To What Extent Do Engineering Students Master and Retain an Understanding of Newtonian Mechanics Throughout Their University Life

Eman Hameed Abdal-Razzaq

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the Degree of

Master of Science
in
Physics

Eastern Mediterranean University
February 2014
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Physics and Chemistry.

Prof. Dr. Mustafa Halilsoy
Chair, Department of Physics

We 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 Science in Physics.

Prof.Dr. Ayhan Bilsel
Co-superviser

Asst. Prof. Dr. Mehmet Garip
Supervisor

Examining Committee

1. Prof. Dr. Ayhan Bilsel
2. Prof. Dr. Mustafa Halilsoy
3. Assoc. Prof. Dr. S.Habib Mazharimousavi
4. Asst. Prof. Dr. Mehmet Garip
5. Asst. Prof. Dr. Mustafa Rıza

ABSTRACT

This research is to assess the conceptual understanding of towards learning Physics courses for master and undergraduate students among the first year and final year. The study examined engineering undergraduates (N = 272) and master students (N=10) for one year at EMU for 2012/2013 session. This is a descriptive quantitative research. Data was collected by using one instrument, namely the Force Concept Inventory (FCI). The data collected was analyzed by using three software package programs SPSS version 20.0, TAP version 12.9.23 and Stat disk version 12.0.2. The findings show that the mean scores obtained by the students 'master and undergraduates" in FCI was 27.8%. The results indicate that there is no statistically significant difference between correct answered and "year, age, CGPA, and program". This means there are no factors affecting on the correct answers of students in EMU. Also the results show that the Mean score for masters students is (M=30.3%), while the Mean score for undergraduate students is (M=26.6%). However, the results indicate that poor conceptual understanding due to misconceptions is detected among students.

Keywords: Force Concept Inventory, Correct answer, Language, and Item analysis.

ÖZ

Bu araştırma birinci ve dördüncü sınıf mühendislik öğrencilerinin Newton mekaniğinin kavramsal anlayışını ölçmeyi amaçlamaktadır. Çalışma Doğu Akdeniz Üniversitesinde, 2012–13 Bahar döneminde 282 öğrenci üzerinde gerçekleştirilmiştir. Bu çalışma tanımlayıcı nicel bir araştırmadır. Bu çalışmada temel veriler, Hestenes ve arkadaşları tarafından tasarlanan Kuvvet Kavramı Ölçeği (FCI) enstrümanı ile toplandı. Test öğrencilere İngilizce Türkçe, Arapça ve Farsça olarak dört farklı dilde sunuldu. Ayrıca, her öğrencinin bazı kişisel verileri de toplandı. Bu veriler öğrencinin yaşı, akademik yılı, kayıtlı olduğu programı, başlangıç Fizik, Kimya ve Matematik derslerinde aldığı not ve genel not ortalaması (CGPA) gibi bilgilerdir. Toplanan veriler SPSS sürüm 20.0, TAP sürüm 12.9.23 ve Statdisk sürüm 12.0.2 kullanılarak istatistiksel olarak analiz edildi. Bulgular öğrencilerin FCI testindeki genel başarılarını ortalama olarak yüzde 27,8 olarak göstermektedir. Ayrıca verilerin analizi FCI testinde gösterilen başarının katılımcıların ' testte seçtikleri dil, eğitim-öğretim yılı, yaş, genel not ortalaması, fen derslerinde almış oldukları not, sınıf ve kayıtlı oldukları program gibi faktörlerden hiç etkilenmediğini, aralarında istatistiksel olarak anlamlı bir ilişkinin bulunmadığını göstermektedir. Bu sonuç, öğrencilerin test başarılarını etkileyen herhangi bir faktör/parametrenin bulunamadığı anlamına gelir. Literatürdeki benzer çalışmalar ile karşılaştırdığımızda, öğrencilerimizin testteki başarıları genelde daha düşüktür. Test sonuçları örneklenen öğrenci gurubunun Newton mekaniğin kavramsal anlayışının zayıf olduğunu ve öğrencilerin konu hakkında yanlış kanılara sahip olduklarını göstermektedir

Anahtar Kelimeler: Kuvvet Kavramı Ölçeği, Doğru cevap, Dil, ve madde analizi.

DEDICATION

To my family

ACKNOWLEDGMENT

In such moments depends Firefly to think before they write letters collected in the words... Scatter the characters and tries in vain to assemble in Brief…. Many lines passing through the imagination and it remains for us in the end only a little bit of memories and images we gathered comrades were on our side…

The duty we thanked them and waving them goodbye as we move our step in the midst of life and singled the contributor acknowledgments to all of the lit candle in the paths of our knowledge and to from the stop on the platforms and gave the proceeds from the idea to enlighten our path to the professors valued in the Faculty of Art and Science and From more gratitude, it is a pleasure and I am delighted to extend my thanks and gratitude to my supervisors Assistant Prof. Dr. Mehmet Garip, who gave me of his knowledge a lot, and who didn't hesitated days to lend a hand to me in all areas, and thank God that make it on my way and it may be that prolongs age to remain guiding light glistening in the light of science and scientists. It also specifically appreciated and thanks to Dr. Ayhan Bilsel to help me in giving useful observations, which helped me a lot to accomplish and end my thesis.

I would also like deepest gratitude to my professors respected members of the Committee for reading my thesis and enrich their proposals beneficial. In the end, I am glad to extend my deepest gratitude to all of a helping hand extended to me in my career Scientific.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

Itemize

EMU	Eastern Mediterranean University
FCI	Force Concept Inventory
PER	Physics Education Research

Chapter 1

INTRODUCTION

1.1 Introduction

A Concept Inventory is a type of test in a given subject that tries to measure students' conceptual understanding of that subject. Specifically the Force Concept Inventory (FCI) that was used in this present study is a tool for assessing conceptual understanding of Newtonian mechanics. This tool has played a significant role in changing attitudes and methods in the teaching of freshman physics courses. [1]

This research is to study the conceptual understanding of Physics among the first year and final year Physics Science undergraduates (N = 272) from EMU for 2012/2013 session also for master students (N=10) among one year. This is a descriptive quantitative research. Our data is collected by using one instrument, namely the Force Concept Inventory (FCI).

The Collected data are analyzed by using three programs SPSS version 20.0, TAP version 12.9.23 and Statdisc version 12.0.2 Results show that the mean scores obtained by the total of master and undergraduates students' in the FCI - test was 27.8%. However, poor conceptual understanding due to misconceptions is detected among them (M = 27.8%, SD = 3.850).

1.2 Objectives of the study

The objectives of this study were:-

1. To answer the thesis question that is “to what extents do engineering students master and retain an understanding of Newtonian mechanic throughout their university life”.

2. To determine the level of conceptual understanding in Newtonian force concept among the engineering student masters and undergraduates.

3. To determine if there is any correlation between the FCI test score and parameters such as (CGPA, grade obtained in introductory science courses such as Physics 1, General Chemistry and Calculus 1).

4. To determine if there is any significant difference between the FCI test score of students when grouped according to:

a) Test-language

b) Registered program

c) Academic year (freshman, sophomore, junior or senior)

d) Student age

1.3 Research Methods

This is a descriptive quantitative analysis based on using a multiple choice test as the instrument to collect the information.

1.4 Instruments

The instrument utilized in this research was the Force Concept Inventory developed by Hestenes and Swachamer. This test tries to measure the extent the students become "Newtonian thinkers" after official education in Classical mechanics course. [2]. To do this, they designed a multiple-choice test. Although in the beginning they started with 29 questions, subsequently (and to this date) the number of questions became 30. For each question there is only one correct answer while there are four alternatives based on most frequently held misconceptions. In the present study, the Force Concept Inventory test was administered to a group of (mostly engineering) students (N = 282).

Table 1. Details of the 282 participants in the present study

Program	**Number of students**				**Master**	**Total**
	Y1	Y2	Y3	Y4	Y5	
EEE	12	7	0	3	0	22
ME	13	18	4	10	3	48
CE	53	55	6	53	7	174
IE	1	0	0	2	0	3
Other	20	10	1	4	0	35
Total	99	90	11	72	10	282

EEE-Electric and Electronic Engineering

ME- Mechanic Engineering

CE-Civil Engineering

IE-Industry Engineering

OTHERS- there are a few students belonging to other such as Information System Engineering, Mathematics etc. Because their numbers are low, they have been included in the category of others.

Y1- Freshman, 1st year; Y2- Sophomore, 2nd year; Y3-Junior, 3rd year; Y4-Senior, 4th year and Y5-Masters.

1.5 Research Procedure

The Force Concept Inventory test in four different languages, namely English, Turkish, Arabic, and Persian, were downloaded from < http://modelinginstruction.org/researchers/evaluation-instruments/fci-and-mbt/> and the password to access the files was obtained from FCIMBT@verizone.net. The test was administered in 10 separate classes to a total of 282 students over a period of two weeks. The students were given the option of choosing from Turkish, English; Arabic and Persian language versions. The respondents were given 30 minutes to answer the Force Concept Inventory.

Table 2 . Distribution of respondents by choice of test Language

Test-language	**Number of students**
English students	131
Turkish students	113
Arabic students	32
Persian students	10

1.6 Data Evaluation

The data collected from the research were analyzed using descriptive and inferential statistical analysis techniques such as Student's t-test, Pearson correlations, ANOVA, Linear Regression, Item analysis, and Kolmogorov–Smirnov was used. Below is a list of the tests used and the information they provided:

a) **Item analysis test** provides

- Test scores for individual respondents.
- Item difficulty of each question.
- Item discrimination index value of each question.

- Option analysis giving information about misconception types by showing response patterns of respondents.

b) T-test- for significance testing between the means of FCI scores and gender.

c) Pearson correlation- to test for relation language between FCI score and parameter such as program, science course grade, and year.

d) Regression linear- to model the relationship between correct answer in FCI score and parameter such as language, age, year, program and CGPA.

e) ANOVA- it like T-test for significance testing between the means of FCI scores belonging to different group’s age, year, program, CGPA.

f) Kolmogorov–Smirnov - utilized to decide if a sample comes from a hypothesized continuous distribution.

