point must be applied in the application procedure of different subject areas, so that, for instance, the applicants of engineering subjects can gain access to college education with lower proficiency test scores compared to their non-engineering counterparts since for these courses a lower level of language competence is needed. However, Wait and Gressel (2009) also believe that using TOEFL score as part of admission to all types of programs may not be an appropriate practice as the underlying language constructs that TOEFL covers might be more relevant to some subject areas than others. This latter idea of the mentioned authors seem to contradict their suggestion regarding using lower cut-off points. To illustrate, setting a lower cut-off point as an admission criterion for engineering students would mean that the applicants would get a lower score in sections of the test whose construct has a low relevance to their subject areas, as well as in those whose construct are highly relevant to the subject areas in question. Therefore, setting a lower cut-off point would still have the risk of admitting students who have not shown an acceptable level of competence in language abilities which are of importance to their success in their engineering fields of study.

Probably, the best solution to this problem would be the use of EPE scores which combine the scores of sections with different weightings. The results of this study showed that some parts of EPE, e.g. reading, and potentially the second task of PTBST, have a good correlation with the GPAs of engineering students. Therefore, a more expedient approach might be applying higher weightings on the scores of the sections of proficiency tests whose underlying construct have more relevance to test-takers' areas of study. Therefore, while continuing to administer the same test, and keeping the costs low, a better implementation of EPE scores can be achieved. Moreover, the data from sections which have lower relevance to the subject areas in question can be used for quality control of the operationalization of the test constructs as well as their further development as well as future research.

CHAPTER 8

LIMITATIONS OF THE STUDY

Perhaps the most salient limitation of this study is the low participation rate. This is a common phenomenon in human sciences (Larson-Hall, 2010), which renders a lot of statistical tests improper or impossible to conduct and is a big disadvantage for quantitative studies. Furthermore, due to the small number of cases, missing data and outliers can distort the realistic results considerably (Larson-Hall, 2010). As a case in point, there was one case of weighted GPA with a value of 0 in both engineering and non-engineering groups. While taking out this case as an outlier did not make much change to the results for the non-engineering group, the same affected the correlation results for the engineering groups significantly by increasing the correlation of the second task of PTBST to a statistically significant value of .617 and .612 for total and weighted GPAs respectively. However, these cases were kept in the data to report only the least desirable results to avoid bias. In order to get a clearer picture and better insight into the nature of PTBST and its predictive power, future studies must be carried out with a higher number of participants who have a more diverse proficiency levels and represent different subjects areas in higher numbers.

The need for a higher number of participants was also corroborated by the results of power analyses for the correlation statistics. Although the correlation between PTBST and GPAs of all participants (engineering and non-engineering together; n=31) showed a very good power (power = 0.7), the power analysis of the correlation between PTBST and both total and weighted GPAs of all participants (engineering and non-engineering together) showed that 40 and 57 participants are needed, respectively, to achieve the optimum power of 0.8. Furthermore, other power analyses of correlations between PTBST and total GPAs of non-engineering showed a relatively good power (power = 0.66). However, to obtain a significant correlation between the PTBST and its third task with the weighted of non-engineering students with a power of 0.8, 40 participants will be ideal. Finally, as to engineering students, the power analyses for the correlations of EPE and its reading section with both total (0.66 and 0.64, respectively) and weighted GPA (0.67 and 0.58, respectively) showed relatively high values, but at least 45 participants are needed to obtain statistically significant correlation values for the second task of PTBST with the total GPA of engineering students. Therefore, a group of at least 80 participants, with 40 engineering and 40 non-engineering, will be an ideal number to obtain statistically powerful results and clearer picture of the results found in this study.

Another limitation of this study concerns methodology. Originally, it was planned to administer the PTBST to all participants at one session, especially since the necessary infrastructure for a simultaneous administration of the test was available. There were enough number of computer stations available at the university. Also, headsets with noise cancellation capability were available and could have been employed to address the problem of test-taker distraction or possible cheating by listening to other test-takers and answering the questions in PTBST as the questions and the tasks were the same for all the participants. However, the busy schedule of participants made this impossible. Hence, there was time gaps between each administration of the test, which is likely to affect the reliability of the test (Bachman, 2004). Moreover, practicality of administering the test to a large population at one session was not possible to probe. Brunfaut (2014) argues that practicality "should be central to language test development from the start of the process and at the same time be integrated in theory." In addition, one major reason for designing a computerized test was to address the practicality issues of speaking test administration, some aspects of which could not be evaluated in practice due to the mentioned reasons.

Another limitation of the study concerns the timing of collecting PTBST data, which was done after the university classes had started. This was due to the fact that potential participants were not on campus before the start of classes, and even if the researcher had contacted all the participants and obtained their consent, he could not have been able to administered PTBST before the start of university classes. Therefore, the input the participants had received in their university classes could arguably have biased the results of the study in favor of a higher correlation with GPA. Although there was no practical solution to address and prevent this methodological problem, the researcher made an attempt to explore the nature of this bias.

Frist of all, before conducting any test to examine the possible bias arising from the late administration of PTBST, one matter must be clarified, and that is this study does not intend to suggest that PTBST is better than EPE, or worse, regarding their predictive power. There is both statistical (as the factor analysis shows) and theoretical support that EPE and PTBST measure different constructs. The operationalization of PTBST based on a sound theoretical model of language assessment shows that this construct is composed of synthesizing the input and speaking. Based on the same theory, synthesizing is not covered in the construct of EPE (as it is in the third task of PTBST for instance), and EPE obviously lacks a speaking section. Since EPE and PTBST test different constructs, the possible positive bias in favor of PTBST resulting from the time lapse between the administration of the test would not suggest that PTBST is stronger (or weaker) than EPE in predicting academic success, and, taking into consideration the difference in construct, this comparison might not be entirely appropriate.

Now, to scrutinize the nature of the possible bias the timing of collecting PTBST data could have brought about, the participants were arranged in a chronological order and were divided into two groups of early and late participants. Then, three correlation of PTBST scores of these two groups with both total and weighted GPAs were conducted. The first one included all engineering and non-engineering participants. The second included only engineering students, and the last covered only non-engineering participants. The results showed that the PTBST scores of the early group had a higher correlation with both total and weighted GPA for all three series of correlations (see Appendix G). This might probably mean that the input from classes did not affect the results in favor of PTBST if the assumption is that the participants could improve the skill or knowledge being tested in their university classes and if the learning had an incremental nature. Yet, this could also mean that the higher correlation between PTBST scores and GPAs of early participants could be accidental, and that the participants whose data could contribute to a higher correlation between PTBST and GPAs simply happened to participate in the study earlier. Moreover, with the low number of participants, even if these results are not accidental, they are not generalizable. Therefore, the best practice to explore the true effect of administration timing of PTBST would be replicating this study with the administration of PTBST before and after the start of the semester, as well as further into the semester, with a high number of participants. If the same results are replicated as this study, a lower correlation of PTBST with GPA further into the semester would be an interesting finding worthy of further research.

One more limitation of this study is arguably the involvement of the researcher at all steps of the study: from the design of PTBST to training the raters. This involvement arose from practicality issues as it was difficult, if not impossible, to solicit the help of a research assistant. However, care was taken to minimize the bias effect of such involvement. As mentioned in the *Designing PTBST* section, a powerful communicative language testing theory proposed by Bachman (1990, p. 87) was used as the basis of test operationalization. The exhaustive explanation in the mentioned section shows the level of care taken by the researcher to ensure a theoretically supported design and operationalization of the test with the minimum of researcher bias introduced into the design. The same approach was taken regarding the design of the rubrics. As to the possible bias of researcher involvement in rater training, even if the researcher had had a research assistant conduct the training, the ideas and training would have been passed down through the researcher to the research assistant (i.e. our hypothetical rater trainer) contributing to little difference from the present design. In addition, at the time of rater training, the researcher had not received the semester grade reports and GPAs and, hence, could not have had any basis to place his bias on during the rater training sessions to obtain desirable results (i.e. getting a higher predictive power of PTBST compared to that of EPE). As to possible bias resulting from the involvement of the researcher in presenting input in the lecture videos of PTBST, there was a minimal contact between the researcher and the participants, and the level of contact was the same for all the participants. This can offset the possible bias effect of the researcher acting in the videos and delivering the input in tasks one and two of PTBST. Besides, the results of MFRM bias analysis showed a minimum rate of bias, which further confirms that the researcher involvement at all steps of the study could not have biased the results considerably, if at all. Hence, although it is hard to argue for a complete lack of any possible bias introduced by the involvement of the researcher at every step of the study, careful measures were taken to bring such an effect to a minimum as the alternative, i.e. the use of a research assistants, was not feasible.

Perhaps, a final limitation is that the researcher could not compare the results of this test to those of an internationally recognized ones, e.g. IELTS or TOEFL iBT, to see how PTBST compares to them in terms of its predictive power as this type of comparison is no uncommon in the literature. Jamieson, Wang, and Church (2013), for instance, found out that an in-house speaking test tended to cover more of the constructs and abilities of interest than a commercial one, namely, Versant designed by Pearson. However, they admit that their in-house test tended to take a considerable amount of staff time and was less practical than the automated commercial test they used. Having found the results of this analogy between their test and the commercial one, Jamieson, Want, and Church (2013) finally report that they decided to keep administering the in-house test due to its low cost. Compared to the test these authors report, PTBST is an automatically delivered test, and since it is an in-house test, it can cover more of the abilities and constructs in question compared to commercial tests as it does not have to be a one size fitting all like IELTS or TOEFL which are design to cater for a wide spectrum of test-takers. Moreover, the low cost and less demand on staff and resources are the possible advantages of PTBST. Nevertheless, a comparison study between PTBST and IELTS or TOEFL iBT speaking is still necessary to give a clear picture of how PTBST's predictive power compare to these commercial tests. This was not possible due to the lack of such data, but future studies are well-worth to conduct such comparisons.

CHAPTER 9

CONCLUSION

Language assessment plays an important role in educational settings, and tests of proficiency are of crucial significance for English-medium universities. Of the most famous of the latter type are TOEFL iBT and IELTS, and many English-medium universities use the scores of these two tests to make admission decisions. Others prefer to use an in-house version of proficiency tests. Considering the characteristics of test-takers and test context while developing a test is emphasized in literature, and Brunfaut (2014) illustrates that these characteristics refer to those that are typical of target language use situation. Perhaps this is one of the most important reasons why some English-medium universities in Turkey and Northern Cyprus have decided to develop and administer their own test. Obviously, such practice is a continuous one and entails rigorous studies and analyses to ensure the validity and reliability of these tests (Alderson, 2009). The current study was an attempt to study the effect of adding a computerized task-based speaking component to the current METU EPE using the development and validation criteria in the literature. Of course, to achieve effective results, such studies must be conducted on a continuous basis in an attempt to achieve optimum results.

On the other hand, the domain of language testing is a dynamic one, and recent years have witnessed new proposals and approaches to testing which are at the development phase and need professional and empirical studies to ensure their appropriateness in practice (Bachman, 2007; Wall & Taylor, 2014). This further adds to the importance of continuous studies into test development and ensuring the application of new approaches to the practice of testing.

The impact of the test on stake-holders is another important aspect of testing which deserves due attention. According to Chapelle (1999), one type of rationale which can be used as a construct validity argument is test consequence or washback effect of the test. Clearly, speaking is an important skill in any language, and failure to capture this skill in testing has serious implications. For one thing, such a practice suggests that speaking skill bears no importance and does not warrant attention. More importantly, a failure to capture the type of speaking which is characteristic of TLU might imply the fact that there is no difference in the type, tone, style, and other characteristics of speaking from context to context and from one subject area to another. Such a negative washback effect can only be counteracted by implementing tests which measure the type of speaking happening in TLU — in this case, spoken communication taking place inside and outside classrooms of an English-medium university campus. PTBST can be considered a first step in introducing such a washback effect to METU NCC and providing a solid justification for time, space, attention, and resource allocation to training this skill in METU NCC SFL. Similarly, this can ensure more attention given by SFL students to this crucial skill during their English preparatory program. Moreover, since synthesizing information is a central part of PTBST's construct, the washback effect of this test, if implemented, can bring about more attention to this crucial academic skill. The results of this study suggest how important this particular skill is in ensuring academic success at METU NCC, especially for non-engineering students.

Another point worthy of attention is the meaningfulness of scores. Reporting a single numerical value as the measure of a test-taker's language competence is not enough information to make important decisions on matters like university admission and course placement. One crucial aspect of test development which concerns the stake-holders and decision makers is setting-standards and giving meaning to scores. Tannenbaum and Cho (2014) propose a framework for setting standards and converting test scores into meaningful descriptors which enable the policymakers and stakeholders to make more informed decisions using test scores. At METU NCC, EPE scores are reported as a single numerical summary with no descriptors attached to it. However, EPE is administered to prospective students of all subject areas, and since designing and administering a test appropriate for each subject area is both costly and challenging, standard descriptors of EPE grades, along with the indication of which section is more important for which subject areas, can help decision makers in each department make a more informed decision on whether students are ready to start their studies at their respective faculties. The framework proposed by Tannenbaum and Cho (2014) is specifically useful as it involves the decision makers in this process and entails documentation procedures which allow for consistency of the test score descriptors and their availability for future use and modifications. Alternatively, as an economical approach, score reports with different weightings of sub-sections can be implemented. Therefore, giving higher weights to test sections which bear more relevance to a certain discipline, test developers and users can still continue to use a single numerical value, but one which is more likely to reflect the ability level necessary for success in that discipline.

More importantly, Douglas (as cited in Brunfaut, 2014) argues that the rubrics, or assessment criteria, must be derived from the criterion situation language analysis. Therefore a more robust approach to test construct of EPE and PTBST would be conducting a more detailed job analysis of TLU to ensure the representation of the major language skills required for different subject areas. For instance, this study showed that there are two distinct forms of synthesizing input, and each were more appropriate than the other for a test of academic English proficiency depending on the subject area of the test-takers (i.e. engineering or non-engineering). Further analysis of TLU can find more distinct forms of synthesizing which are appropriate for different subject areas, and this can inform the test construct better resulting in academic English proficiency tests with better predictive power. This can be achieved through analysis of the content and linguistic levels of tests and assignments given by professors at different faculties. This latter action might be too difficult to take, but is sure to yield interesting and effective results. Alternatively, a reversed approach can be taken in informing decision makers. In other words, the test rubrics can be used to report meaningful scores describing what a test-taker can and cannot do based on the test results, and by evaluating those abilities, decision makers in each department and faculty can make informed decisions as to who to admit or vice versa.

Closely related to the meaningfulness of test scores is their application to English-medium university settings. At an English-medium university, both the content and assessment are presented through English. Therefore, students' success depends on their language proficiency as well as content mastery. In other words, both learning and assessment rely on English and having language competence below the required level can affect both learning and test performance of students in their subject areas (Shaw & Imam, 2013). Therefore, the minimum language competence required to successfully learn the content of the subject area and to perform well on subject area tests, against which the success of students are measured, must be identified. Shaw and Imam (2013) conducted a study to find out what the minimum required language level is necessary to successfully pass International General Certificate of Secondary Education exams. They also looked into the different language competences and skills required by each subject area. They found that a general B1 level of proficiency in Common European Framework Reference for Languages (CEFR) was adequate in this particular context. However, they found that a higher language competence level is required for humanity subject areas like history than that for natural sciences such as biology. Shaw and Imam’s (2013) study has two implications. First of all, the same cut-off score for all the faculties may not yield optimum results. Clearly, some subject areas need a higher competence level. In the case of EPE, only EFL students are required to obtain a higher score (70 and above), and a score of 60 and above is required of all the other subject areas, while some of these fall into human sciences category and need to meet a higher cut-off score. Second, some parts of academic English proficiency tests must be given a higher weight than the other parts depending on the subject area of the test-taker, and this was evident in this study, although the low number of participants inhibits a definite reliance on the results of this study to make such a claim. However, other studies also show that particular test types and sections seem to have more predictive power than others in terms of academic success, and the predictive power of these sections are not the same across all subject areas (Al-Musawi & Al-Ansari, 1999; Ayers & Quanttlebaum, 1992, as cited in Cho & Bridgeman, 2012; Vinke & Jochemes, 1993; Wait & Gressel, 2009). Therefore, after all, weighted proficiency scores, with higher weighting of the sections that are more relevant to certain subject areas, seem to be an appropriate approach.