Chapter 2

LITERATURE REVIEW

In this chapter, we will present a survey of the literature on how the FCI was developed, how it has been applied to science and engineering students and how it has influenced the teaching of Classical mechanics. We shall also survey recent attempts and devising similar instruments in other disciplines such as Biology, Chemistry. Also, we shall explore specific issues like Cumulative GPA and language. We are going to later examine these important parameters in our analysis.

2.1 The Force Concept Inventory

Currently, the FCI is the most frequently used instrument for the purpose of assessing students' conceptual understanding of Newtonian mechanics [3]. What this thirty item test has effectively shown is that although students may be able to solve typical quantitative problems, they fail to show any understanding of the relevant concepts contained in these questions [4].

The FCI test is designed to measure understanding in six different areas, called dimensions, of the Newtonian force concept (see table 3 below) [2]. Groups of questions in the test measure each specific dimension. Each question has been designed to test only one concept without requiring any calculations. For each item, there are five choices. Only one of these choices is correct. The remaining four choices are distractors which have been selected from commonly held misconceptions. In the beginning, many physics instructors considered the items in the FCI to be trivial or easy. However, when they applied the test to their students, they found that their students lacked basic understanding of the concepts. In fact from the beginning, FCI test results were showing that even students who completed a semester of introductory Physics courses were only managing a success rate of sixty three to seventy seven percent.

Table 3. Newtonian Concept in the Inventory [2]

	Inventory Item
0.Kinematics	
Velocity discriminated from position	20E
Acceleration discriminated from velocity	21D
Constant acceleration entails	
Parabolic orbit	23D,24E
Changing speed	25B
Vector addition of velocities	7E
I. First Law	
With no force	4B,6B,10B
Velocity direction constant	26B
Speed constant	8A,27A
With cancelling forces	18B,28C
2. Second Law	
Impulsive forces	6B,7B
Constant force implies	
Constant acceleration	24E,25B
3.Third Law	
For impulsive force	2E,11E
For continuous forces	13A,14A
4. Superposition Principle	
Vector sum	19B
Cancelling force	9D,18B,28C
5. Kinds or force	
5S.Solid contact	
Passive	9D,(12B,D)
Impulsive	15C
Friction opposes motion	29C
5F. Fluid contact	
Air resistance	22D
Buoyant (air pressure)	12D
5G. Gravitation	5D,9D,(12B,D),17C,18B,22D
Acceleration independent of weight	1C,3A
Parabolic trajectory	16B,23D

2.1.1 Review of the FCI

Huffman and Heller [5] made the first review on the FCI, and they looked at the validity of dividing the test in to six dimensions. They conducted factor analysis of the data presented by Hestenes et al, and concluded that the students didn't poses a mental perception of force in the six dimensions. They also considered FCI to be unsuitable or ineffective at measuring student understanding. The reply to this criticism from Hestenes et al. was that they agreed with the author's conclusion that the students didn't think about force within the six dimensions precisely because they were not Newtonian thinkers! But they argued that the FCI results were valid and the test was able to assess the difference between "Newtonian" and student perception.

This discrepancy has remained unresolved and still causes divisions in how the FCI results are interpreted. It is clear that there will always be disagreements among educators as to the effectiveness of assessing conceptual understanding by using such inventories.

Another criticism of the FCI is its format as a multiple-choice test. By design, the FCI was aimed to minimize false-positives; that is to prevent a non-Newtonian thinker to select answers like a Newtonian-thinker and vice versa.

Hestenes and Halloun considered an 85% score in FCI as the threshold level for mastery in Newtonian mechanics and 60% as the threshold level for entry to Newtonian physics" [6]. By mastery they mean the individual is a Newtonian thinker and by entry level they mean the individual is beginning to think like a Newtonian thinker.

Rebello and Zollman [7] wanted to assess the effect of the distractors. They administered the FCI test to a group of students by removing all the choices and simply presenting it as an open ended question set. They then compared responses of the students to the open-ended questions with those from the multiple choices FCI test. They found that the incorrect solutions to the open-ended test did not correlate well with distractors in the multiple-choice FCI.

2.1.2 Impact of the Force Concept inventory

Three distinct uses for the FCI test have been proposed by its developers [2]. One use is as an aid to instructors to check which concepts have not been understood by students or which misconceptions prevail. Another use is for placing student’s in appropriate sections/groups for instruction. However, Hestenes et al., warns that since the FCI does not test how well a student copes with calculations in physics, he suggests that an additional mathematics test be also administered in order to make a better decision on placement. The third use suggested is for assessing how effective is the instruction in teaching students to become Newtonian thinkers. This can best be achieved by giving the test as a pretest in the beginning of the semester and as a posttest given at the end. It is argued that comparison of pre and posttest results provide the evidence if there has been any changes in the conceptual understanding of the student’s because of the instruction. Out of these three uses it is this last one that has had the biggest effect on physics instruction.

As soon as first results of the FCI started to appear it began to show how ineffective the traditional way of teaching physics by lectures was [2]. Many instructors were finding that their students scored a lot lower than what they expected. They rationalized this by assuming that their students’ experience in physics was minimal or non-existent. However, score results showed that prior physics experience had no effect on pretest scores either. Therefore the conclusion had to be that the traditional way of teaching physics had no effect on the post test results. The tiny variation between pre and post test scores was a shock to many educators.

Hake [8], provide a summary of the FCI results collected for 6542 students taking introductory physics courses from 62 different university, college and high school [9]. Using this data Hake compared the test scores for students receiving traditional instruction (passive learning) with those involved in classes where there was engagement and interaction among students and instructors (active learning). He defined *relative gain* as;

$$\boldsymbol{relative\ gain} = \frac{\boldsymbol{posttest\ score - pretest\ score}}{\boldsymbol{100 - pretest\ score}}$$

Then he calculated the class average of students' relative gain for each course and he used these averages to assess the efficacy of teaching. The average relative gain for those courses that were interactive and engaging, were two standard deviations higher than that for traditional lecture-based courses. Interestingly enough, Hakes' results correlated well with the Mechanics Baseline test [2]. This is a test which aims to assess "problem solving ability" as opposed to "conceptual understanding". To this day, instructors in many institutions continue to utilize the Force Concept Inventory for the purpose of studying and assessing their own methods of teaching.

2.2 Concept Inventory Development

In addition to the lack of consistency in concept inventory development, there are specific concerns about using these assessments as indicators of student understanding. First, there are complications introduced by having distractors as multiple choice answer alternatives. Distractors are included with the intention of determining whether students have overcome common sense misconceptions, indicating a true understanding of the correct physics explanation. Because of this, distractors must be carefully composed, reflecting typical student misunderstandings, if the question is to be an accurate reflection of a student's grasp on physics concepts.

Dean Zollman and Sanjay Rebello explored the alignment of responses on force concept inventory problems and equal open-ended problems with a sample student population of non-majors who generally had some physics background at Kansas State University. After administering the FCI to one randomly chosen group and the same questions in an open-ended format to another randomly chosen group, the open-ended answers were sorted based on naturally occurring categories in the responses. Comparing these answers, it is apparent that misconceptions presented in the multiple choice format differ from the misconceptions that appear in the open-ended format [7]. While there is only one right answer, there are many possible wrong answers. It seems that the distractors in the FCI do not necessarily reflect the misconceptions of the students. Therefore, conclusions about student misunderstanding based on the FCI distractors may not be accurate. This is a fundamental limitation of all multiple choice assessments.

Furthermore, researchers presented revised multiple choice questions in which the misconceptions resulting from the open-ended questions replaced irrelevant FCI distractors. Upon comparing the number of students who chose the original FCI distractors verses the revised distractors, the latter tended to dominate. Thus, it can again be concluded that, "an analysis of the incorrect responses to FCI problems cannot be an effective way to set which parts of the students' conceptual understanding are incomplete" [7]. Furthermore, a warning is included in the final discussion of this study cautioning that distractors are transient; misconceptions change as the students learn physics jargon and confuse content throughout the semester [7].

Considering these two results, it is clear that not all distractors are useful in identifying misconceptions; in fact, most are not accurate. Yet, distractors are still included in concept inventories to fulfill their originally intended purpose: differentiate between a student's true understanding of physics concepts and some common prevailing misconception.

In order to develop a concept inventory, it is first necessary to make a list of the concepts intended for learning by students. But to understand how a student perceives a concept, it is essential to interact directly with the student. This can be achieved through surveys, focus groups or by face-to-face interviews

A concept inventory is not an unchanging and static set of items. An inventory that has be developed goes through cycles of being administered, and analyzed, and on the basis of these analyses, the items may be revised, removed or new items added in to the inventory.

2.3 Other Concept Inventories

Because of the success of the FCI, educators and researchers from other disciplines, such as the sciences and engineering fields, are developing concept inventories for their own areas. Examples include the Foundation Coalition involved in various engineering subjects [10]; in chemistry [11], dynamics, electricity and magnetism [12], fluid mechanics [13].

One example is the Materials Concept Inventory (MCI). The developers of this inventory state that the “overall goal is to analytically link relationships of scientific fundamentals to macroscopic materials behavior” [11]. The items are from the topics of atomic structure and bonding, band structure, crystal geometry, defects, microstructure, and phase diagrams for metals, ceramics, polymers and semiconductors. The MCI contains 30 questions, with ten based on previous knowledge of chemistry and geometry, and 20 based on content from a materials course.

Another example is the Statics Concept Inventory. This inventory has been designed with the aim to “detect errors associated to incorrect concepts, not with other skills (e.g., mathematical) necessary for Statics”[14]. In designing this inventory, developers prepared items that required very simple calculations such that an incorrect answer would as a result incorrect assumption and conception of the subject matter and not because of any calculation errors.