PTBST is not free from drawbacks. The results show that this test still does not have a strong predictive power regarding the academic success of engineering students. While, this might be due to the low number of participants, it can simply be due to the fact that, as Graham (1987) and Oliver, Vanderford, and Grote (2012) suggest, there are many factors contributing to the success or failure of a student at an English-medium university than English proficiency, and PTBST did not have the potential to capture those factors. However, taking an empirical approach and following the literature in developing it proved to be effective and yielded promising results, at least for non-engineering students. Moreover, Bachman (1991) believes that both task content and task method of a test task are major players in deciding the resultant performance of test-takers. Therefore, with a more careful design of tasks which reflect the common task types of various engineering disciplines and with integration of content shared by the majority of engineering subject areas, PTBST can also demonstrate a sensibly strong predictive power of academic success for engineering disciplines, as well.

Finally, further studies with larger sample sizes need to be done to obtain a clearer and generalizable results. Moreover, job analyses which gather information on the tasks and abilities required by different subject areas should be conducted to further fine tune the operationalization of the test in the future.

REFERENCES

Alderson, J. C. (2009). Test review: Test of English as a Foreign Language ™: Internet-based Test (TOEFL iBT®). *Language Testing*, *26*(4), 621-631. doi:10.1177/0265532209346371

Al-Musawi, N., & Al-Ansari, S. (1999). Test of English as a foreign Language and first certificate of English tests as predictors of academic success for undergraduate students at the University of Bahrain. *System*, *27*(3), 389-399. doi:10.1016/s0346-251x(99)00033-0

Aydın, G. (2012). *The role of English proficiency level, personal and affective factors predicting language preparatory school students' academic success* (Master's thesis, Middle East Technical University, Ankara, Turkey). Retrieved from http://etd.lib.metu.edu.tr/upload/12614711/index.pdf

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*(1), 61. doi:10.2307/3586563

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition*, *10*(02), 149-164. doi:10.1017/s0272263100007282

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*(4), 671. doi:10.2307/3587082

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (2007). what is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. D. Fox, M. Wesche, D. Bayliss, L. Cheng, & C. Coe (Eds.), *Language testing reconsidered* (pp. 41-71). Ottawa, Ontario: University of Ottawa Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257. doi:10.1177/026553229501200206

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74. doi:10.1080/15434300903464418

Brooks, L., & Swain, M. (2014). Contextualizing performances: comparing performances during TOEFL iBT TM and real-life academic speaking activities. *Language Assessment Quarterly*, *11*(4), 353-373. doi:10.1080/15434303.2014.947532

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1-25. doi:10.1191/0265532203lt242oa

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (RR-05-05, TOEFL-MS-29). Princeton, NJ: Educational Testing Service.

Brown, G., & Yule, G. (1983). *Teaching the spoken language: An approach based on the analysis of conversational English*. Cambridge, UK: Cambridge University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York, NY: Longman.

Brunfaut, T. (2014). Language for Specific Purposes: Current and Future Issues. *Language Assessment Quarterly, 11*(2), 216-225. doi:10.1080/15434303.2014.902060

Burgess, T. C., & Greis, N. A. (1970). *English language proficiency and academic achievement among students of English as a second language at the college level* (ED074812). Retrieved from ERIC database.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (RM-00-06, TOEFL-MS-20). Princeton, NJ: Educational Testing Service.

Bygate, M. (1987). *Speaking*. Oxford, UK: Oxford University Press.

Byrnes, H. (2002). The role of task and task-based assessment in a content-oriented collegiate foreign language curriculum. *Language Testing*, *19*(4), 419-437. doi:10.1191/0265532202lt238oa

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *I*(1), 1-47. doi:10.1093/applin/i.1.1

Champely, S. (2015). *Pwr: Basic functions for power analysis v 1.1-3* [Computer software]. Retrieved from http://cran.r-project.org/package=pwr

Chapelle, C. A. (1999). Validity in language assessment. *Annual review of applied linguistics*, *19*, 254-272. doi:10.1017/s0267190599190135

Chapelle, C., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000* (RM-97-03, TOEFL-MS-10). Princeton, NJ: Educational Testing Service.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT(R) scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421-442. doi:10.1177/0265532211430368

Cohn, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (RR-06-06, TOEFL-MS-33). Princeton, NJ: Educational Testing Service.

Coulthard, M. (1985). *An introduction to discourse analysis* (2nd ed.). New York, NY: Routledge.

Crossley, S., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*(3), 250-270. doi:10.1080/15434303.2014.926905

Cubilo, J., & Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. *Language Assessment Quarterly*, *10*(4), 371-397. doi:10.1080/15434303.2013.824972

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, U.K: Cambridge University Press.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research*, *4*(3), 193-220. doi:10.1177/136216880000400302

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: L. Erlbaum Associates.

Enginarlar, H. (2006). *METU-EPE (2006) validity studies*. Middle East Technical University. Retrieved from http://ydyom.metu.edu.tr/en/metu-epe-reports.

Enginarlar, H. (2007). *A research report on METU-EPE (2007/June) (English proficiency exam)*. Middles East Technical University. Retrieved from http://ydyom.metu.edu.tr/en/metu-epe-reports.

Enginarlar, H. (2009). *Validity studies on METU-EPE (English proficiency exam)*. Middle East Technical University. Retrieved from http://ydyom.metu.edu.tr/en/metu-epe-reports.

Enginarlar, H. (2012). *A research report on January 2012 EPE*. Middle East Technical University. Retrieved from http://ydyom.metu.edu.tr/en/metu-epe-reports.

Ferris, D., & Tagg, T. (1996). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly, 30*(2), 297. doi:10.2307/3588145

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2-32. doi:10.1177/026553229901600102

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133-167. doi:10.1191/0265532202lt225oa

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly*, *21*(3), 505. doi:10.2307/3586500

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, *11*(2), 186-197. doi:10.1080/15434303.2014.895829

Heil, D. K., & Aleamoni, L. M. (1974). *Assessment of the proficiency in the use and understanding of English by foreign students as measured by the test of English as a foreign language* (ED093948). Retrieved from ERIC database.

Hughes, A. (1988). Introducing a needs-based test of English language proficiency into an English-medium university in Turkey. In A. Hughes (Ed.), *Testing English for university study* (pp. 134-153). London, UK: Modern English Publications in association with the British Council.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). London, UK: Penguin.

Hymes, D. (1972). Models of interaction of language and social life. In J. J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 35-71). New York, NY: Holt, Rinehart and Winston.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, *18*(3), 213-241. doi:10.1016/s0889-4906(97)00053-7

Jamieson, J., Wang, L., & Church, J. (2013). In-house or commercial speaking tests: Evaluating strengths for EAP placement. *Journal of English for Academic Purposes*, *12*(4), 288-298. doi:10.1016/j.jeap.2013.09.003

Jones, W. (2012). Assessing students' grammatical ability. In C. A. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to* *second language assessment* (pp. 247-256). Cambridge, UK: Cambridge University Press.

Kermond, C. (2012). *Psychology 101: Introductory psychology syllabus and class information 1st summer session 2012* [Class handout]. Retrieved from Department of Psychology, Michigan State University, MI, USA

Kirkgoz, Y. (2007). English language teaching in Turkey: Policy changes and their implementations. *RELC Journal*, *38*(2), 216-228. doi:10.1177/0033688207079696

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System*, *20*(3), 373-386. doi:10.1016/0346-251x(92)90047-7

Lebeau, I., & Rees, G. (2008). *Language leader coursebook series*. Harlow, Essex, UK: Pearson Education Limited.

Lim, P. L., Kurtin, M., & Wellman, L. (2001). *Grammar workbook for the TOEFL exam*. New York, NY: Arco.

Linacre, J. M., & Wright, B. D. (1999). *A user's guide to FACETS: Rasch measurement computer program*. Chicago, IL: Mesa Press.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

McKay, S. L. (2003). Toward an appropriate EIL pedagogy: re-examining common ELT assumptions. *International Journal of Applied Linguistics*, *13*(1), 1-22. doi:10.1111/1473-4192.00035

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

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446-466. doi:10.1093/applin/18.4.446

METU Cyprus. (2016, June 8). Admission Criteria for Your Country- METU NCC. Retrieved from http://international.ncc.metu.edu.tr/admission-criteria-for-your-country

METU Ranking. (2016, June 8). METU in World Universities Ranking METU Cyprus. Retrieved from http://international.ncc.metu.edu.tr/metu-in-world-universities-ranking

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(3). doi:10.1515/iral.2007.011

Morgan, G. B., Zhu, M., Johnson, R. L., & Hodge, K. J. (2014). Interrater reliability estimators commonly used in scoring language assessments: A Monte Carlo investigation of estimator accuracy. *Language Assessment Quarterly*, *11*(3), 304-324. doi:10.1080/15434303.2014.937486

North, B. (2012). *The Development of a Common Framework Scale of Language Proficiency*. New York, NY: Lang, Peter, Publishing Inc.

ODTÜ Kuzey Kıbrıs Kampusu. (2016, June 8). Taban Puan ve Sıralamalar. Retrieved from http://tanitim.ncc.metu.edu.tr/taban-puan-ve-siralamalar/

Oliver, R., Vanderford, S., & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research & Development*, *31*(4), 541-555. doi:10.1080/07294360.2011.653958

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, U.K: Cambridge University Press.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370-395. doi:10.1111/j.2044-8325.2011.02045.x

RStudio Team. (2015). *RStudio: Integrated development for R* [Computer software]. Retrieved from RStudio Inc., Boston, MA. http//www.rstudio.com

Selvi, A. F. (2011). The non-native speaker teacher. *ELT Journal*, *65*(2), 187-189. doi:10.1093/elt/ccq092

Shaw, S., & Imam, H. (2013). Assessment of International Students Through the Medium of English: Ensuring Validity and Fairness in Content-Based Examinations. *Language Assessment Quarterly*, *10*(4), 452-475. doi:10.1080/15434303.2013.866117

Shehadeh, A. (2012). Task-based language assessment: Components, development, and implementation. In C. A. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 156-163). Cambridge, UK: Cambridge University Press.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*(3), 233-249. doi:10.1080/15434303.2013.869815

Taylor, C., & Angelis, P. (2008). The evolution of the TOEFL. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 27-54). New York, NY: Routledge.

Van den Branden, K., Depauw, V., & Gysen, S. (2002). A computerized task-based test of second language Dutch for vocational training purposes. *Language Testing*, *19*(4), 438-452. doi:10.1191/0265532202lt239oa

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489. doi:10.2307/3586922

Vinke, A. A., & Jochems, W. M. (1993). English proficiency and academic success in international postgraduate education. *High Educ*, *26*(3), 275-285. doi:10.1007/bf01383487

Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, *10*(2), 178-195. doi:10.1080/15434303.2013.769552

Wait, I. W., & Gressel, J. W. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education*, *98*(4), 389-398. doi:10.1002/j.2168-9830.2009.tb01035.x

Wall, D., & Horák, T. (2011). The impact of changes in the TOEFL examination on teaching and learning in a sample of countries in Europe: Phase 3, the role of the coursebook. Phase 4, describing change. *ETS Research Report Series, 2011*(2), i-181. doi:10.1002/j.2333-8504.2011.tb02277.x

Wall, D., & Taylor, C. (2014). Communicative language testing (CLT): reflections on the “issues revisited” from the perspective of an examinations board. *Language Assessment Quarterly, 11*(2), 170-185. doi:10.1080/15434303.2014.902058

Weir, C. J. (1990). *Communicative language testing*. New York, NY: Prentice Hall.

Wiebe, J. (2013). *9 ways to make your expensive product look like a total steal* [Blog post]. Retrieved from http://blog.kissmetrics.com/a-total-steal/

Yapar, T. (2003). *A study of the predictive validity of the Başkent university English proficiency exam through the estimates use of the two-parameter IRT model's ability* (Master's thesis, Middle East Technical University, Ankara, Turkey). Retrieved from http://etd.lib.metu.edu.tr/upload/1217629/index.pdf

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79-94.

Yule, G. (1996). *Oxford introductions to language study: Pragmatics*. Oxford, UK: Oxford University Press.

Zhang, Y., & Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31-50. doi:10.1177/0265532209360671

APPENDICES

A. Rubrics

Table 17 – Rubrics for Linguistic Competence

0	1	2	3	4
The speaker does not speak, or barely a few sentences are understandable.	There are many grammar, vocabulary, and pronunciation mistakes. I can understand only around 30% of the response.	The grammar, vocabulary, and pronunciation mistakes the speaker makes renders around 50% of the speech difficult or impossible to understand.	The speaker makes a few grammatical, vocabulary, or pronunciation mistakes which make understanding around 20% of sentences difficult to understand.	The grammar, vocabulary, and pronunciation are near perfect. I have no difficulty understanding the speaker.
Delivery				
0	1	2	3	4
There is no fluency. The response is comprised of disconnected phrases or, at best, clauses.	There are pauses more than 60% of the time and no fillers are used to compensate for discontinuity of speech. Difficulty in understanding the speech because of disconnected flow of speech covers almost 60% of the response.	There are pauses around 40% of the time and no fillers are used to compensate for discontinuity of speech. Difficulty in understanding the speech because of disconnected flow of speech covers almost 40% of the response.	There are some pauses with no attempt to smooth out with fillers. Difficulty in understanding the speech because of disconnected flow of speech happens a few times.	There is a natural and sustained flow of speech. At the same time, everything is comprehensible. If there are pauses, appropriate fillers are used to smooth out the speech.

Table 18 – Rubrics for Discourse Competence and Task Accomplishment

	Discourse Competence				
	0	1	2	3	4
	There is no organization to the response	There are no pre-stories (Q1) and no meta-statements present. The organization can be vaguely understood by meaning. It means that intonation and linkers are not present to support the organization.	There are no pre-stories (Q1) and no meta-statements present. The organization can be vaguely understood by meaning, linkers or intonation.	There are pre-story (Q1) and meta-statement for general organization, but it is vague or incorrect. Meaning, intonation, and linkers make following the discourse easy though.	Pre-story (Q1), meta-statements for general organization, meta-statements/linkers/intonation patterns for subtopics are present that show a clear organization of the response.
	Task Accomplishment				
	0	1	2	3	4
Q 1 & 2	None of the main idea are presented, and the speaker does not answer the question.	Only one main ideas is presented with limited inclusion of its subtopics.	Two main ideas are presented with satisfactory inclusion of subtopics.	All the main ideas are present, but the subtopics are not fully developed.	All the main ideas and relevant subtopics are present.
Q 3	The speaker does not answer the question, or talks about an irrelevant experience.	The speaker mentions the advertisement technique, but gives the account of an irrelevant experience.	The speaker does not fully develop the story and makes no connection to the advertisement technique. The speaker may even forget to mention the technique.	The speaker mentions the advertisement technique and a related experience, but does not fully develop it to make a clear connection to the technique.	The speaker mentions the advertisement technique and fully develops the account of experience with clear connection with the technique.

Table 19 – Rubrics for Sociolinguistic Competence and *Questions Asked*

	Sociolinguistic Competence				
	0	1	2	3	4
Q 2 & 3	The speaker does not speak more than a few sentences to establish a style.	The response is mostly like an informal talk characteristic of friendly conversations.	There is a balance of both formal and informal styles and language.	The style is not completely academic, but those lapses are condonable.	The style of speaking is academic.
Q 1	The speaker does not speak more than a few sentences to establish a style.	The response is mostly like an academic lecture.	Half of the talk becomes like an academic lecture.	The style sometimes becomes formal as if the speaker is giving a lecture.	The style is completely appropriate to an informal and friendly context.
	Questions Asked				
	0	1	2	3	4
	The candidate asks no questions	**If one or more of the following are evident:** • The question has serious language problems. • It has little relevance to the topic, if any. • The pronunciation is mostly difficult to understand. • There is no formality in the question.	**If one or more of the following are evident:** • The question has serious language problems but it is relevant. • The language is generally good, but the question has little relevance to the topic. • There are instances of unclear pronunciation which make understanding difficult. • It is asked with little formality without undermining the teacher's authority with tone, vocabulary, or sarcasm.	**If one or more of the following are evident:** • The question has some language errors, but it is understandable. • It is relevant, but not completely. • There are some pronunciation problems, but it is understandable. • It is asked with appropriate formality.	**If ALL of the following are evident:** • The questions is grammatically correct. • It is relevant. • It is clear and understandable. • It is asked with appropriate formality.

B. Research Data

Table 20– Research Data of Engineering Participants

Participant	Participation Date	GPA	Weighted GPA	PTBST	Task1	Task2	Task3	EPE	Listening	Reading	Note-Taking	Writing	Cloze	Dialogue and Situation
1	07-Nov-2015	1.66	1.57	3.18	2.25	3.60	3.70	70.50	23.00	22.00	2.00	9.00	7.50	7.00
2	07-Nov-2015	3.11	3.00	3.73	3.60	3.70	3.90	88.00	27.00	25.00	5.00	13.00	8.50	9.50
3	08-Nov-2015	.74	.96	2.56	2.92	2.25	2.50	62.00	19.00	20.00	3.00	11.50	4.00	4.50
4	09-Nov-2015	1.56	1.42	2.69	2.17	2.90	3.00	61.00	17.00	19.00	4.00	9.50	5.50	6.00
5	19-Nov-2015	1.69	1.32	3.25	3.58	3.17	3.00	69.00	23.00	15.00	5.00	10.00	6.50	9.50
6	24-Nov-2015	3.71	3.65	2.94	3.25	3.17	2.40	83.00	28.00	23.00	4.50	10.00	8.00	9.50
7	29-Nov-2015	2.91	3.04	3.77	3.80	3.90	3.60	78.50	28.00	25.00	3.50	11.00	4.50	6.50
8	10-Dec-2015	2.94	2.77	2.97	2.42	3.08	3.40	77.00	21.00	26.00	4.00	11.00	5.50	9.50
9	13-Dec-2015	1.50	1.35	1.80	1.50	2.70	1.20	66.50	23.00	18.00	3.50	9.00	6.00	7.00
10	14-Dec-2015	1.12	1.46	2.57	2.33	2.67	2.70	68.00	21.00	18.00	3.50	11.50	4.50	9.50
11	21-Dec-2015	2.00	1.86	2.43	3.00	2.60	1.70	81.50	27.00	23.00	5.00	12.00	7.00	7.50
12	23-Dec-2015	.88	.00	3.24	3.33	3.70	2.70	77.00	26.00	20.00	4.50	13.50	5.00	8.00
13	27-Dec-2015	3.07	2.50	2.67	3.00	3.10	1.90	77.00	22.00	24.00	5.00	12.00	5.00	9.00

Table 21 – Research Data of Non-Engineering Participants

Participant	Participation Date	GPA	Weighted GPA	PTBST	Task1	Task2	Task3	EPE	Listening	Reading	Note-Taking	Writing	Cloze	Dialogue and Situation
14	10-Oct-2015	2.41	2.38	2.86	2.67	3.60	2.30	63.50	20.00	15.00	4.00	11.50	6.50	6.50
15	16-Oct-2015	.82	.62	2.31	2.33	2.40	2.20	67.00	18.00	20.00	4.00	11.50	5.50	8.00
16	05-Nov-2015	3.75	3.68	3.23	3.70	2.70	3.30	86.00	27.00	29.00	3.50	10.00	7.50	9.00
17	11-Nov-2015	1.38	.50	2.90	3.10	2.90	2.70	71.50	21.00	20.00	5.00	12.50	6.00	7.00
18	11-Nov-2015	3.91	3.88	3.41	3.33	3.40	3.50	79.50	24.00	25.00	4.00	12.00	7.00	7.50
19	11-Nov-2015	1.12	1.00	2.97	3.00	3.00	2.90	74.50	22.00	21.00	5.00	12.50	6.50	7.50
20	11-Nov-2015	1.64	1.39	2.57	3.10	2.70	1.90	59.50	16.00	15.00	4.00	12.50	3.00	9.00
21	15-Nov-2015	2.17	2.21	3.10	3.30	3.50	2.50	64.00	19.00	19.00	3.00	10.50	3.50	9.00
22	19-Nov-2015	.85	.35	2.63	2.50	2.80	2.60	59.50	18.00	18.00	3.00	11.50	4.00	5.00
23	27-Nov-2015	.35	.00	2.79	2.58	3.30	2.50	76.00	25.00	24.00	2.50	11.50	6.00	7.00
24	10-Dec-2015	3.75	3.68	3.13	2.80	3.40	3.20	78.00	24.00	24.00	4.50	9.00	6.50	10.00
25	15-Dec-2015	1.66	.88	3.90	3.90	4.00	3.80	82.50	28.00	26.00	4.00	11.00	7.00	6.50
26	17-Dec-2015	2.38	2.50	3.11	3.42	3.60	2.30	83.00	27.00	27.00	4.00	10.00	8.00	7.00
27	17-Dec-2015	3.85	3.78	3.08	3.33	2.90	3.00	90.00	30.00	28.00	4.00	11.50	8.00	8.50
28	23-Dec-2015	1.65	1.65	2.67	3.20	3.00	1.80	59.50	17.00	20.00	3.00	9.50	4.00	6.00
29	23-Dec-2015	2.88	2.85	3.27	3.50	3.40	2.90	77.50	26.00	24.00	4.50	9.00	5.50	8.50
30	25-Dec-2015	2.71	2.62	2.93	2.10	3.40	3.30	63.50	21.00	18.00	4.50	9.00	5.50	5.50
31	30-Dec-2015	4.00	4.00	3.13	3.10	3.00	3.30	59.50	14.00	18.00	4.00	9.50	5.50	8.50

C. Speaking Grading Forms

Student Number	
Rater's Name	

Question 1 Rubrics					
Linguistic Competence (Grammar, Vocabulary, and Pronunciation)	**0**	**1**	**2**	**3**	**4**
Delivery (Fluency and Comprehensibility)	**0**	**1**	**2**	**3**	**4**
Discourse--Coherence & Organization (pre-stories, meta-statements for general organization, meta-statements for subtopics, linkers, intonation)	**0**	**1**	**2**	**3**	**4**
Task Accomplishment (Cohesion and Inclusion of Major and Supporting Ideas)	**0**	**1**	**2**	**3**	**4**
Sociolinguistic Competence (Correct Use of Jargons, Informal and friendly Style)	**0**	**1**	**2**	**3**	**4**
Questions Asked	**0**	**1**	**2**	**3**	**4**

Strategic Competence

Do not penalize a student on any of the above features if he/she uses strategies like use of repairs, backtracks, synonyms, etc.

Rater's Comments

Fold

Question 2 Rubrics					
Linguistic Competence (Grammar, Vocabulary, and Pronunciation)	**0**	**1**	**2**	**3**	**4**
Delivery (Fluency and Comprehensibility)	**0**	**1**	**2**	**3**	**4**
Discourse--Coherence & Organization (meta-statements for general organization, meta-statements for subtopics, linkers, intonation)	**0**	**1**	**2**	**3**	**4**
Task Accomplishment (Cohesion and Inclusion of Major and Supporting Ideas)	**0**	**1**	**2**	**3**	**4**
Sociolinguistic Competence (Correct Use of Jargons, Formal and Friendly Style)	**0**	**1**	**2**	**3**	**4**
Questions Asked	**0**	**1**	**2**	**3**	**4**

Strategic Competence

Do not penalize a student on any of the above features if he/she uses strategies like use of repairs, backtracks, synonyms, etc.

Rater's Comments

Question 3 Rubrics					
Linguistic Competence (Grammar, Vocabulary, and Pronunciation)	**0**	**1**	**2**	**3**	**4**
Delivery (Fluency and Comprehensibility)	**0**	**1**	**2**	**3**	**4**
Discourse--Coherence & Organization (pre-stories, meta-statements for general organization, meta-statements for subtopics, linkers, intonation)	**0**	**1**	**2**	**3**	**4**
Task Accomplishment (Cohesion and Inclusion of Major and Supporting Ideas)	**0**	**1**	**2**	**3**	**4**
Sociolinguistic Competence (Correct Use of Jargons, Formal and friendly Style)	**0**	**1**	**2**	**3**	**4**

Strategic Competence

Do not penalize a student on any of the above features if he/she uses strategies like use of repairs, backtracks, synonyms, etc.

Rater's Comments

D. Task Descriptions Provided To Raters

Questions Number 1

Your classmate missed the first session of the class. He wants to know what the course syllabus and requirements are. Get ready to leave a voicemail on his phone.

1. Exam
   - There will be two exams
   - They will form 60% of the final grade
   - Each exam will cover half of the book
   - 80 percent of the questions will be in multiple-choice format
   - One exam will be in the middle and the other will be at the end of the semester
2. Assignments
   - Will form 30% of the final grade
   - Have to be in hard-copy
   - Must be submitted not after the due date
3. Class attendance
   - Will form 10% of the final grade
   - There will be 16 sessions
   - Three sessions of absenteeism is allowed
   - Take part in class activities and take notes
4. Sources
   - Coursebook and lectures
   - Coursebook is downloadable
   - Hard copies are available
   - Sharing of the digital version is not allowed
   - No pamphlets or notes will be provided

Questions Number 2

Using professor's examples and the points in the reading passage, explain the three methods advertisers use to make an expensive product seem cheap.

Three advertisement techniques

1. Contrast Effect
   - A more expensive product is placed next to the current (already expensive) product to create the misperception that the advertised product is not really expensive.

Example: a bread-maker in a magazine was placed next to a more expensive one. Although it was expensive, people started buying.

2. Partial Price Quoting
    - The total price of the product is not quoted. Instead, it is divided over several months or days and the monthly or daily payments are quoted, making the product look cheap.
      Example: buy iPhone 6 for 3 dollars a day.
3. Normalizing
    - Showing the number of people who have bought the product, they implant the impression that it is alright to pay a high price for a product.
      Example: Trying bungee jumping after seeing others do it; buying a pair of jeans for 300 dollars simply because others have done it and it is OK to pay such an outrageous price.

Question 3

Think about an experience of buying an expensive product that seemed cheap to you. Describe your experience with details and examples. Explain what advertisement technique convinced you to by the product.

A clear reference to one of the techniques above and providing a personal example.

- The personal example should be well constructed.
- There should be clear connections between the technique and the example.

E. Reading Passage Of The Second Task

READING PASSAGE FOR QUESTIN 2

Three Ways to Make Expensive Products Look Like Steal

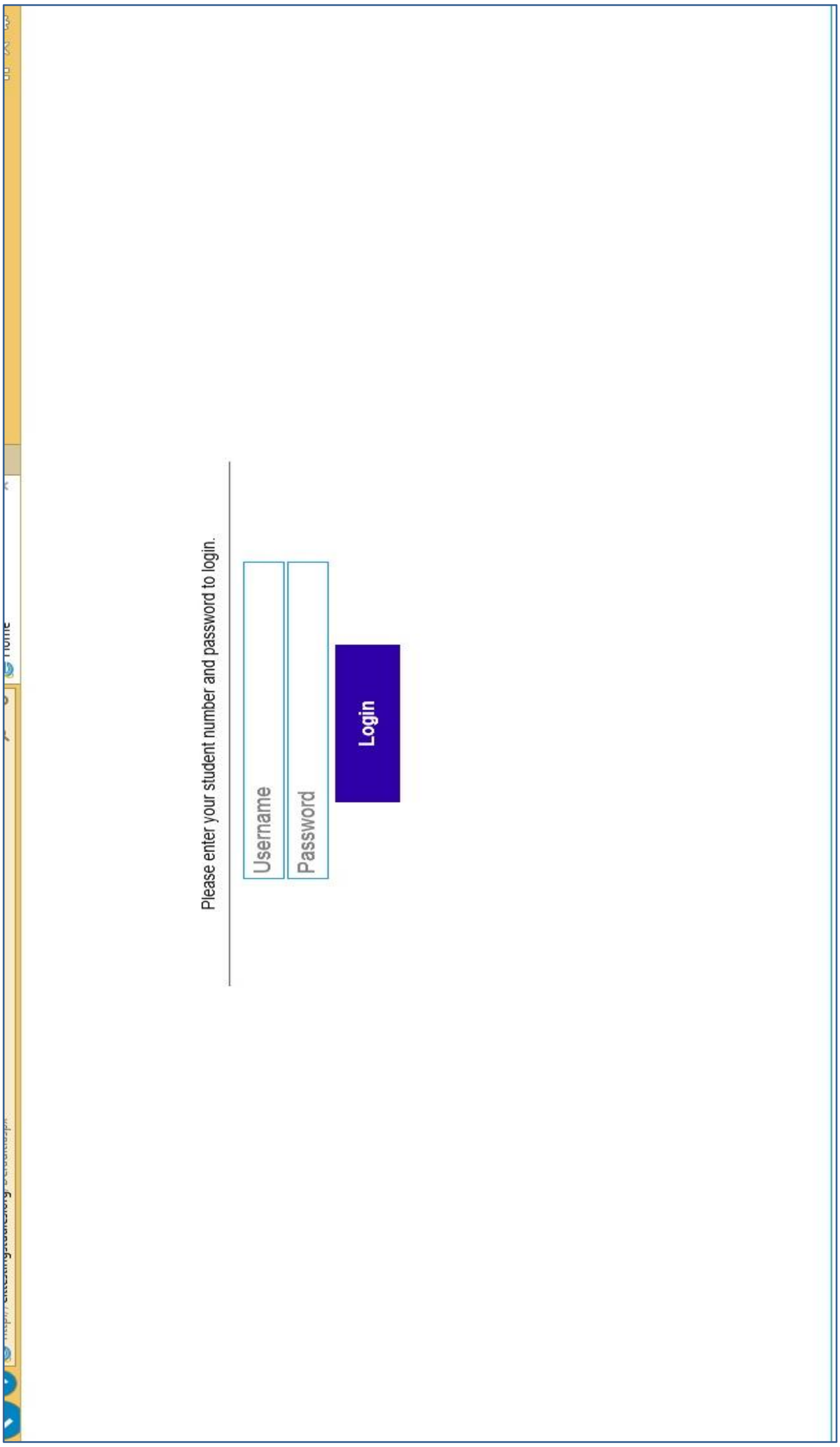
Businesses and companies do not want their products to seem either cheap or expensive. Customers associate low price with lower quality. Likewise, when the price seems too expensive, every customer is bound to think twice. So, advertisement agencies try to shift potential customers' attention away from price and onto product value. They normally use three main techniques.

The first technique is called "Contrast Effect". Using this technique, a similar product is chosen with a higher price and introduced next to the advertised one. This way, potential customers subconsciously come to the conclusion that no matter how expensive the product is, there is a more expensive one out there and as a result worry less about the high price.

The second technique is "Partial Price Quoting." Basically, in advertisement, agencies and companies do not quote the total price which might seem very expensive at first glance. Instead, they divide the price into monthly or even daily payments. Since the focus of attention is the price, not the time period, the price seems to be very low.

The final technique is "normalizing." This method simply convinces the potential customers that the mentioned price is a normal one, even if it really is not, by showing them the large number of people who have bought the product. As a result, paying a high price does not seem abnormal anymore

F. PTBST Screenshots

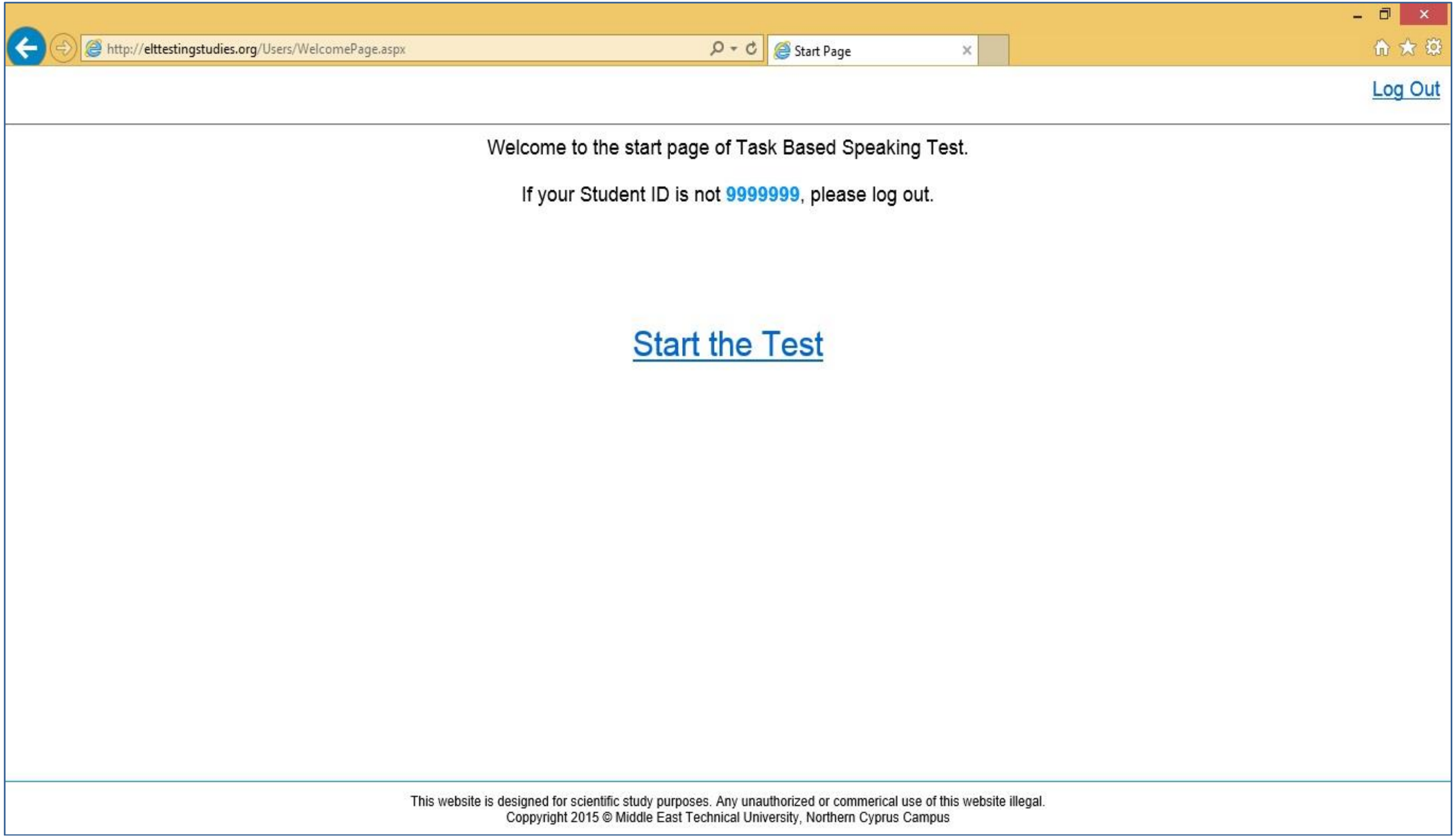


Please enter your student number and password to login.