2.4 Cumulative GPA and Grade Predictive Schemes

As an alternative, a student’s incoming cumulative GPA has been found to be a predictive factor for student success. Scott Freeman et al. At the University of Washington, Seattle, developed prediction schemes for grades in introductory biology courses based on incoming GPA and SAT scores[17]. Like Freeman et al., Lai shows that, in addition to gender, a student’s cumulative GPA at the beginning of the class greatly predicts their performance. In fact, a student’s cumulative GPA is the most correlated parameter to the final grade received in an introductory physics course. Because of this, cumulative GPA can be used to develop grade prediction schemes for the University of EMU which accurately predict what the student will receive in three courses (PHYS101, MATH151, and CHEM101). Examining 31 terms of data (Winter 1996 through Winter 2008 from the data set we will use for our research), Lai finds that "a student's physics grade tends to be lower than their cumulative GPA" [18]. Furthermore, a gender gap can also be seen in introductory physics courses: "… the average grades of females are consistently lower than the average grades of males" [18]. Based on this high correlation between cumulative GPA and student performance, Lai developed grade prediction schemes by course and by gender. Plotting course grades vs. incoming cumulative GPA and fitting a quadratic (see Figure 1); equations that predict course grades were developed for each introductory physics course at the University of Michigan.

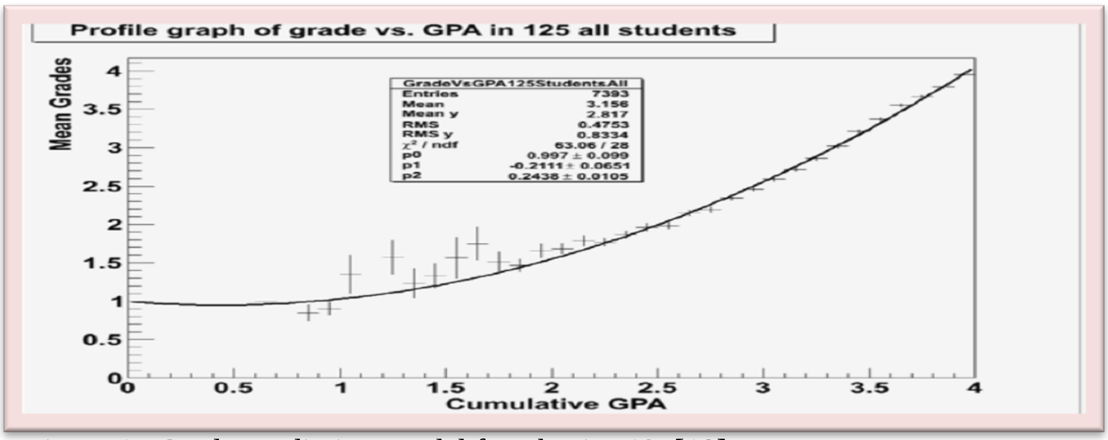


Figure 1. Grade Prediction Model for Physics 125[18]

These predictions are taken to form a basis for further comparisons as they are highly accurate and largely independent of the instructor. We will use these schemes as a point of reference in our analysis. Just as concept inventories use gain scores to account for differences in initial levels of understanding, grade prediction schemes allow us to determine if a student does better than, remains the same as, or does worse than expected as we vary a parameter.

2.5 Difficulties with particular representations: Language

In prior studies indicate the language, either spoken or written, is another way of representing physics concepts or situations, Lemke has studied patterns of language particular to the physics classroom, and how sharing or failing to share these patterns leads to productive or unproductive discussion[18].

Some linguistic work has been done in PER, as well. Williams [19] notes that many of the words that we use to represent physics concepts (force, speed, work) also represent common-language concepts that are much less precisely defined. Thus, students and teachers can be using the same words to represent may different ideas. Brookes [20], investigated the role of language in learning physics in much more detail, often in the context of quantum mechanics. He interviewed a number of students and faculty regarding such topics in quantum mechanics as the infinite square well and the Bohr-model, in addition to studying the textbook language. He found that much of language use, both expert and novice, takes the form of metaphor and/or analogy. Brookes identified a number of specific metaphorical ideas used (such as "the potential well step is a physical object"), and noted that much of the difference he observed in success could be attributed to correct or incorrect applications of metaphors. Students have a strong tendency to construct overly literal metaphors, treating potential steps as physical steps, or thinking of a particle as a truly solid and localized object. The physicists studied were capable of applying literal interpretations of the language when appropriate, and ignoring these interpretations otherwise. In short, experts were aware of the limitations of the linguistic representations that they were using, while students were not.

Chapter 3

METHODS

This chapter discusses the various methods that were adopted for this study. These include analysis of the all data and also the effectiveness of supplementary study groups at EMU. First, we summarize our data set, together with a discussion of the content and structure of the courses studied as well as an overview of internal and external parameters that have been compiled. After, we describe five key analytical tools that will be used in our analysis: The T- Test, Pearson correlation, Item analysis, ANOVA, and Linear Regressions.

Next, we begin to explore important factors, as suggested by the literature, and their effects on all our data. We focus on the parameters we have in our data: number of correct answers, Cumulative GPA, language, science course grade, year, age, and gender of the respondents.

3.1 Data

The University offers three main introductory courses: PHYS101, MATH151 and CHEM101. The PHYS101 is an algebra-based Classical Mechanics course whereas CHEM101 is general chemistry and MATH151 is Calculus I.

3.2 Analysis tools

We will now provide a brief overview of fundamental analytical techniques. This section will serve as a reference for the tools we use in our analysis.

3.2.1 T-test

The t-test is a statistical hypothesis test for the equality of the means of normally distributed data. The distribution of the means for small samples follows the Student's t distribution. Thus the test is used to see if two means are significantly different from each other [21]. This is done by calculating the *t*-statistic:

$$t = \frac{\overline{x_1^2} - \overline{x_2^2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where

$\overline{x_1^2} - mean\ of\ sample\ 1$

$\overline{x_2^2} - mean\ of\ sample\ 2$

$n_1 - number\ of\ subjects\ in\ sample\ 1$

$n_2 - number\ of\ subjects\ in\ sample\ 2$

$s_1^2 - variance\ of\ sample\ 1 = \frac{\sum(x_1 - \bar{x}_1)^2}{n_1}$

$s_2^2 - variance\ of\ sample\ 1 = \frac{\sum(x_2 - \bar{x}_2)^2}{n_2}$

3.2.2 Pearson’s correlation

This test is used to measure the strength of a linear association with Pearson’s correlation coefficient, r, between two variables. A value of 1 for r indicates perfect positive correlation and a value of -1 means perfect negative correlation. The coefficient measures the degree of linear relationship between two variables. The Fisher r-to-t test is used to measure the statistical significance of Pearson’s r value. [22].

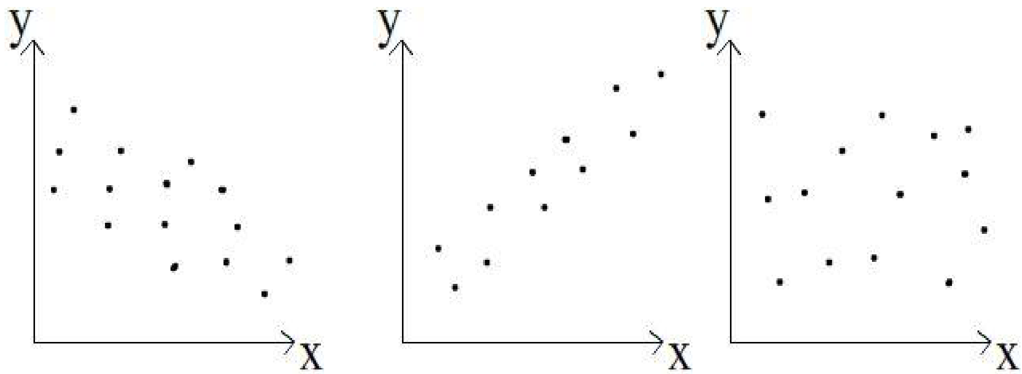
We can categorize the kind of correlation by considering the behavior of the “dependent” variable as the other (independent) variable increases:

 Positive correlation – the “dependent” variable tends to increase;

 Negative correlation – the “dependent” variable tends to decrease;

 No correlation – the “dependent” variable neither increase nor decrease.

Visually, we can best observe the relationship by plotting the data as a scatter plot. The three plots below exemplify negative, positive and no correlation [22].



Negative correlation Positive correlation No correlation

Mathematically, the correlation coefficient can be calculated using the equation:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{[\sum x^2 - \frac{\sum (x)^2}{n}][\sum y^2 - \frac{\sum (y)^2}{n}}}$$

n- Number of pairs of scores

$$\sum xy - Sum\ of\ the\ products\ of\ paried\ scores$$

$$\sum x - Sum\ of\ x\ scores$$

$$\sum y - Sum\ of\ y\ scores$$

$$\sum x^2 - Sum\ of squared\ \ x\ scores$$

$$\sum y^2 - Sum\ of squared\ \ y\ scores$$

3.3 Item Analysis

Item analysis is a method that checks student responses to individual test items (problems) so as to evaluate the quality of these items and of the test as a full. Item analysis is very valuable because it enables us to check the difficulty of the items and the discriminating ability of each question, and in this way it helps us to decide which items to eliminate because they may be unclear or misleading. Additionally, item analysis is effective for increasing instructors' skills in test construction, and identifying specific areas after all content that require greater affirmation or clarity. Item analysis contains Item Difficulty, Item Discrimination, Difficulty and Discrimination Distributions, and Reliability Coefficient [23].

3.3.1 Item Difficulty

Item difficulty is the percentage of students who answered a test item correctly. This means that low item difficulty value (e.g., 28 %) indicate difficult items, since only (28 %) small percentage of students got the item correct. Conversely, high item difficulty values (e.g., 84) indicate easier items, as a greater percentage of students got the item correct [23].

3.3.2 Discrimination Index

Item discrimination measures how well a particular question/item discriminates between high scoring and low scoring students. A high value for the discrimination index means that a bigger proportion of the high scoring students are answering the item correctly than the low scoring students. [23].

Another parameter that can be used as a discrimination index for items is the Point-Biserial. It’s value can vary between -1.00 to 1.00. A strong and positive correlation suggests that high scoring students are able to answer the particular item correctly. This is to be expected since students who know the content and are scoring well on the test generally should also be doing well on individual items. However, there's a problem if students are answering a test correctly even though they do not know the content. This situation is discovered by low or negative

3.3 One way of ANOVA

The one-way analysis of variance (ANOVA) is used to test if there is any significant difference between the means of three or more independent samples. For example, you may use a one-way ANOVA to understand whether eye color of students have any effect on the mean exam score. To do this the students are grouped according to eye color and the mean exam scores are compared by ANOVA to see if they are (statistically) different from each other or not. This test however simply says if they are the same or not, but it can't indicate which mean is larger/smaller than others [24].

3.4 Linear Regression

This is similar to Pearson's Correlation Coefficient, in the sense that it considers the relation between two variables. It tries to answer if the changes in one variable are (linearly) related to the changes in the other variable. It is also possible to consider more than one independent variable affecting a dependent variable. In this case the method is called Multiple Linear regression. Linear regression models employ the least squares technique in which the deviations of the data from the model are minimized [25].

3.5 Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test (K–S test) is a non-parametric test used for testing the distribution of one dimensional continuous data. It can be used as a one sample test or as a two sample test. In a one sample test an actual distribution is tested against a reference distribution. In a two sample test two sets of distributions are compared against each other. The K–S statistic measures distance between the distribution functions. The null hypothesis is set up to state that the two samples are drawn from the same distribution in the case of two-sample test, and the sample is drawn from the reference distribution in the case of one-sample test [26].

Chapter 4

ANALYSIS

The research question asks "To what extent do engineering students master and retain an understanding of Newtonian mechanic throughout their university life". In order to answer this question, we used the FCI test in three languages, namely English, Turkish and Arabic, because for the majority of our students' English is not their native or mother tongue. We administered the FCI to the undergraduate students of the Faculty of Engineering at the Eastern Mediterranean University (N=272), and 10 Masters students. Our objective was to assess the conceptual understanding of Newtonian force concept amongst undergraduates. Therefore the results for the Master students are briefly mentioned at the end of this chapter.

In our investigation, we first looked at test scores (rate of correct answers) in the FCI-Test; second, the effect of test language on score; third, relation between science course grade with the test scores and fourth we considered the responses of test takers to the individual questions (item analysis).

4.1 Study of Correct answer in FCI-Test at EMU

We want to explore the test scores of the students in the FCI-Test along with their various attributes, and see if there are any relationships between these variables. The respondents were divided in to groups according to the following six categories and their scores tested or correlated. First grouping was "test language"; second was “respondent age”; third was the “academic year”, fourth was the respondents “degree program”, fifth with “cumulative grade point-CGPA” and finally with gender.

To begin with we will first consider the test scores in general. Result show that the mean score is 7.98 with a standard deviation of 3.53. The maximum score was 22 while the minimum was 1” in FCI-Test. The histogram of the test scores is given in figure 4.2 below.

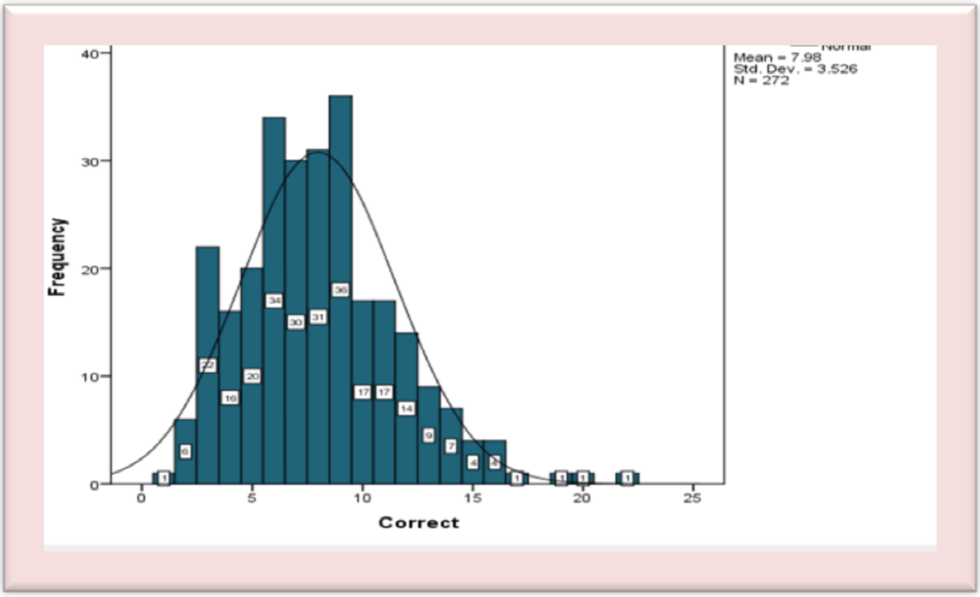


Figure 2. Histogram of the correct answers given by students in FCI-Test

It is worth noticing from the histogram that the bars closely follow the normal distribution line. All the data do fall inside the bounds of natural variability. To see if the test score of the respondents are normally distributed, we used the Kolmogorov-Smirnov test, results of which are given in table 4.1 below. Since significance is 0.004, which is less than 0.05, we conclude that our test scores are normally distributed.

Table 4 . Kolmogorov-Smirnov test showing that test scores are normally distributed

		Correct answer
Normal Parameters [a,b]	Mean	7.98
	Std. Deviation	3.526
Most Extreme Differences	Absolute	0.107
	Positive	0.107
	Negative	-0.053
Kolmogorov-Smirnov Z		1.764
Asymp. Sig. (2-tailed)		0.004

4.1.1 Effect of Test Language on FCI-Test scores:

To study the effect of test language, we plotted individual test scores against test language (1 is English, 2 is Turkish and 3 Arabic) as shown in figure 4.2.

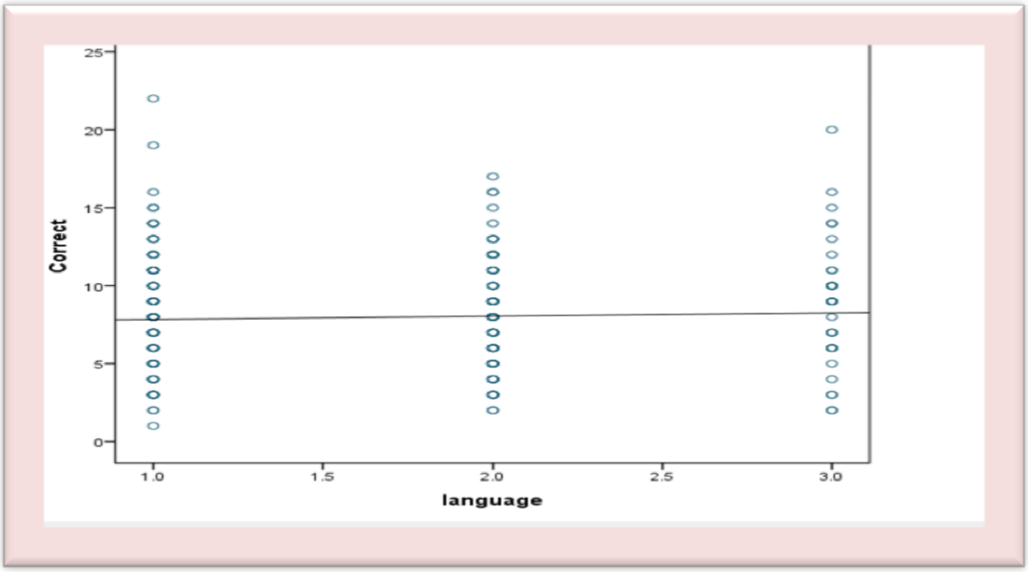


Figure 3 .The relationship between language and correct answer in FCI-Test

Table 5 . Shows the Test Score sample statistics grouped by test language

Language	Mean Score	Std. Dev.	Percent Score	Maximum Score	N
English	8.397	3.828	28.00%	22	120
Turkish	8.451	3.668	28.20%	17	113
Arabic	6.969	3.836	23.20%	20	32

From table 5, we note that the mean score for English and Turkish are very close to each other. Therefore we calculate the 95% confidence interval to see if we can accept or reject the hypothesis that there is no difference in the mean score for English, Turkish, and Arabic language tests. At the 95% confidence level, the difference between each pair of means include zero, therefore at 95% level, we cannot reject the null hypothesis that $\mu_E = \mu_T$; $\mu_E = \mu_A$; $\mu_T = \mu_A$. This means that at the 95 % level there is, statistically, no significant difference in the achievement of students solving the FCI in different languages.

In figures 4A, 4B and 4C we plotted the total number of students answering an item correctly against item number. This plot shows us the answering pattern of students doing the test in different languages and also allows us to compare if there are noticeable differences in the answering pattern. In figure 4A we see that among the students answering Turkish-FCI, the number of students correctly answering questions 5, 13, 24, 25 and 30 are very low in comparison with the remaining questions. In figure 4B for English-FCI, we see that the numbers of students answering questions 20, 25, 26, and 30 correctly are low, and in figure 4C, the number of students answering correctly the questions 19, 20, 25, and 26 in the Arabic-FCI is also very low. We will discuss this issue in more detail in the section “4.4 Questions”.