Figure 3 – Login Page of PTBST

Figure 4 – PTBST control panel

The user is redirected to this page after logging in. The user name is the 7-digit student number.



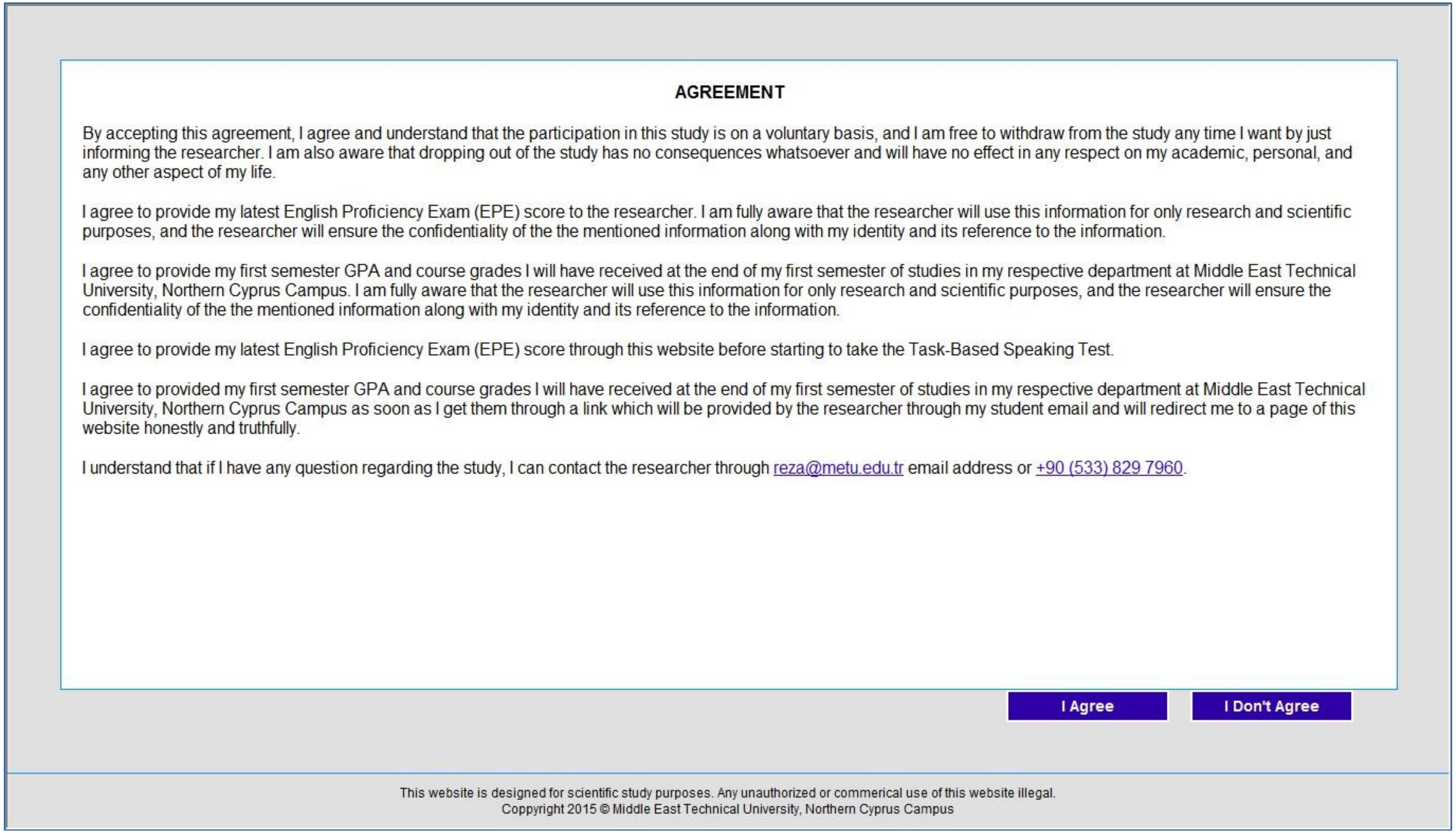
Welcome to the start page of Task Based Speaking Test.

If your Student ID is not 9999999, please log out.

Start the Test

This website is designed for scientific study purposes. Any unauthorized or commerical use of this website illegal.
Copyright 2015 © Middle East Technical University, Northern Cyprus Campus

Figure 5 – Digital Consent Form



AGREEMENT

By accepting this agreement, I agree and understand that the participation in this study is on a voluntary basis, and I am free to withdraw from the study any time I want by just informing the researcher. I am also aware that dropping out of the study has no consequences whatsoever and will have no effect in any respect on my academic, personal, and any other aspect of my life.

I agree to provide my latest English Proficiency Exam (EPE) score to the researcher. I am fully aware that the researcher will use this information for only research and scientific purposes, and the researcher will ensure the confidentiality of the the mentioned information along with my identity and its reference to the information.

I agree to provide my first semester GPA and course grades I will have received at the end of my first semester of studies in my respective department at Middle East Technical University, Northern Cyprus Campus. I am fully aware that the researcher will use this information for only research and scientific purposes, and the researcher will ensure the confidentiality of the the mentioned information along with my identity and its reference to the information.

I agree to provide my latest English Proficiency Exam (EPE) score through this website before starting to take the Task-Based Speaking Test.

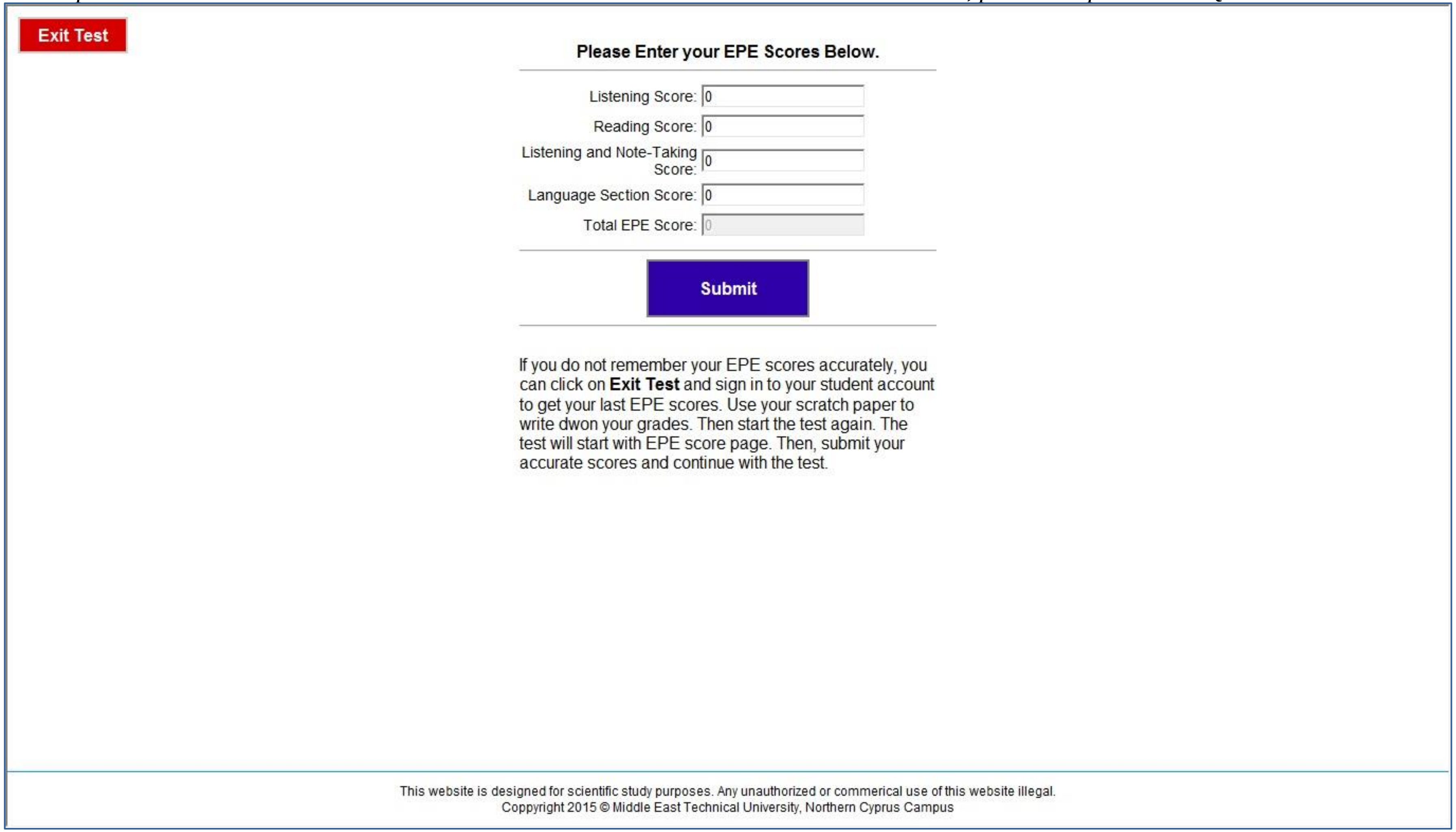
I agree to provided my first semester GPA and course grades I will have received at the end of my first semester of studies in my respective department at Middle East Technical University, Northern Cyprus Campus as soon as I get them through a link which will be provided by the researcher through my student email and will redirect me to a page of this website honestly and truthfully.

I understand that if I have any question regarding the study, I can contact the researcher through reza@metu.edu.tr email address or +90 (533) 829 7960.

This website is designed for scientific study purposes. Any unauthorized or commerical use of this website illegal.
Coppyright 2015 © Middle East Technical University, Northern Cyprus Campus

Figure 6 – EPE Score Data Collection Page

Participants are asked to enter their latest EPE scores here. The data are stored in a remote, password-protected SQL Server database.



Please Enter your EPE Scores Below.

Listening Score: 0

Reading Score: 0

Listening and Note-Taking Score: 0

Language Section Score: 0

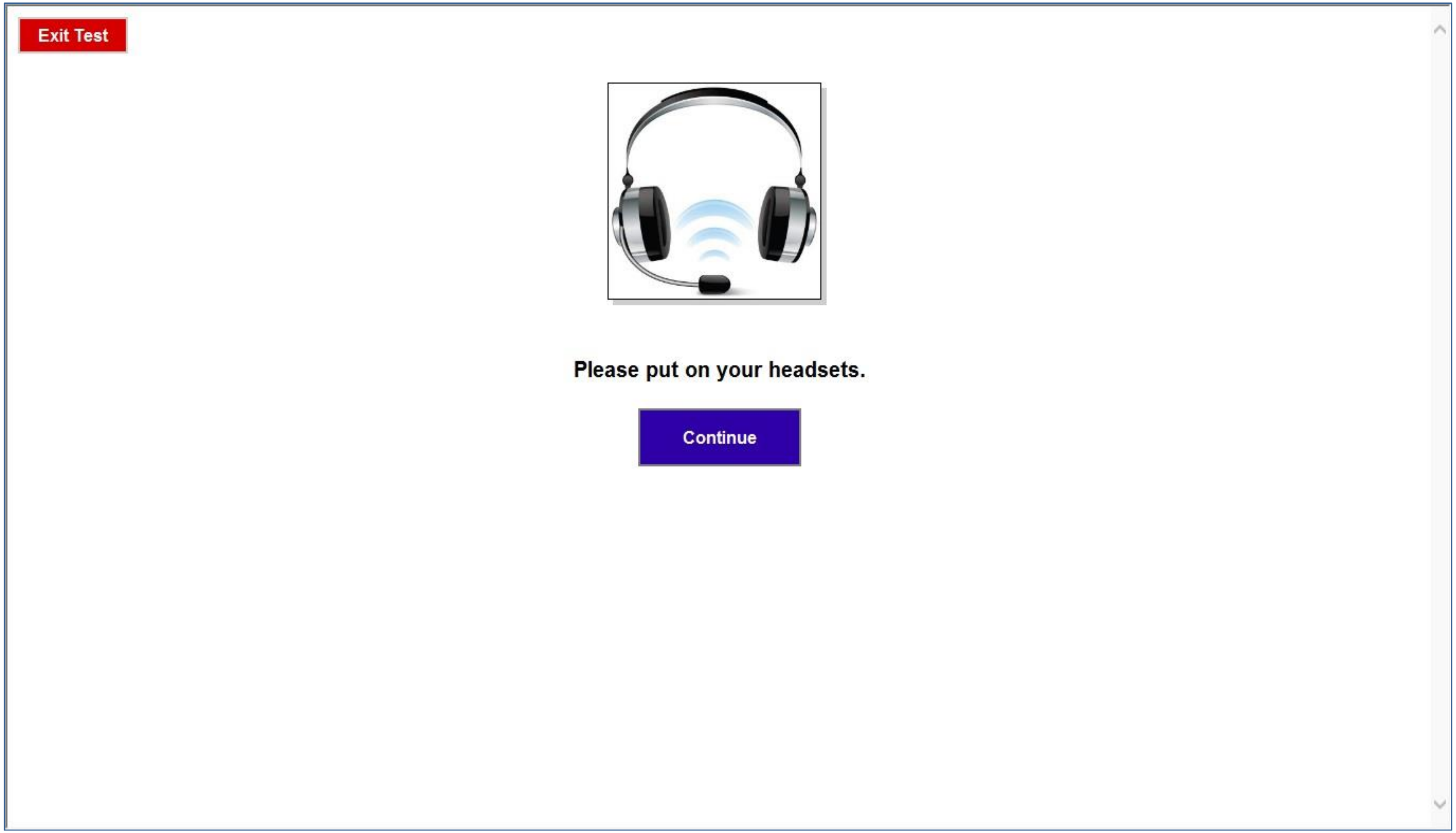
Total EPE Score: 0

If you do not remember your EPE scores accurately, you can click on **Exit Test** and sign in to your student account to get your last EPE scores. Use your scratch paper to write dwon your grades. Then start the test again. The test will start with EPE score page. Then, submit your accurate scores and continue with the test.

This website is designed for scientific study purposes. Any unauthorized or commerical use of this website illegal.

Figure 7 – Headset Notification Page

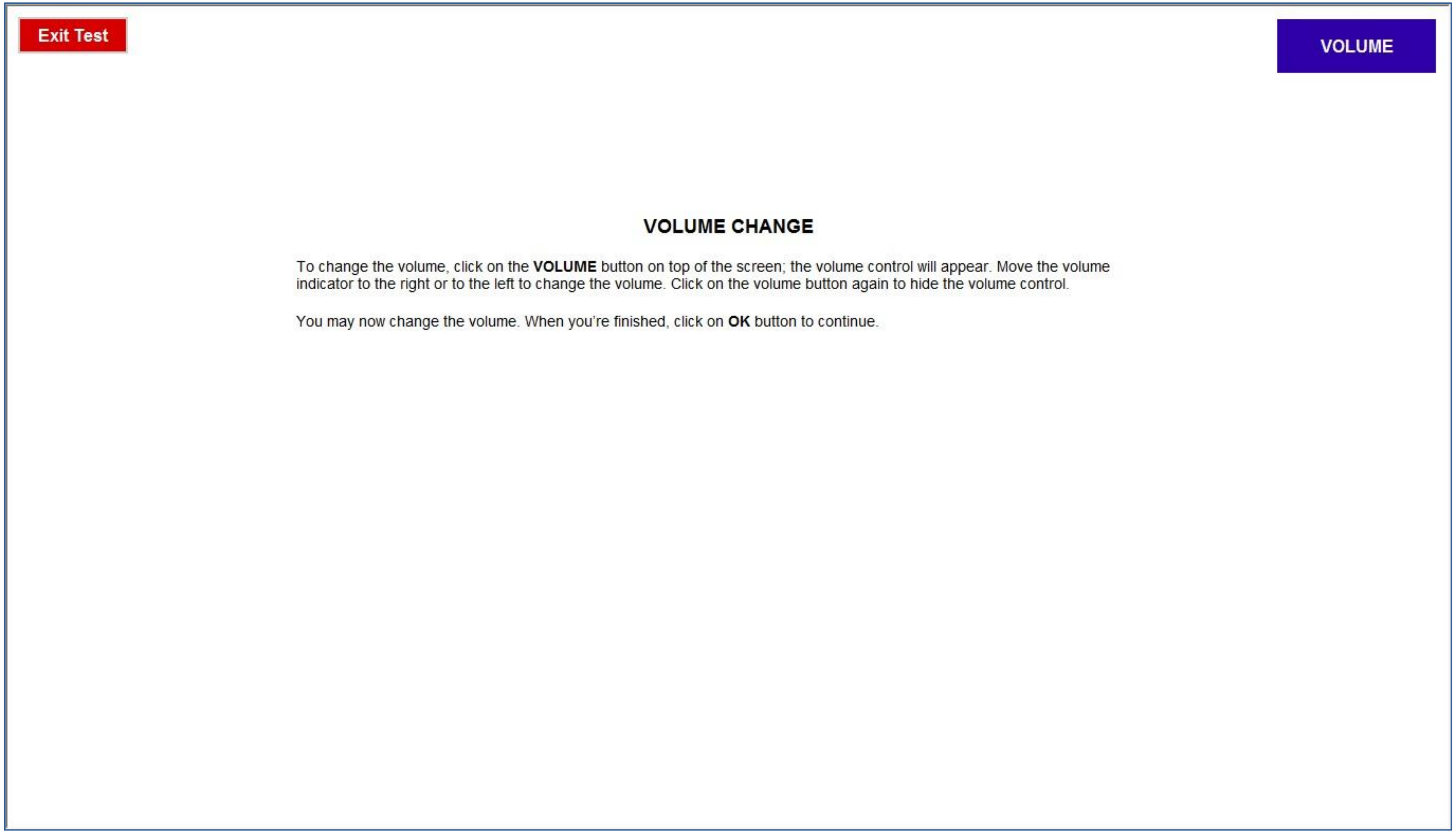
This is to let the participants know that they need to have their headsets on and that the hardware check is about to begin.



Please put on your headsets.

Figure 8 – Volume Adjustment Page

The participant can set the volume to a level comfortable to his/her ears. The sound files are all normalized using Adobe Audition software.



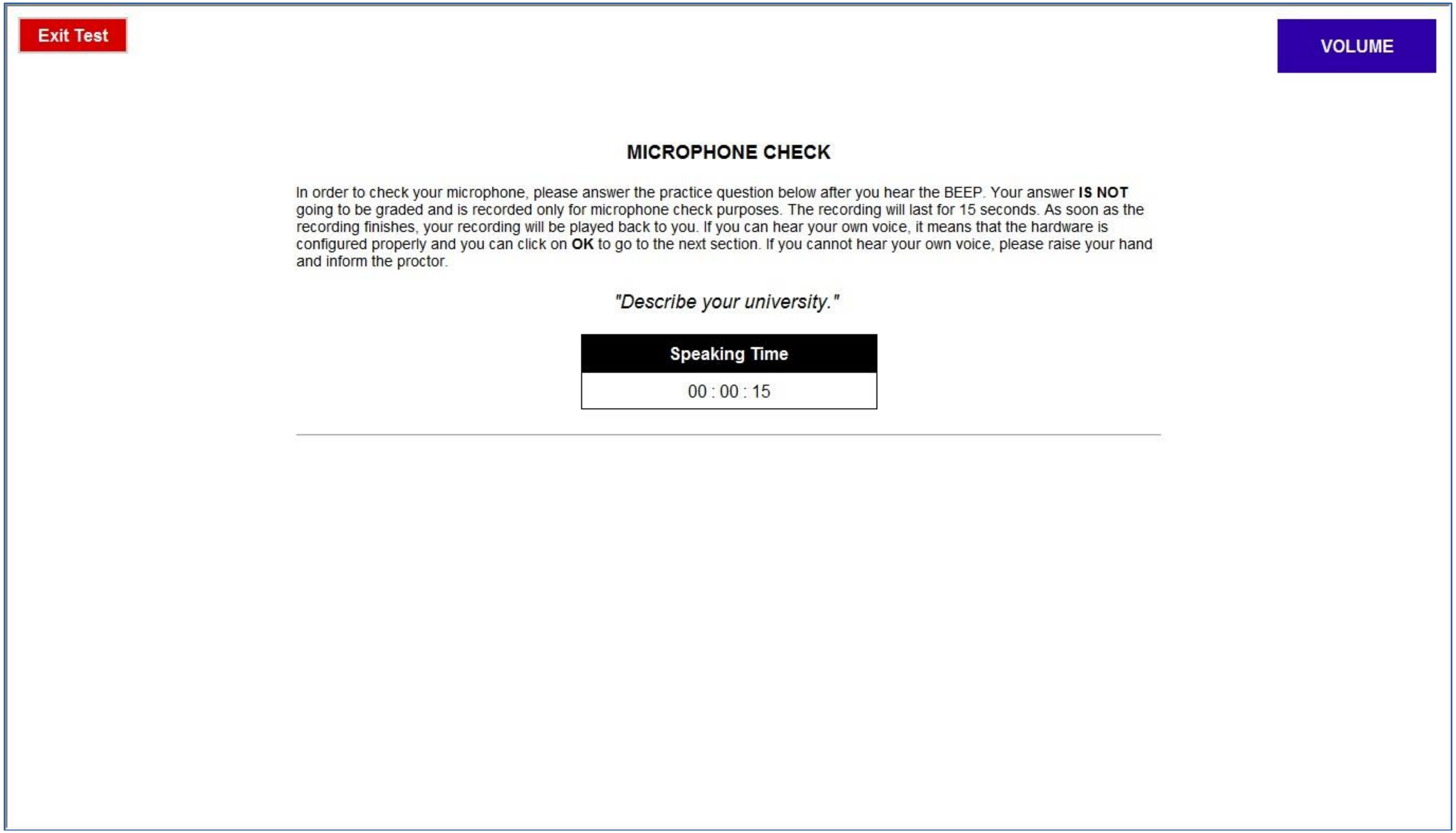
VOLUME CHANGE

To change the volume, click on the **VOLUME** button on top of the screen; the volume control will appear. Move the volume indicator to the right or to the left to change the volume. Click on the volume button again to hide the volume control.

You may now change the volume. When you're finished, click on **OK** button to continue.

Figure 9 – Recording Hardware Check Page

This page checks if the microphone is recording and sets the recording setting for an optimum sound quality.



MICROPHONE CHECK

In order to check your microphone, please answer the practice question below after you hear the BEEP. Your answer **IS NOT** going to be graded and is recorded only for microphone check purposes. The recording will last for 15 seconds. As soon as the recording finishes, your recording will be played back to you. If you can hear your own voice, it means that the hardware is configured properly and you can click on **OK** to go to the next section. If you cannot hear your own voice, please raise your hand and inform the proctor.

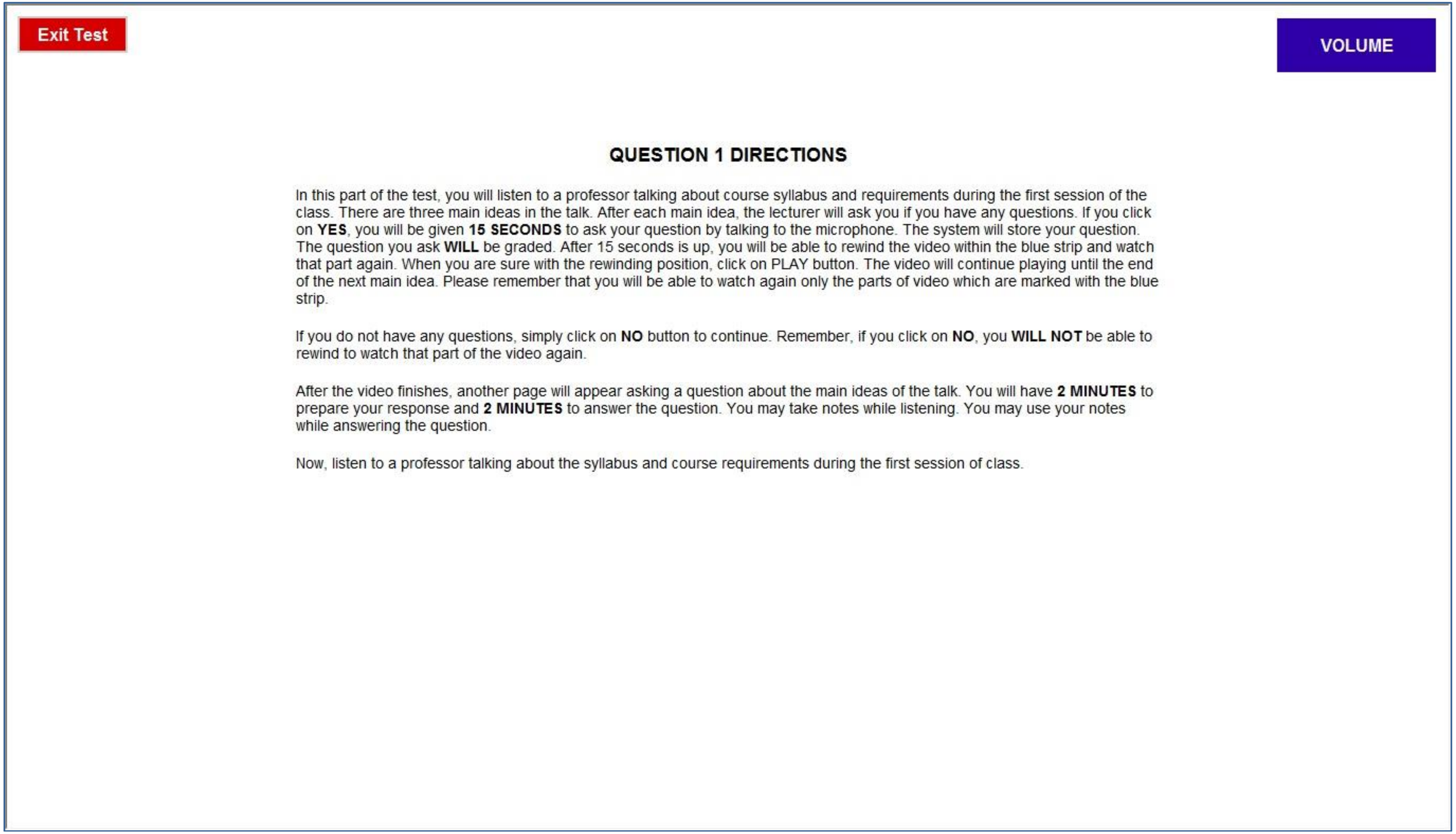
"Describe your university."

Speaking Time

00 : 00 : 15

Figure 10 – Directions of Question One

A recording reads through the instructions to ensure that the participant pays attention to all instructions. It also provides a warm-up for



QUESTION 1 DIRECTIONS

In this part of the test, you will listen to a professor talking about course syllabus and requirements during the first session of the class. There are three main ideas in the talk. After each main idea, the lecturer will ask you if you have any questions. If you click on **YES**, you will be given **15 SECONDS** to ask your question by talking to the microphone. The system will store your question. The question you ask **WILL** be graded. After 15 seconds is up, you will be able to rewind the video within the blue strip and watch that part again. When you are sure with the rewinding position, click on PLAY button. The video will continue playing until the end of the next main idea. Please remember that you will be able to watch again only the parts of video which are marked with the blue strip.

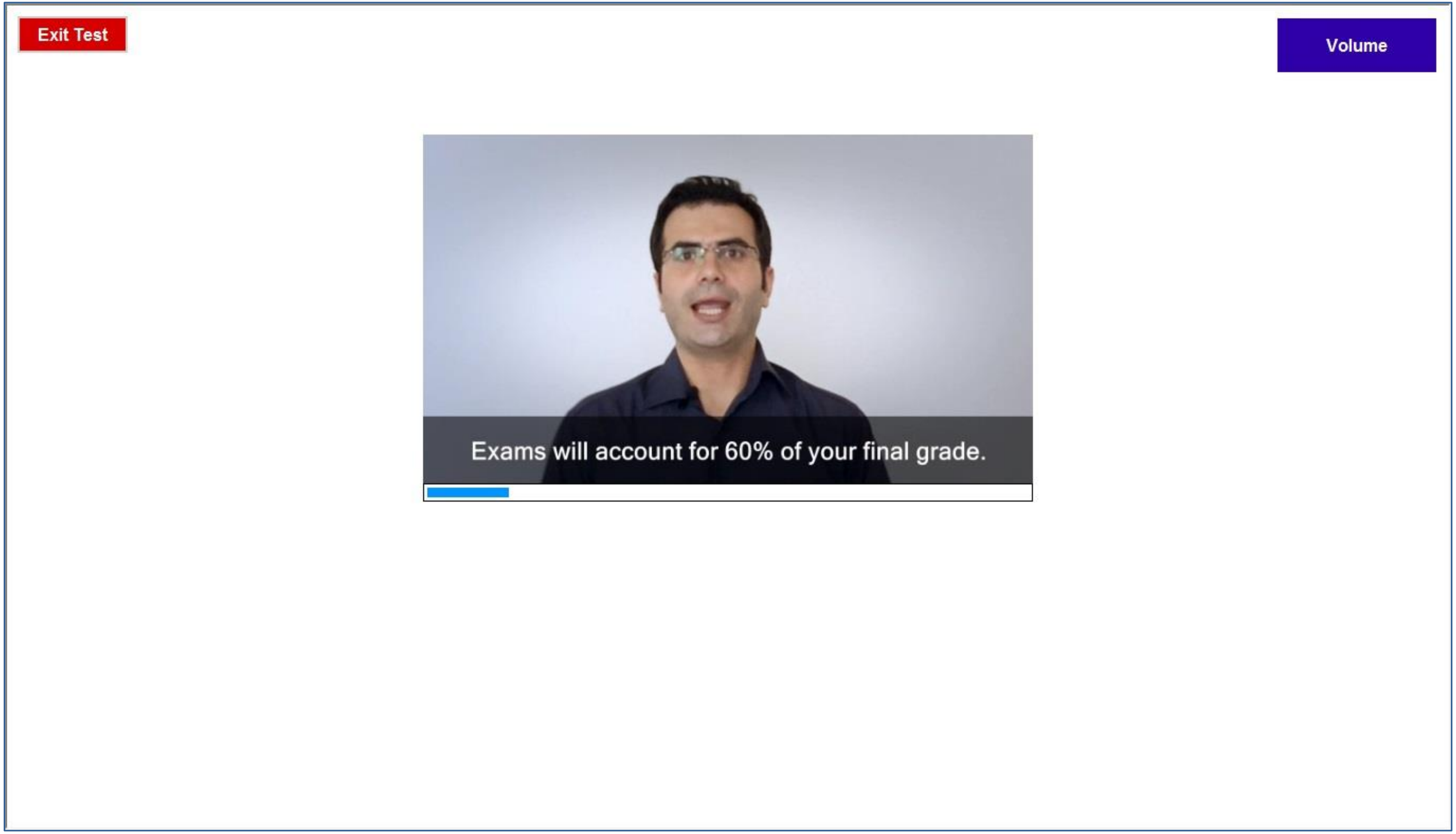
If you do not have any questions, simply click on **NO** button to continue. Remember, if you click on **NO**, you **WILL NOT** be able to rewind to watch that part of the video again.

After the video finishes, another page will appear asking a question about the main ideas of the talk. You will have **2 MINUTES** to prepare your response and **2 MINUTES** to answer the question. You may take notes while listening. You may use your notes while answering the question.