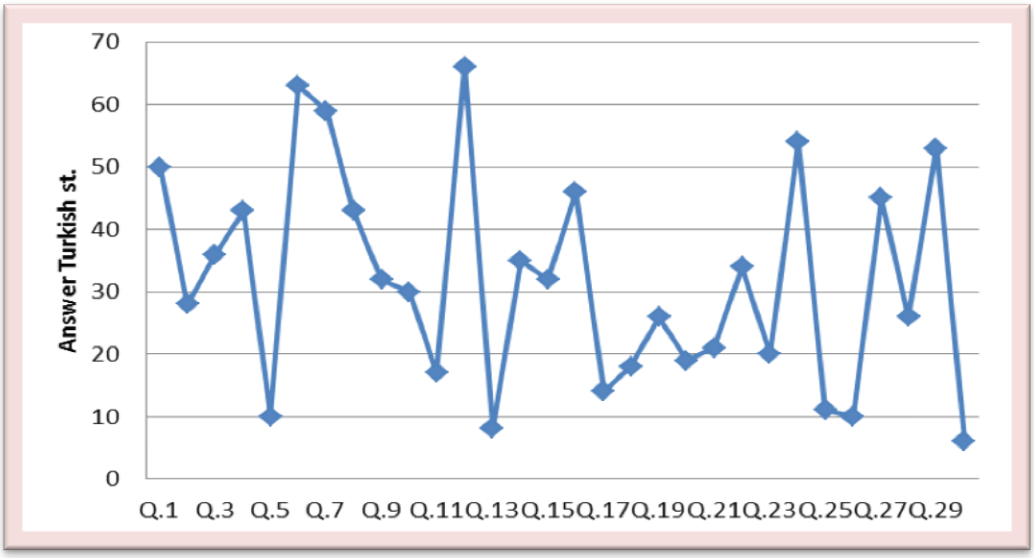


Figure 4 . A Number of correct answers to Items in Turkish test.

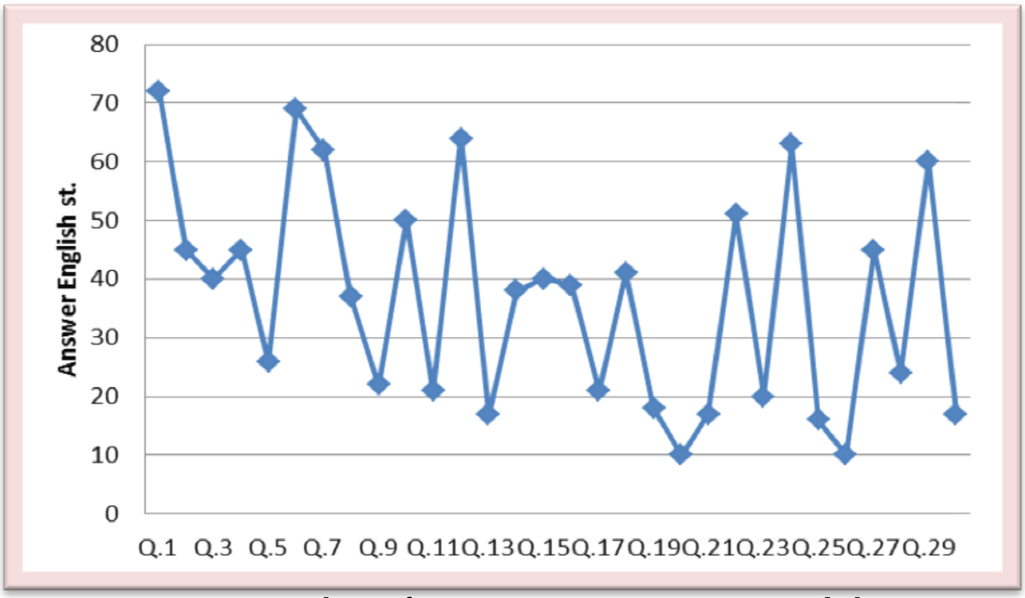


Figure 4 . B Numbers of correct answers to Items in English test.

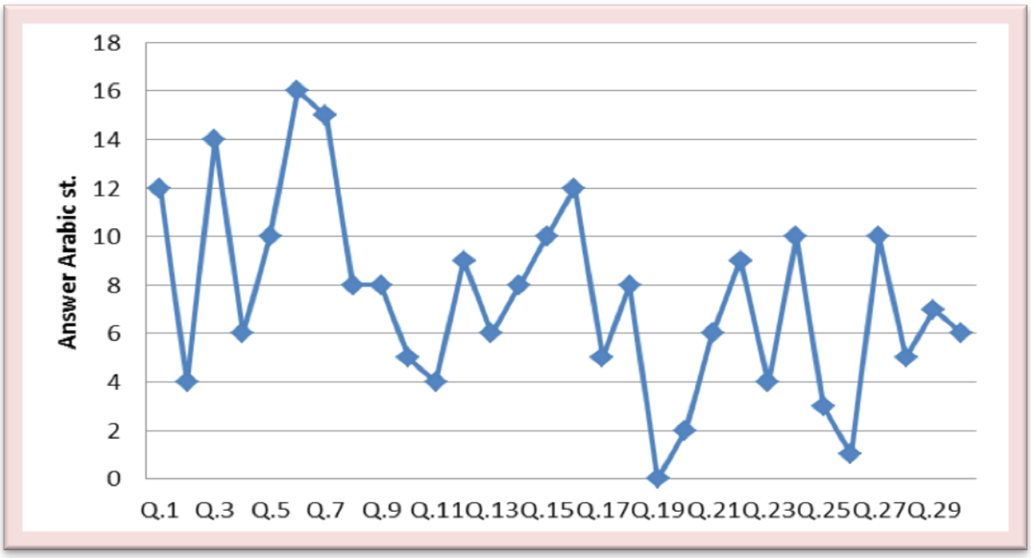


Figure 4 . C Number of correct answers to Items in Arabic test.

4.1.2 The relationship between Age and Score

To study the effect of student age on test score, we divided the respondents into three age groups called 1, 2 and 3. Those born during 1995 to 1998 were coded as Age = 1; those born during 1991 to 1994 coded as Age = 2, and those born before 1991 coded as Age = 3 (see appendix B - for all the codes used in SPSS). Table 6 shows the statistics for Age. Note that the respondents in Age 2 group constitute the largest sample size with N=138.

Table 6.Shows the sample statistics for Age

Age	**Mean**	**Std. Dev.**	**N**
1. 1995-1998	7.43	3.214	28
2. 1994-1991	8.43	3.87	138
3. Before 1991	7.54	3.056	106
Total	7.98	3.526	272

One way analysis of variance test results for the scores for each age group is given in table 7.

Table 7.Shows ANOVA test results for scores in each age group.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	57.789	2	28.894	2.347	***0.098***
Within Groups	3311.12	269	12.309		
Total	3368.91	271			

The test result is significant at 0.098. Since this value is greater than 0.05, we conclude that statistically there is no significant difference between the test score means for the three age groups at the 0.05 level. Also we plotted individual test scores in each age group in figure 8 and drew the best-fit line through the data. The flatness of the least squares fit line indicates that there is no correlation between age and test score in this instance.

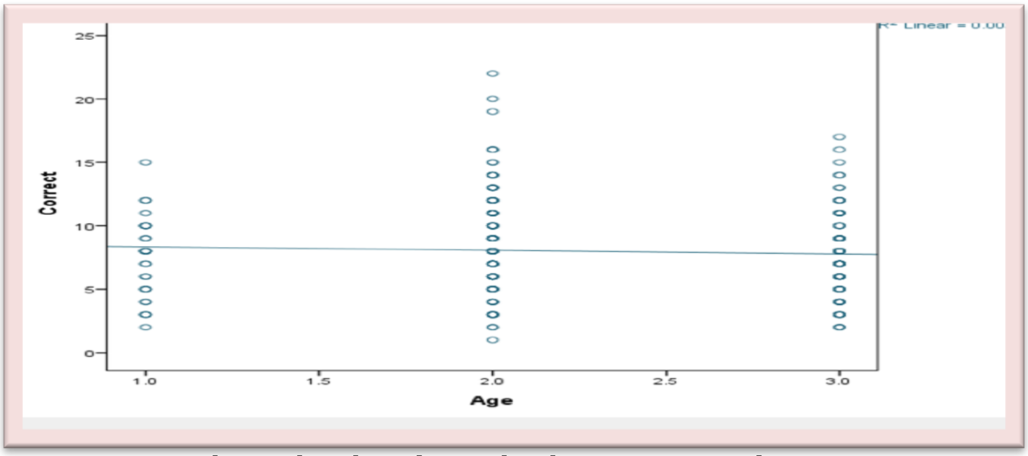


Figure 8. Shows that the relationship between age and correct answer

4.1.3 The relationship between YEAR and test score

Next, we wanted to see if there was any difference in the test scores of respondents who were in different academic years. The students fall in to one of four years, namely year 1, 2, 3 or 4. We carried out an ANOVA test whose outcome is given in table 9.

Table 9. Shows that ANOVA methods between YEAR and score

	Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Between Groups	71.947	3	23.982	1.949	0.122
Within Groups	3296.96	268	12.302		
Total	3368.91	271			

Also, statistics for test score bye year are shown in table 10. The ANOVA test results above show that the significance of the test is 0.122. Since this value is much greater than 0.05, we accept that there is no significant difference between the mean score for each YEAR.

Table 10.Shows that sample statistics for year

Year	**Mean**	**Std. Dev.**	**N**
1	8.11	3.583	99
2	8.31	3.594	90
3	6.73	2.649	11
4	7.58	3.463	72
Total	7.98	3.526	272

Also the linear regression plot of correct answers against YEAR is shown in figure 5. The correlation coefficient, R^2, indicates that there does not appear to be any relationship between the YEAR in which a student is in and his score in FCI test.