Now, listen to a professor talking about the syllabus and course requirements during the first session of class.

Figure 11 – Lecture Video Page for Question One

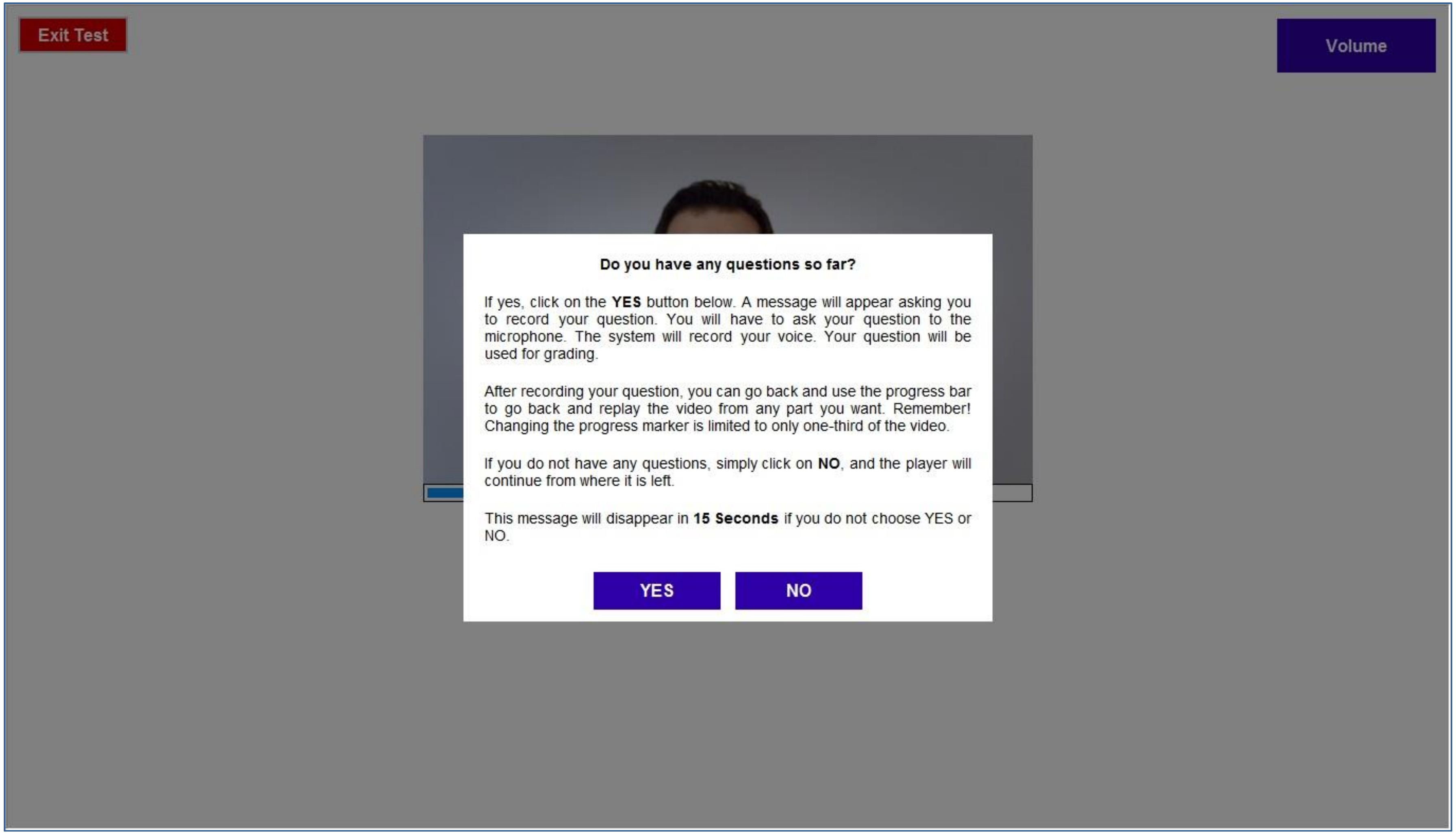
Captions and illustrations are included to add to the diversity of stimuli as it is the case in real world communication.



Exams will account for 60% of your final grade.

Figure 12 – Prompt Page for *Questions Asked* Feature of PTBST

This is the message which appears once after each main idea of the lecture. The speaker in the video asks: "Do you have any questions?"



Do you have any questions so far?

If yes, click on the **YES** button below. A message will appear asking you to record your question. You will have to ask your question to the microphone. The system will record your voice. Your question will be used for grading.

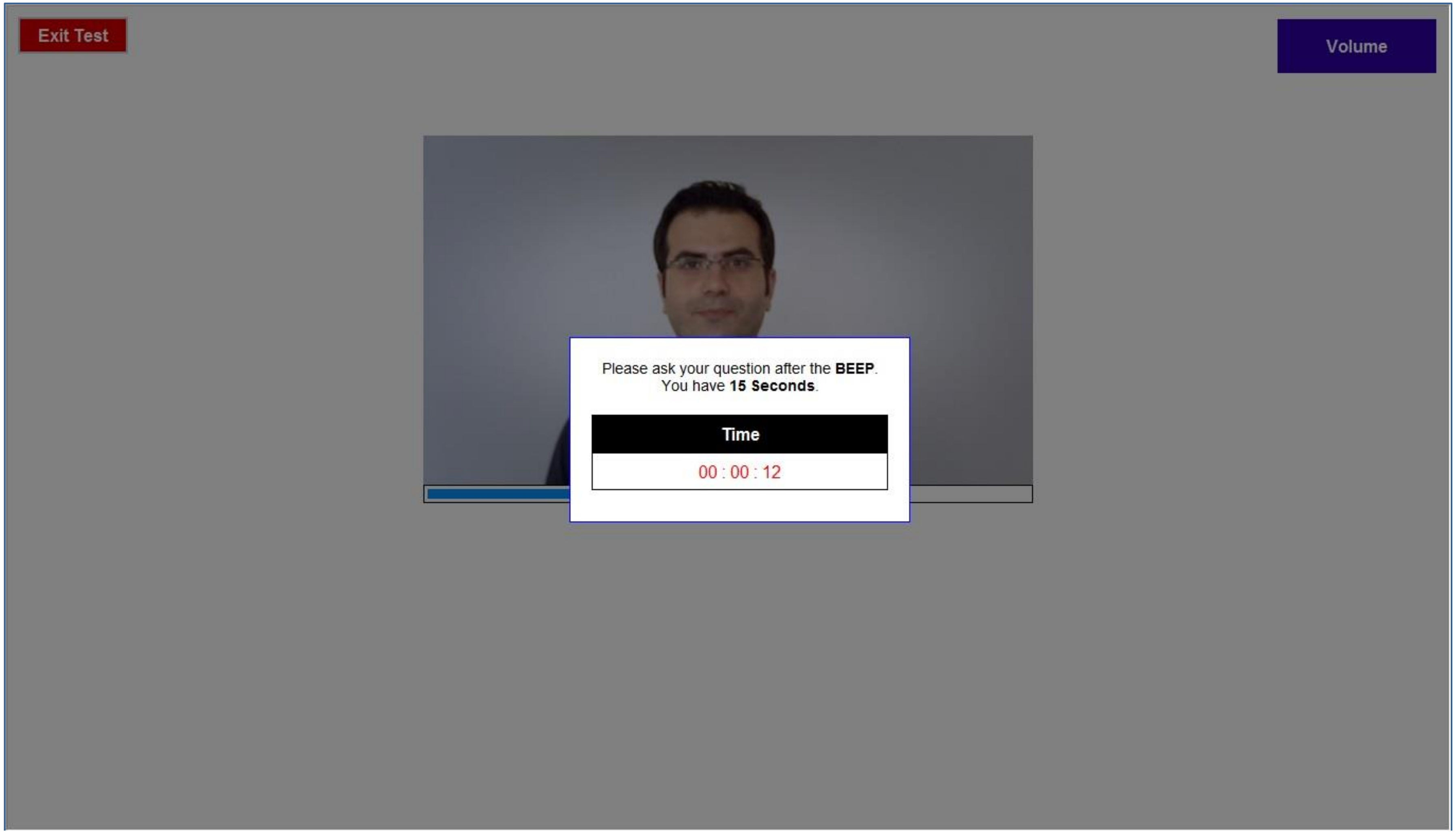
After recording your question, you can go back and use the progress bar to go back and replay the video from any part you want. Remember! Changing the progress marker is limited to only one-third of the video.

If you do not have any questions, simply click on **NO**, and the player will continue from where it is left.

This message will disappear in **15 Seconds** if you do not choose YES or NO.

Figure 13 – *Questions Asked* Recording Timer

The countdown timer shows how much time test-takers have to ask their quesiotn.



Please ask your question after the **BEEP**.
You have **15 Seconds**.

Time
00 : 00 : 12

Figure 14 – Progress Bar Tweaking Control

The blue strip shows the allowed boundary for tweaking. It covers the part of the video related to the main idea which has just finished.

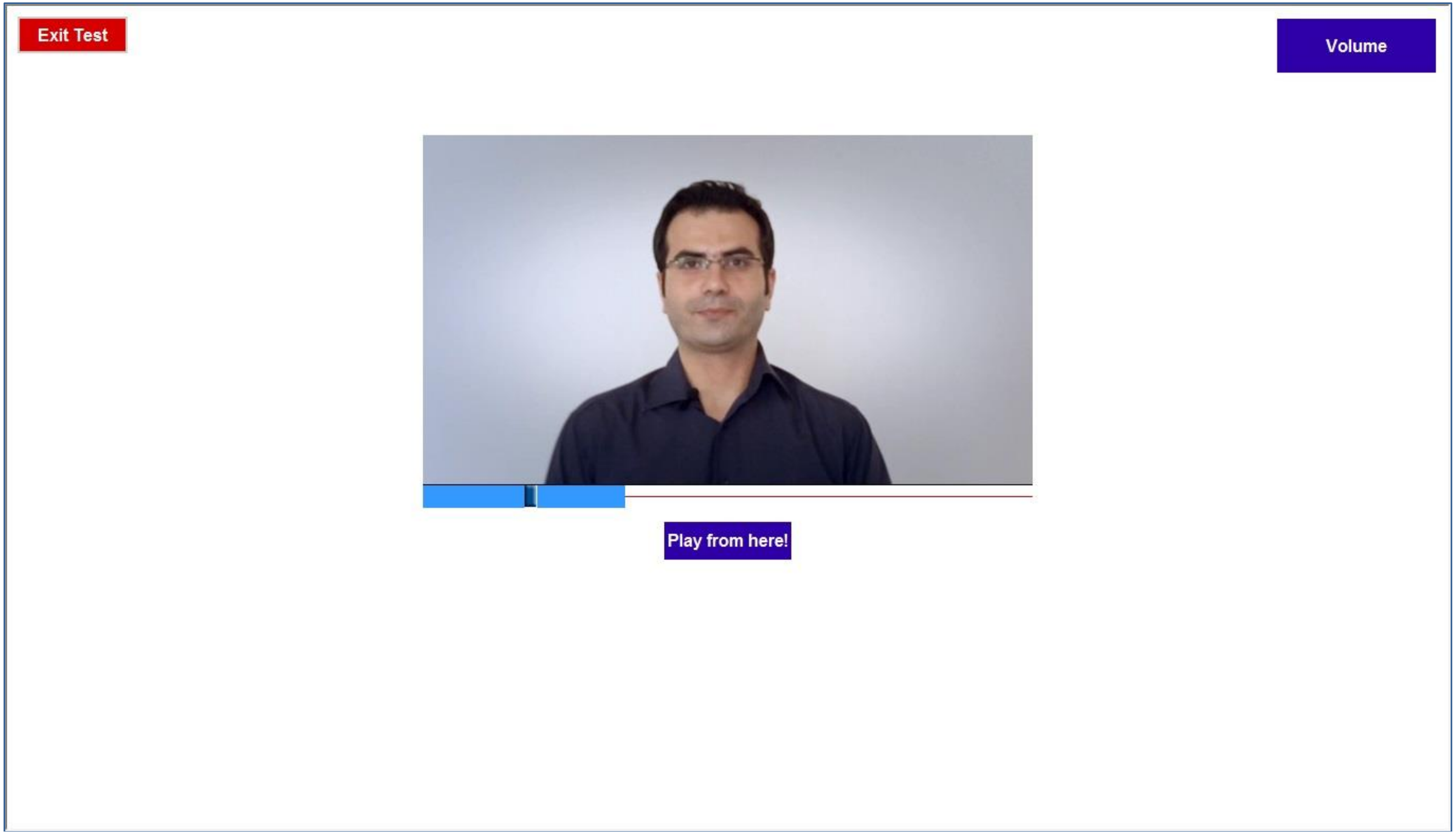
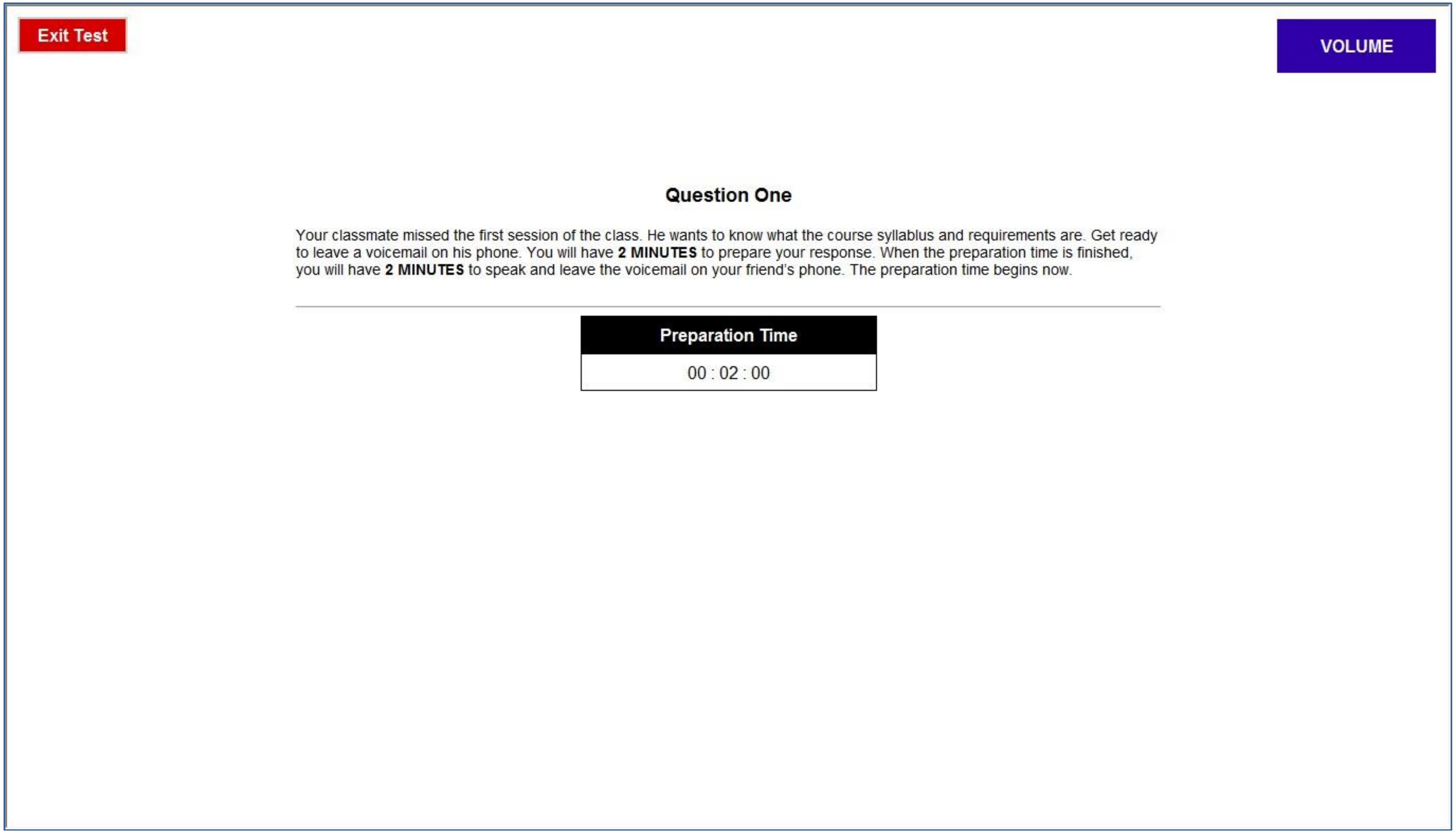


Figure 15 – Task One Question Prompt Page

This page contains the question which the participant is required to answer about the lecture. The prompt is also read aloud by a recording.