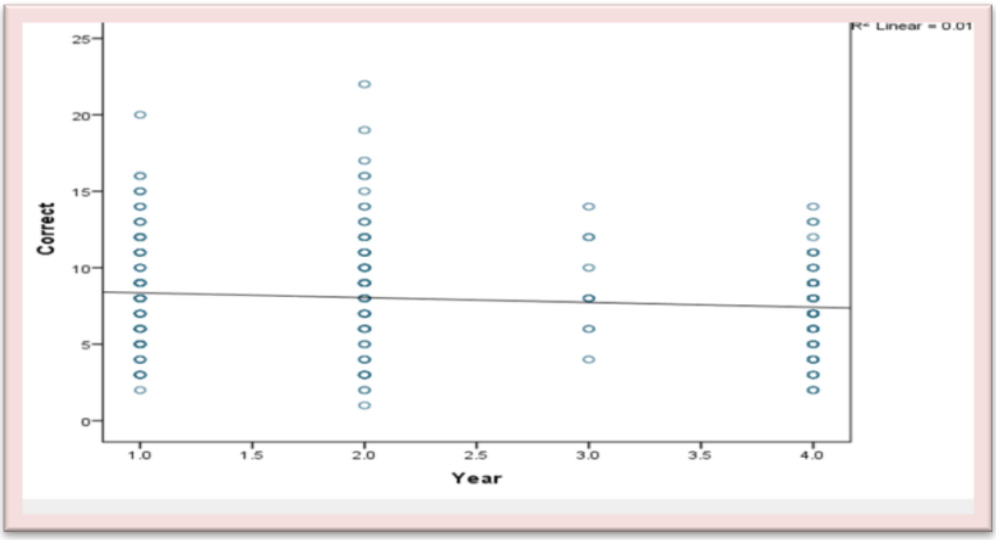


Figure 5.Shows that the relationship between year and correct answer

4.1.4 The relationship between Program and correct answer

To seek the relationship between the program and correct answer, we carried out the ANOVA test on the test scores grouped by respondents program whose results are in table 11. The sample statistics for test scores by program are given in table 12. In table 11, we see that the ANOVA test significance is 0.543, which again means that there is no significant difference between the students' registered program and their FCI score.

Table 11.Shows the ANOVA method between program and the correct answer

	Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Between Groups	74.698	7	10.671	0.855	0.543
Within Groups	3294.21	264	12.478		
Total	3368.91	271			

In order to confirm that there is no relationship between students' program and their score, we also conducted a linear regression of the data as shown in figure 6.

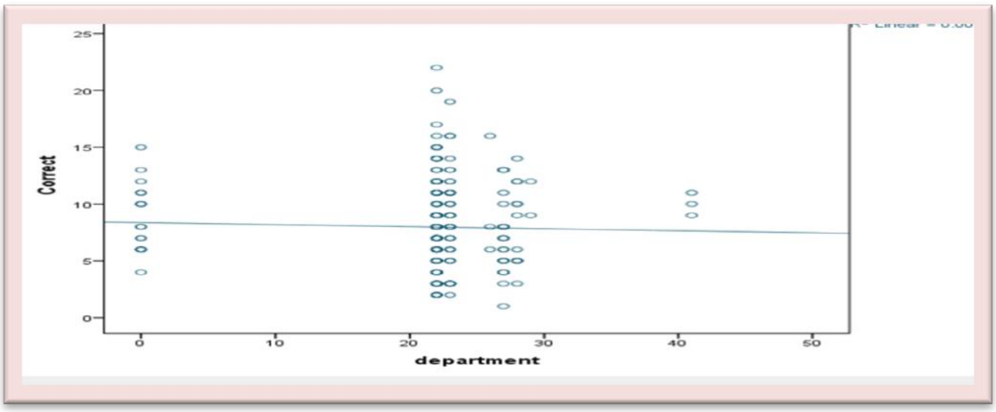


Figure 6. Shows the relationship between program and correct answer

Table 12.Shows sample statistics for program and correct answer

Program	Mean	N	Std. Deviation
CE	7.86	167	3.473
ME	8.30	44	4.044
EEE	7.00	23	3.261
IE	10.0	3	5.292

4.1.5 The relationship between "CGPA" and FCI score

Also, we probed CGPA and score by plotting individual FCI scores against CGPA of each student in figure 7 and calculated the linear regression between them

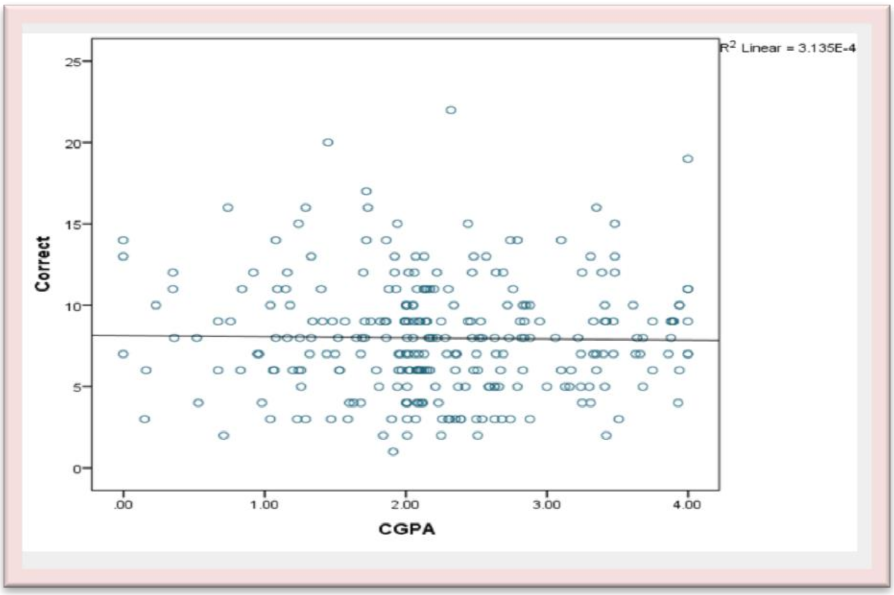


Figure 7.Shows the relationship between CGPA and correct answer

The linear regression line in figure 7 show, rather surprisingly, that there is no discernable relationship or correlation between CGPA and FCI test score ($R^2=3\times10^{-4}$).

4.1.6 The relationship between gender and FCI score

Finally, we tested whether there is any significant difference between gender and test score and for this we conducted a Student's t-test between male and female groups, by calculating the 95 % confidence interval for the difference between the mean test scores for male and female students, as shown in table 14.

Table 13.Shows T-Test between gender and correct answer

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Correct	Equal variances assumed	.066	.797	-.822	270	.412	-.538	.655	-1.828	.752
	Equal variances not assumed			-.866	42.809	.391	-.538	.622	-1.792	.716

Both confidence intervals, calculated by assuming equal and non-equal variances, include the value zero. Therefore we conclude that no significant difference between them exists. Table 15 below, gives the sample statistics

Table 14. Shows sample statistics between gender and test score

Gender		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Correct	male	239	7.92	3.555	.230
	female	33	8.45	3.317	.577

4.2 Evaluation of Courses at EMU

Among the objectives of this study was to see if there is any significant difference between mean course grade in the introductory science courses of PHYS101, MATH151 and CHEM101 whose summarized data are given in table15.

Table 15.Shows the sample statistics for courses

		PHYS101	**MATH151**	**CHEM101**
N	Valid	264	264	181
	Missing	8	8	91
Mean		1.6883	2.0716	1.9017
Std. Deviation		1.33376	1.32604	1.19905

When we compare these means, we find that at the 95% confidence level the mean for PHYS101 is different, in fact lower than both MATH151 and CHEM101.

Also, in order to see if the course grades are distributed normally we plotted the letter grade histogram for each course in figures 8, 9 and 10. The x-axis shows the corresponding numerical value of each letter grade. Note that the data seem to agree with the line representing a normal distribution except at the extremities where they are much greater than would be expected for normally distributed data. (The Kolmogorov-Smirnov normality test data are given in appendix B).