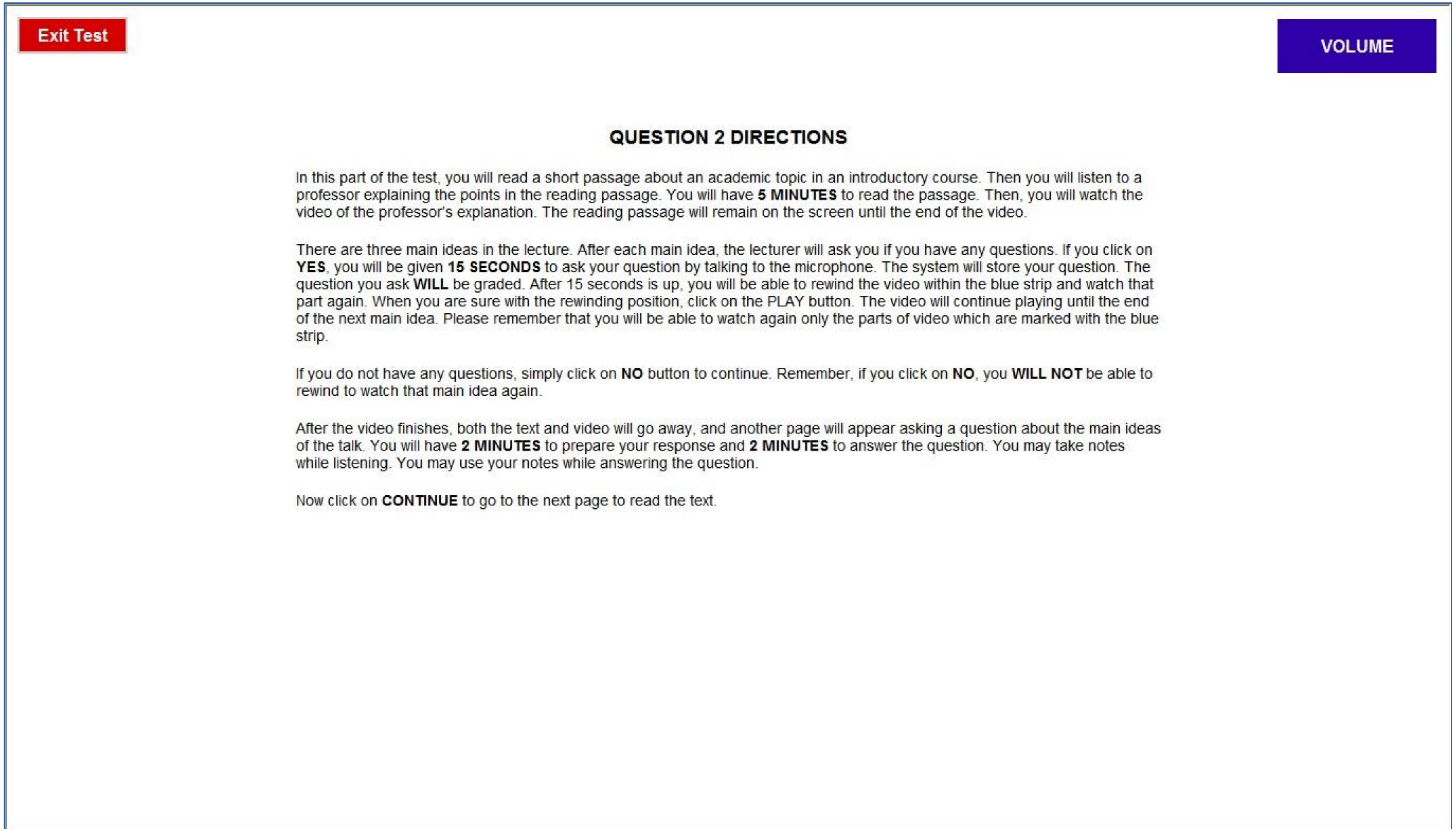


Question One

Your classmate missed the first session of the class. He wants to know what the course syllablus and requirements are. Get ready to leave a voicemail on his phone. You will have **2 MINUTES** to prepare your response. When the preparation time is finished, you will have **2 MINUTES** to speak and leave the voicemail on your friend's phone. The preparation time begins now.

Preparation Time
00 : 02 : 00

Figure 16 – Directions of Task Two



QUESTION 2 DIRECTIONS

In this part of the test, you will read a short passage about an academic topic in an introductory course. Then you will listen to a professor explaining the points in the reading passage. You will have **5 MINUTES** to read the passage. Then, you will watch the video of the professor's explanation. The reading passage will remain on the screen until the end of the video.

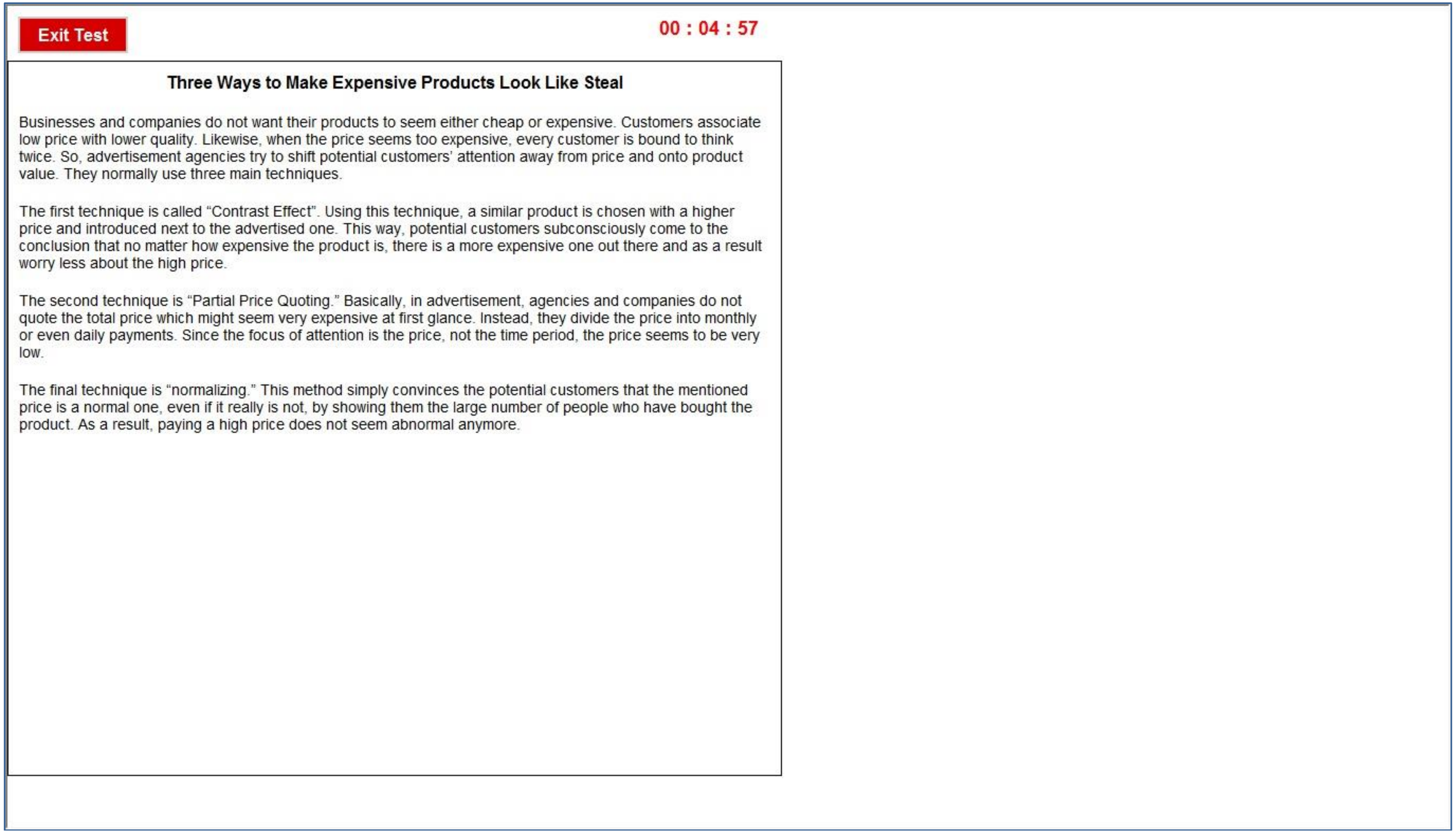
There are three main ideas in the lecture. After each main idea, the lecturer will ask you if you have any questions. If you click on **YES**, you will be given **15 SECONDS** to ask your question by talking to the microphone. The system will store your question. The question you ask **WILL** be graded. After 15 seconds is up, you will be able to rewind the video within the blue strip and watch that part again. When you are sure with the rewinding position, click on the PLAY button. The video will continue playing until the end of the next main idea. Please remember that you will be able to watch again only the parts of video which are marked with the blue strip.

If you do not have any questions, simply click on **NO** button to continue. Remember, if you click on **NO**, you **WILL NOT** be able to rewind to watch that main idea again.

After the video finishes, both the text and video will go away, and another page will appear asking a question about the main ideas of the talk. You will have **2 MINUTES** to prepare your response and **2 MINUTES** to answer the question. You may take notes while listening. You may use your notes while answering the question.

Now click on **CONTINUE** to go to the next page to read the text.

Figure 17 – The Reading Passage Page of Task Two
The participants have five minutes to read the whole passage and take notes.



Three Ways to Make Expensive Products Look Like Steal

Businesses and companies do not want their products to seem either cheap or expensive. Customers associate low price with lower quality. Likewise, when the price seems too expensive, every customer is bound to think twice. So, advertisement agencies try to shift potential customers' attention away from price and onto product value. They normally use three main techniques.

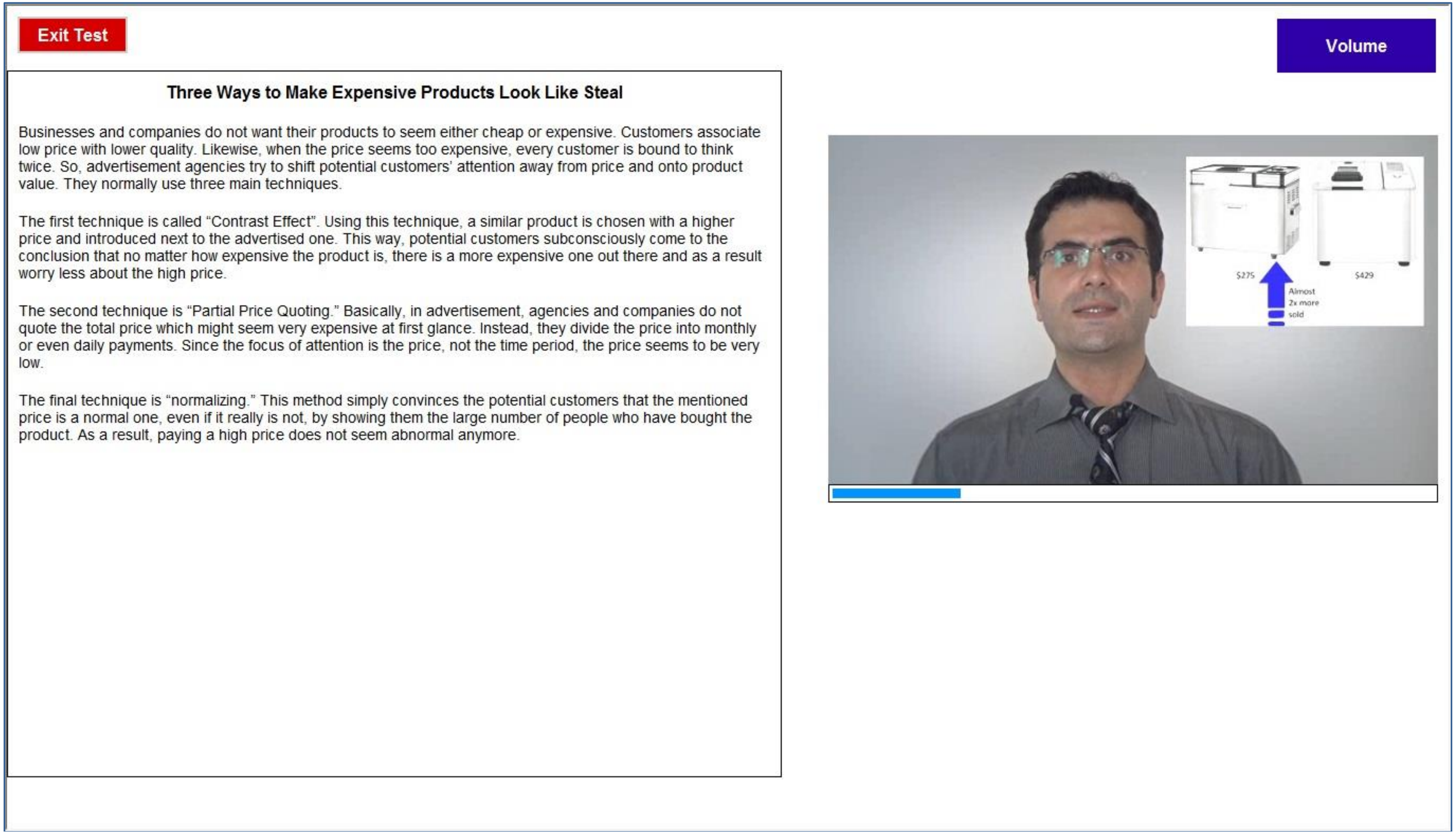
The first technique is called "Contrast Effect". Using this technique, a similar product is chosen with a higher price and introduced next to the advertised one. This way, potential customers subconsciously come to the conclusion that no matter how expensive the product is, there is a more expensive one out there and as a result worry less about the high price.

The second technique is "Partial Price Quoting." Basically, in advertisement, agencies and companies do not quote the total price which might seem very expensive at first glance. Instead, they divide the price into monthly or even daily payments. Since the focus of attention is the price, not the time period, the price seems to be very low.

The final technique is "normalizing." This method simply convinces the potential customers that the mentioned price is a normal one, even if it really is not, by showing them the large number of people who have bought the product. As a result, paying a high price does not seem abnormal anymore.

Figure 18 – Lecture Page of the Second Task

The reading passage is accessible to the test-takers while listening to the lecture.



Three Ways to Make Expensive Products Look Like Steal

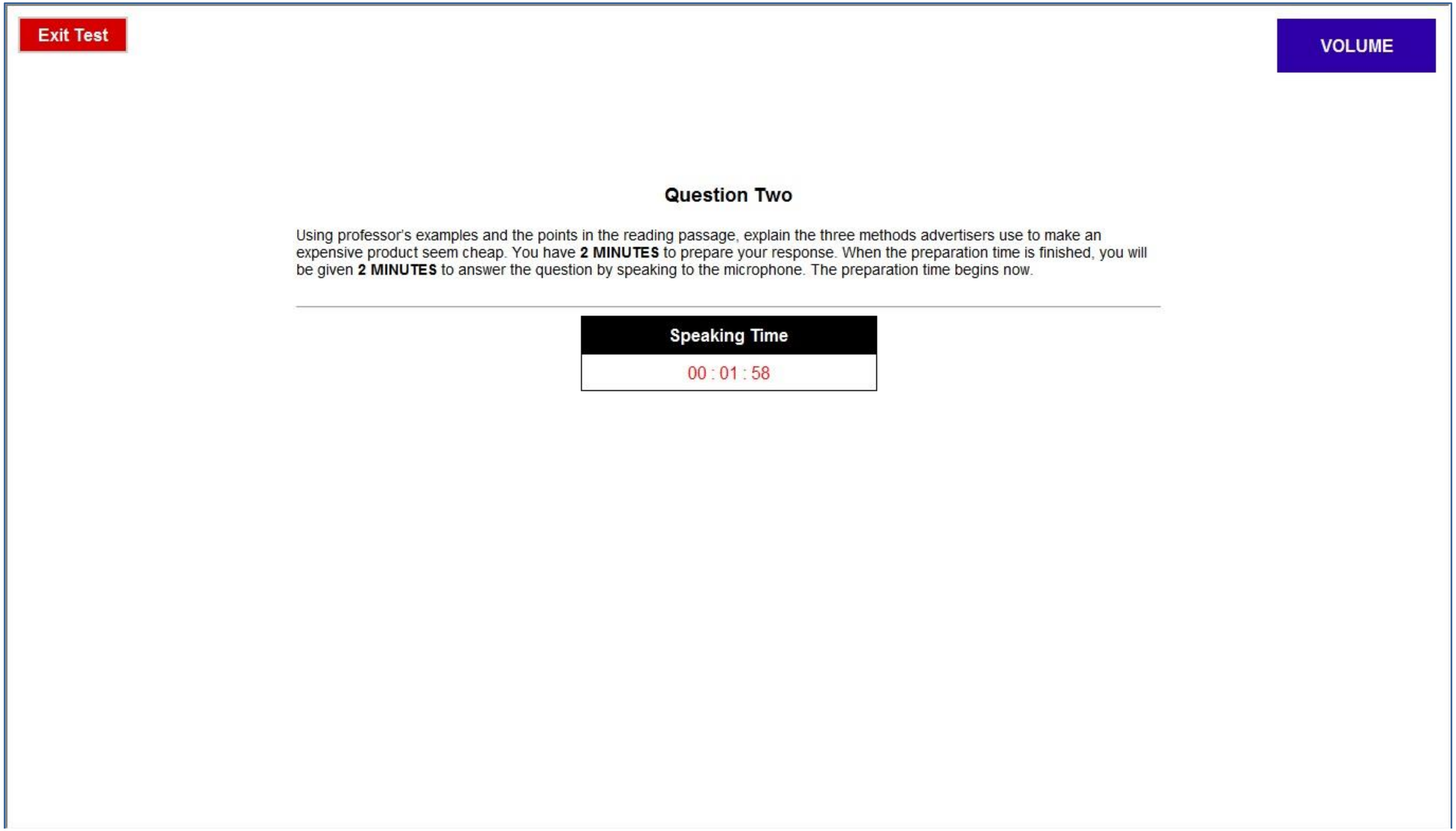
Businesses and companies do not want their products to seem either cheap or expensive. Customers associate low price with lower quality. Likewise, when the price seems too expensive, every customer is bound to think twice. So, advertisement agencies try to shift potential customers' attention away from price and onto product value. They normally use three main techniques.