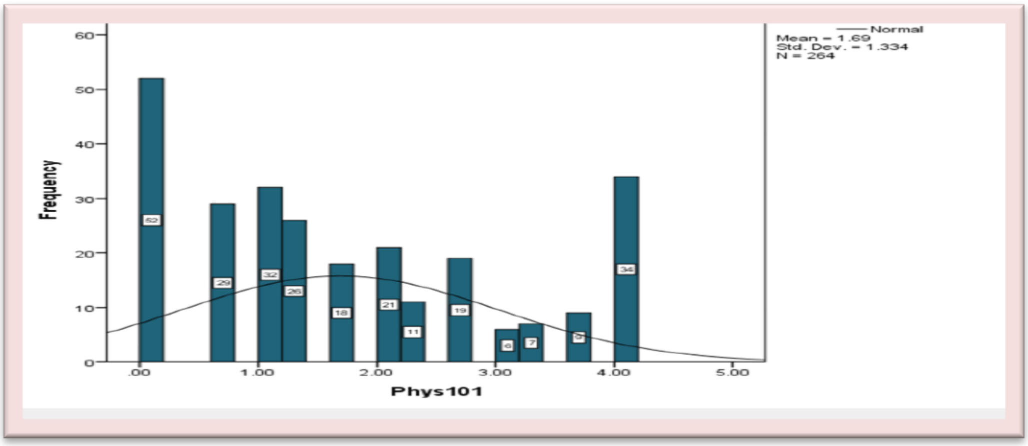


Figure 8. shows normality assessment for PHYS101

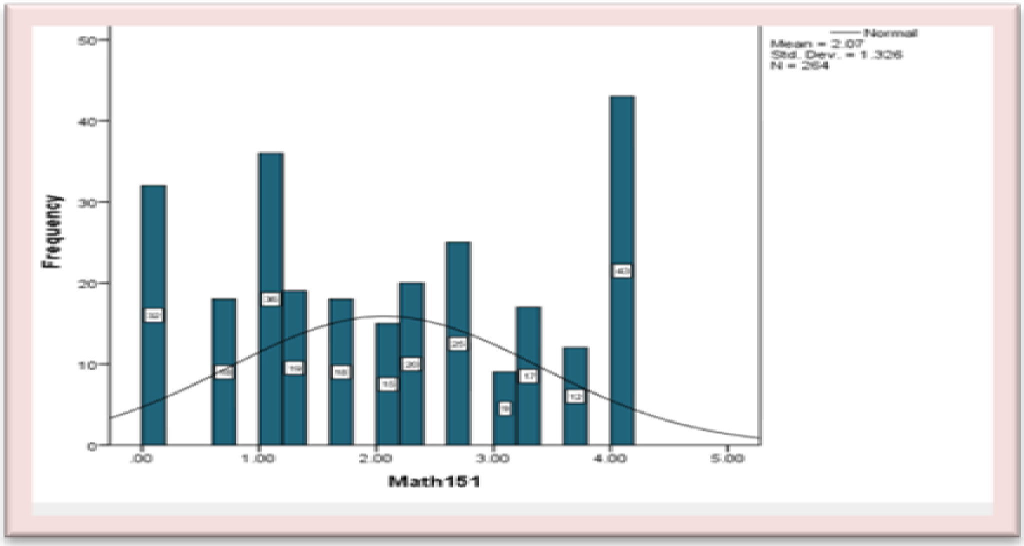


Figure 9. shows normality assessment for MATH151

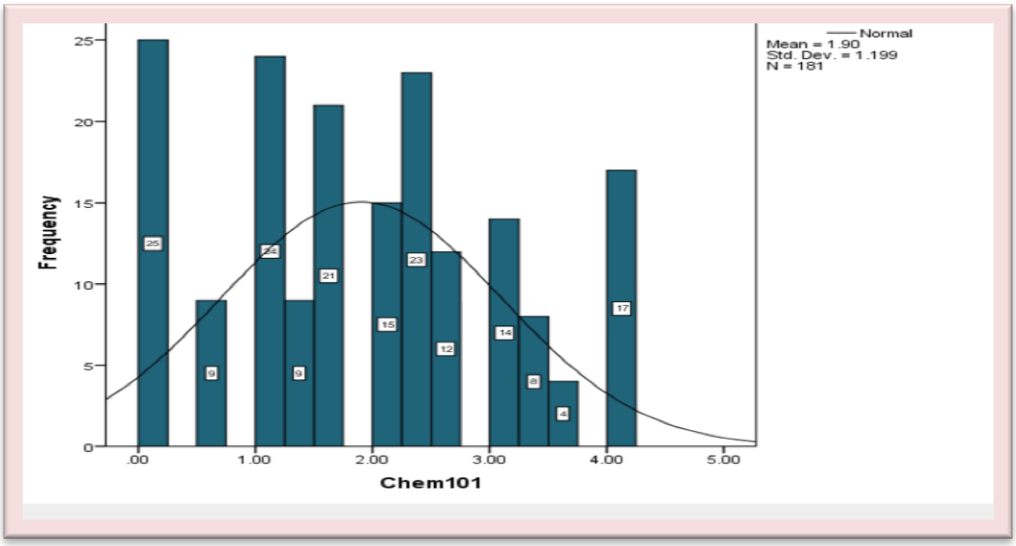


Figure 10. shows normality assessment for CHEM101

4.2.1 The relationship between Course Grade and Year

Summary of the course grades by year for each of the science courses PHYS101, CHEM101 and MATH151are given in Table 16.

Table 16. Course grade averages for each year, for the science course.

Year		**Phys101**	**Math151**	**Chem101**
first year	Mean	1.6747	2.1242	2.1056
	N	95	95	54
	Std. Deviation	1.51615	1.47141	1.39466
second year	Mean	1.4080	1.9091	1.5426
	N	88	88	54
	Std. Deviation	1.20833	1.28740	1.22098
third year	Mean	2.4300	1.9100	1.8714
	N	10	10	7
	Std. Deviation	1.08531	1.43639	1.31240
last year	Mean	1.9493	2.2254	2.0318
	N	71	71	66
	Std. Deviation	1.17642	1.14427	.92838
Total	Mean	1.6883	2.0716	1.9017
	N	264	264	181
	Std. Deviation	1.33376	1.32604	1.19905

To see if there is any significant difference between course grades of a course in different years, we conducted an ANOVA test. The ANOVA results for PHYS101 in Table17 shows that not all the course grade means are the same for each year. In other words the mean grades for each year differ.

Table 17. ANOVA of grades in each year for PHYS101

PHYS101	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17.271	3	5.757	3.322	.020
Within Groups	450.582	260	1.733		
Total	467.854	263			

The ANOVA results for MATH151 given in table 18 show that that there is no difference between the mean grades for each year. This also implies that there is no correlation between course grade and year.

Table 18. ANOVA of grades in each year for MATH151

MATH151	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.527	3	1.509	.857	.464
Within Groups	457.930	260	1.761		
Total	462.457	263			

Similarly, when we conduct ANOVA for CHEM101 whose outcome is given in table 19, we again find that there is no difference at the 5 % significance level between the yearly grade averages for CHEM101.

Table 19. ANOVA of grades in each year for CHEM101

CHEM101	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10.332	3	3.444	2.453	.065
Within Groups	248.458	177	1.404		
Total	258.790	180			

4.2.2 The relationship between Language and science course performance

We also looked at the success profile of the respondents in the three science courses based on their choice of test language. Table 20 shows the numbers of students in each language category who have obtained a particular grade in PHYS101 at EMU.

Interesting points in table 20 are

- Highest percentage of students (31.2 %) getting A and A- in the three language groups are those who chose the Arabic FCI test!
- The largest percentage of students (42.5 %) receiving failing grades (D− and F) in the three language groups are those who chose the English FCI test.

Table 20.Shows the student evaluation of PHYS101 by language

		English N= 117	**Percent**	**Turkish N=116**	**Percent**	**Arabic N=31**	**Percent**	**Total N=264**	**Percent**
A,A-	Count	13	10.84	20	17.7	10	***31.20***	43	15.80
	Expected	9.55		8		2.55		34	
B+,B	Count	5	4.17	7	6.19	1	3.125	13	4.78
	Expected	2.9		2.85		0.75		7	
B-,C+	Count	13	10.84	16	14.16	1	3.125	19	11.1
	Expected	6.65		6.55		1.75		15	
C,C-	Count	14	11.7	22	19.5	3	9.4%	21	14.4
	Expected	8.65		8.55		2.3		19.5	
D+,D	Count	21	17.5	31	27.5	6	18.75	58	21.32
	Expected	12.85		12.75		3.45		29	
D-,F	Count	51	***42.5***	20	17.7	10	31.25	81	29.8
	Expected	17.95		17.75		4.75		40.5	
Total	Count	117		116		31		264	

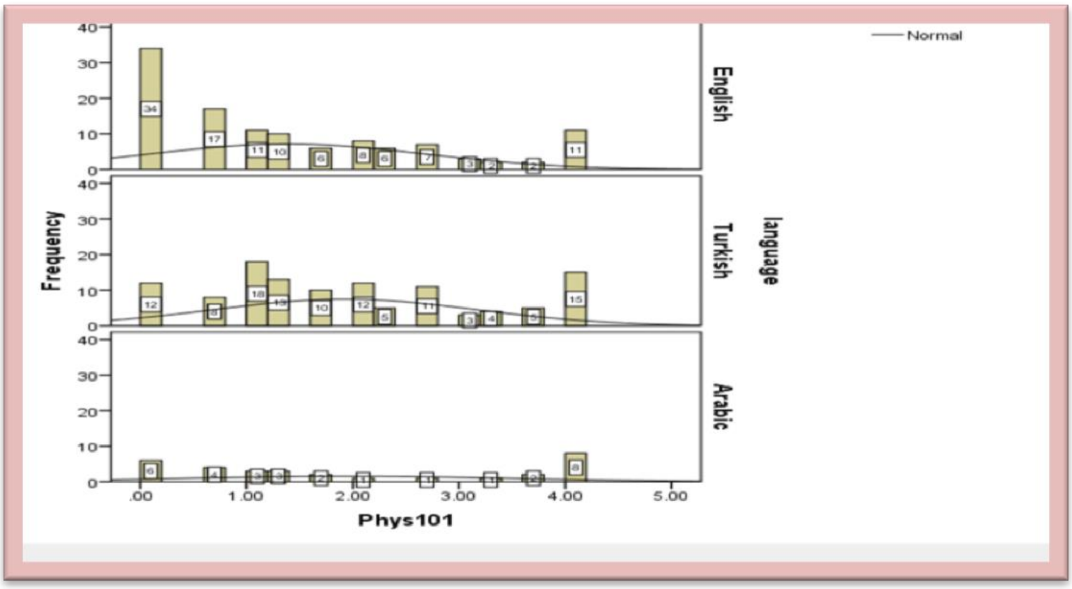


Figure 11. shows the distribution for PHYS101 according to language

Similarly, table 21 shows the numbers of students in each language category who have obtained a particular grade in MATH101. Again we see that highest percentage of A and A− is from the Arab language group and the highest percentage of D− and F is from the English language group.