The first technique is called "Contrast Effect". Using this technique, a similar product is chosen with a higher price and introduced next to the advertised one. This way, potential customers subconsciously come to the conclusion that no matter how expensive the product is, there is a more expensive one out there and as a result worry less about the high price.

The second technique is "Partial Price Quoting." Basically, in advertisement, agencies and companies do not quote the total price which might seem very expensive at first glance. Instead, they divide the price into monthly or even daily payments. Since the focus of attention is the price, not the time period, the price seems to be very low.

The final technique is "normalizing." This method simply convinces the potential customers that the mentioned price is a normal one, even if it really is not, by showing them the large number of people who have bought the product. As a result, paying a high price does not seem abnormal anymore.

Figure 19 – Task Two Question Prompt Page



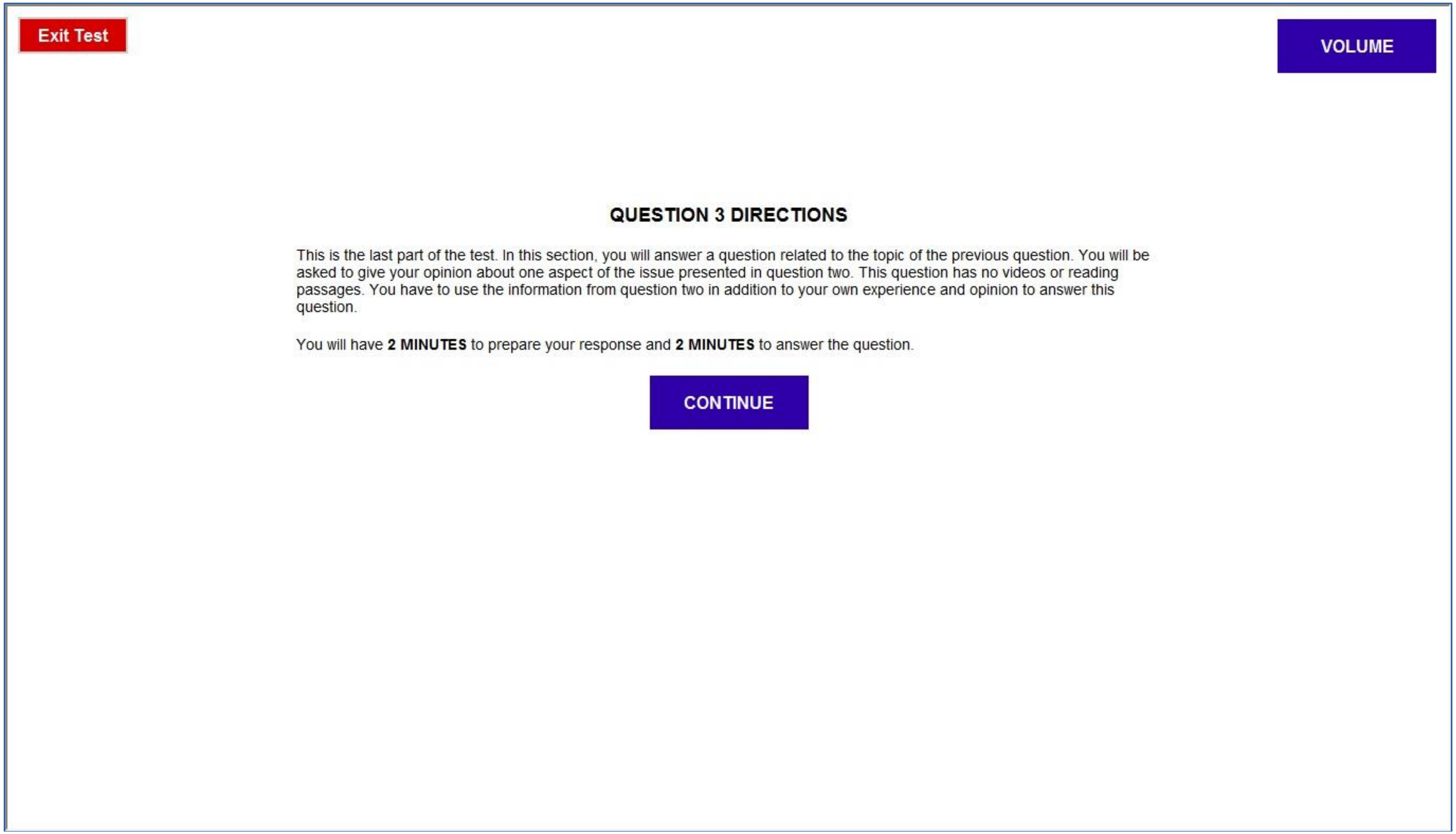
Question Two

Using professor's examples and the points in the reading passage, explain the three methods advertisers use to make an expensive product seem cheap. You have **2 MINUTES** to prepare your response. When the preparation time is finished, you will be given **2 MINUTES** to answer the question by speaking to the microphone. The preparation time begins now.

Speaking Time

00 : 01 : 58

Figure 20 – Directions of Task Three

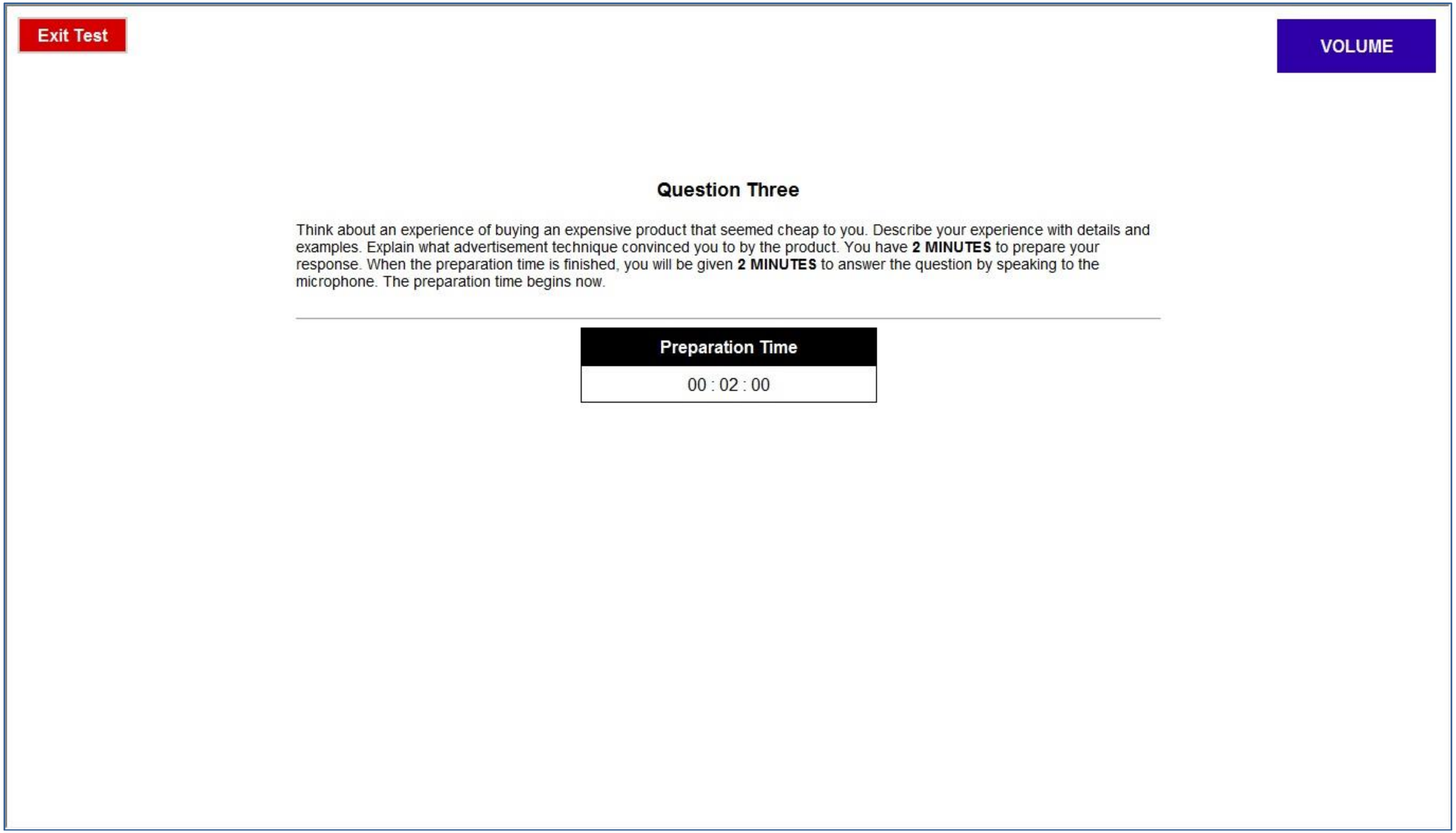


QUESTION 3 DIRECTIONS

This is the last part of the test. In this section, you will answer a question related to the topic of the previous question. You will be asked to give your opinion about one aspect of the issue presented in question two. This question has no videos or reading passages. You have to use the information from question two in addition to your own experience and opinion to answer this question.

You will have **2 MINUTES** to prepare your response and **2 MINUTES** to answer the question.

Figure 21 – Task Three Question Prompt Page



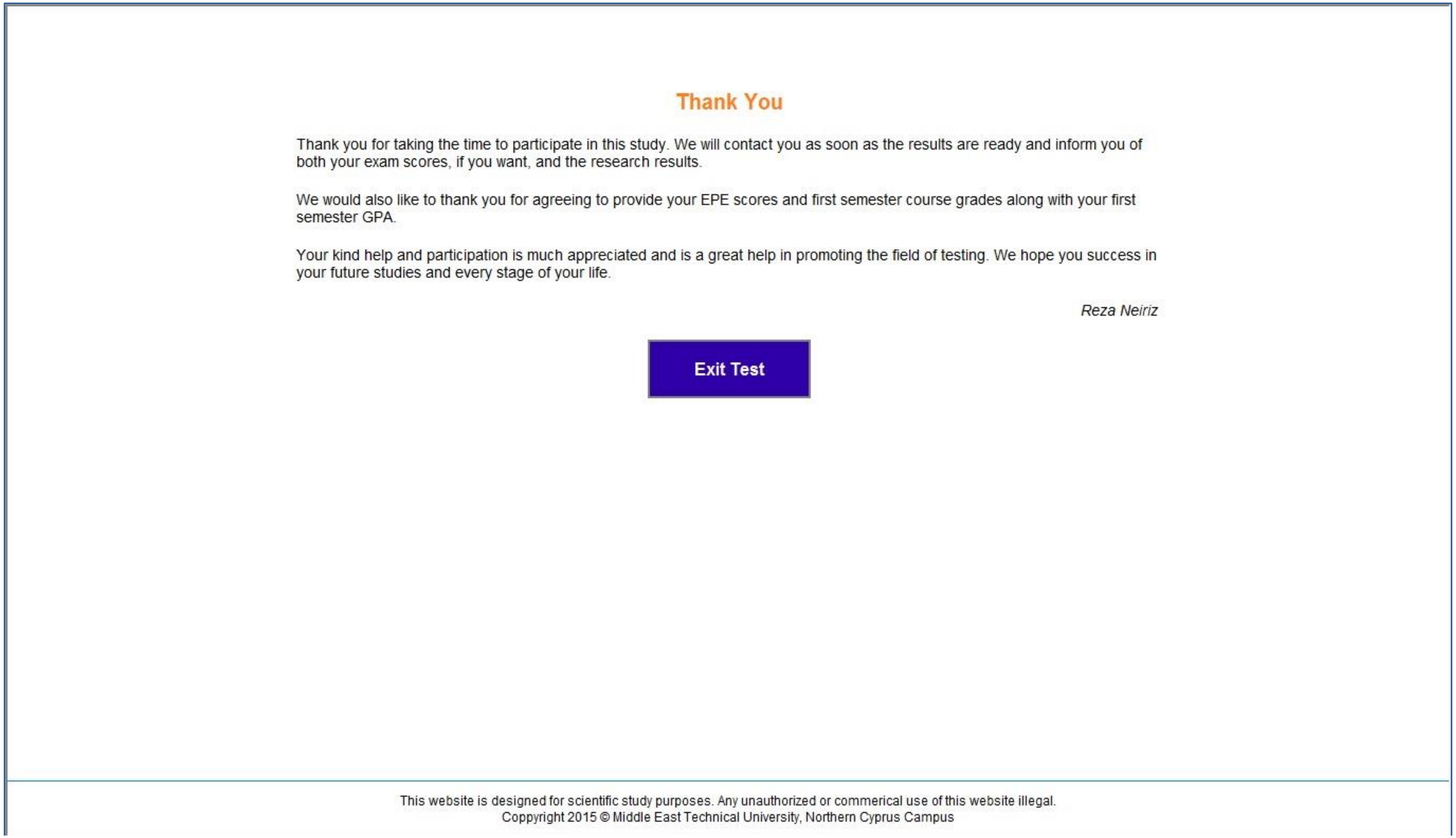
Question Three

Think about an experience of buying an expensive product that seemed cheap to you. Describe your experience with details and examples. Explain what advertisement technique convinced you to by the product. You have **2 MINUTES** to prepare your response. When the preparation time is finished, you will be given **2 MINUTES** to answer the question by speaking to the microphone. The preparation time begins now.

Preparation Time

00 : 02 : 00

Figure 22 – The Final Page of PTBST



Thank You

Thank you for taking the time to participate in this study. We will contact you as soon as the results are ready and inform you of both your exam scores, if you want, and the research results.

We would also like to thank you for agreeing to provide your EPE scores and first semester course grades along with your first semester GPA.

Your kind help and participation is much appreciated and is a great help in promoting the field of testing. We hope you success in your future studies and every stage of your life.

Reza Neiriz

This website is designed for scientific study purposes. Any unauthorized or commerical use of this website illegal.
Coppyright 2015 © Middle East Technical University, Northern Cyprus Campus

G. Correlations Of Early And Late Participants

Correlations are all Pearson Product-Moment. Statistically significant values are indicated by an asterisk.

Correlations Non-Engineering (Early)

	GPA	Weighted GPA
PTBST	.821*	.789*
	.007	.011
	9	9

Correlations Non-Engineering (Later)

	GPA	Weighted GPA
PTBST	.126	-.016
	.747	.966
	9	9

Correlations Engineering (Early)

	GPA	Weighted GPA
PTBST	.582	.556
	.171	.195
	7	7

Correlations Engineering (Later)

	GPA	Weighted GPA
PTBST	.057	-.158
	.915	.765
	6	6

Correlations Total (Early)

	GPA	Weighted GPA
PTBST	.678*	.644*
	.005	.010
	15	15

Correlations Total (Later)

	GPA	Weighted GPA
PTBST	.270	.182
	.312	.500
	16	16