Table 21. shows the student evaluation of MATH151 according to language

		English N=117	**Percent**	**Turkish N=116**	**Percent**	**Arabic N=31**	**Percent**	**Total n=264**	**Percent**
A,A-	Count	19	15.8	26	23	10	***31.25***	55	20.22
	Expected	12.2		12.1		3.2		27.5	
B+,B	Count	13	10.53	11	9.74	2	6.25	26	9.55
	Expected	5.75		5.75		1.55		13.0	
B-,C+	Count	18	15	24	21.3	3	9.4	45	16.6
	Expected	10		11.0	9.9	2.6		22.5	
C,C-	Count	14	11.7	15	13.3	4	12.5	33	12.2
	Expected	7.3		7.25		1.95		16.5	
D+,D	Count	23	19.2	26	23	6	18.75	55	20.2
	Expected	12.2		12.05		3.2		27.5	
D-,F	Count	30	***25***	14	12.4	6	18.75	50	18.4
	Expected	11.1		11		2.95		25	
Total	Count	117	97.5	116	102.7	31	96.9	264	97.1

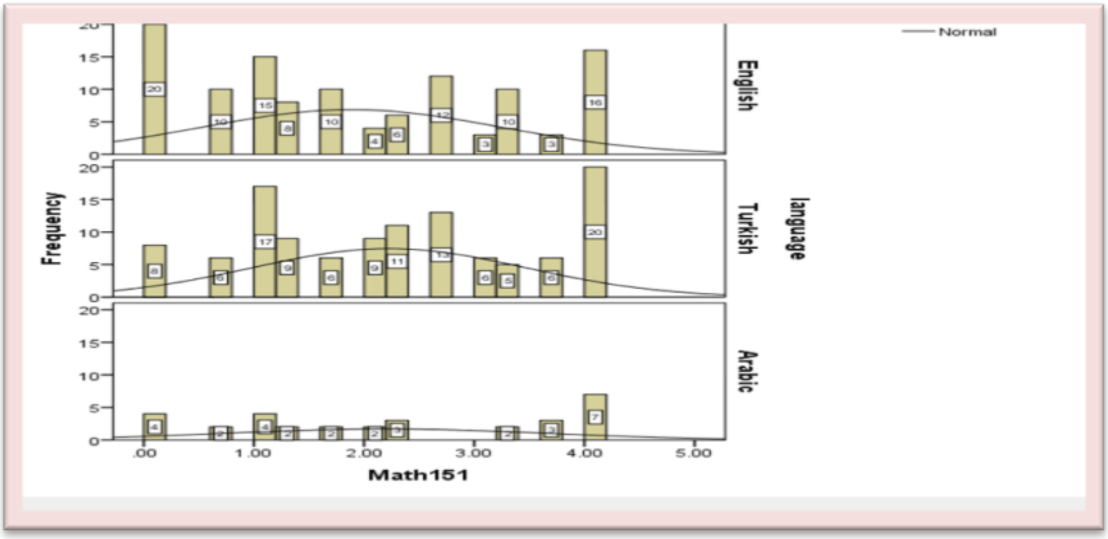


Figure 12. shows the distribution for MATH151 according to language

Finally, the data for CHEM101 is given in table 22. Again we observe a similar trend that biggest proportion of students getting A and A− are Arab language test-takers and highest proportion of students receiving D− and F are English language test-takers.

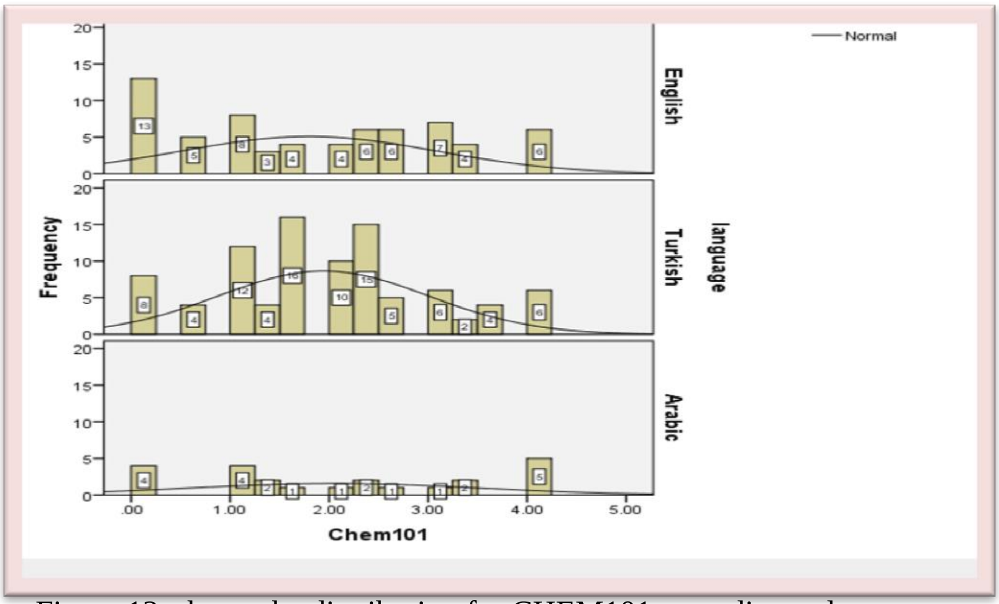


Figure 12. shows the distribution for CHEM101 according to language

Table 22. shows the student evaluation of CHEM101 according to language

		English N=66	Percent	Turkish N=116	Percent	Arabic N=23	Percent	Total N=181	Percent
A,A-	**Count**	6	5	10	8.85	5	***15.6***	21	7.7
	Expected	3.85		503		1.35		10.5	
B+,B	**Count**	11	9.2	8	7.1	3	9.4	22	8.1
	Expected	4		5.6		1.4		11	
B-,C+	**Count**	12	10	20	17.7	3	9.4	35	12.9
	Expected	6.4		8.9		2.2		17.5	
C,C-	**Count**	8	6.7	26	23.1	2	6.25	36	13.3
	Expected	6.6		9.15		2.3		18	
D+,D	**Count**	11	9.2	16	14.2	6	18.8	33	12.2
	Expected	6.05		8.4		2.05		16.5	
D-,F	**Count**	18	***15***	12	10.6	4	12.5	34	12.5
	Expected	6.2		3.9		2.15		17	
Total	**Count**	66		92		23		181	
	Expected	66		92		23		181	

4.3 Evaluation of performance by dimensions in the FCI-Test

In order to explore the individual questions in the FCI - test and performance by test language we carried out Item analysis. In table 29 we list for each item and for each test language, the number of correct answers, and total number of responses, the item difficulty and the discrimination index of the item.

In addition to the classification by Hestenes, 1992 for all the questions in FCI-Test (see table 23) below, we used the classification of representational coherence in grouping the items in the FCI. Although many items in the inventory are written in the same context, they nevertheless separate in to differing categories of the representations and dimensions of the concept of force.

Table 23. shows the classification of FCI questions in terms of dimensions and representations of FCI [2]

Kinematics	**Newton's first law**		**Newton's second law**	**Newton's third law**	**Kinds of force** Gravitation	Contact
Diagram	Verbal	Diagram	Verbal	Verbal	Verbal	Verbal
12, 14, 19, 20	10,17,24,25	6,7,8,23	22,26,27	4,15,16,28	1,2,3,13	5,11,18,29,30

Next, we shall probe each of these dimensions for the Turkish and English language groups because their sample size is large. In Table 24, we list the percent score for the Kinematics dimension for the two language groups and for comparison we also give the test-score % by those respondents who have a CGPA corresponding to “A, A-, B+, B, third those get fail grade “D, D-, F”.

Table 24. shows correct answer % of Kinematics- Diagram (12, 14, 19, and 20)

Kinematics – Diagram (12,14,19,20)			
Language	FCI-Score %	A,A-,B+,B Correct %	D,D-,F Correct %
Turkish students N=113	32.3%	39.1%	69.7%
English students N=131	24.8%	43.3%	77.4%

We can observe that the correct answers by those answering Turkish FCI is greater than those answering English FCI, with a difference of 7.5 points between their mean score percentages. In the “Newton’s first law- Verbal” dimension group (items 10, 17, 24 and 25), Turkish language test-takers are again scoring higher than English ones with a difference of 5.1 points in their mean percentage scores (see table 25 below). But, in the “Newton’s first law- Diagram” dimension (items 6, 7, 8 and 23), we see English language test-takers scored higher than Turkish ones with a difference of 4.5 points in their mean percentage score.

Table 25. shows the correct answered % of Newton's first law- Verbal and Diagram

Newton’s first law – Verbal (10,17,24,25)			
Language	FCI Score %	A,A-,B+,B Correct %	D,D-,F Correct %
Turkish students N=113	40.9	39.1	69.7
English students N=131	35.8	43.3	77.4

Newton’s first law – Diagram (6,7,8,23)			
Language	FCI Score %	A,A-,B+,B Correct %	D,D-,F Correct %
Turkish students N=113	24.1	39.1	69.7
English students N=131	28.6	43.3	77.4

In table 26, data for the third dimension group, “Newton’s second law- Verbal”, shows that English test-takers score higher than Turkish language test-takers, with a difference of 4.8 points.

Table 26. shows the correct answered % for Newton's second law- Verbal

Newton’s second law – Verbal (22,26,27)			
Language	FCI Score %	A,A-,B+,B Correct %	D,D-,F Correct %
Turkish students N=113	22.1%	52.2%	92.9%
English students N=131	26.9%	57.7%	103.3%

From the fourth dimension group “Newton’s third law – Verbal (items 4, 15, 16 and 28), which is given in table 27, we observe that Turkish language test-takers scored higher than English ones with a difference of 3.3 points.

Table 27. shows of correct answered % of Newton's third law- verbal

Newton’s third law – Verbal (4,15,16,28)			
Language	FCI Score %	A,A-,B+,B Correct %	D,D-,F Correct %
Turkish students N=113	32.5%	39.1%	69.7%
English students N=131	29.2%	43.3%	77.4%

The final dimension group “Kinds of force” is subdivided into two categories; (i Gravitation) and (ii Contact forces). The results for these two subgroups are given in table 28. We see that English test-takers have a higher score than Turkish test-takers. In gravitation the difference between English and Turkish percentage means is 8.4 points for Gravity-verbal and 5.1 points for Contact-verbal.

Table 28. shows the correct answered % for Kinds of force

Kinds of force – Gravitation Verbal (1,2,3,13)			
Language	FCI Score %	A,A-,B+,B	D,D-,F
Turkish students N=113	23%	39.1%	69.7%
English students N=131	33.2%	43.3%	77.4%
Kinds of first – Contact Verbal (5,11,18,29,30)			
Language	FCI Score %	A,A-,B+,B	D,D-,F
		Correct %	Correct %
Turkish students N=113	21.5%	31.3%	55.7%
English students N=131	26.6%	34.7%	61.9%

Because the number of correct answers for the items 5, 11, 13, 25, 26 and 30 are the lowest, we compared them with the other questions.

In question 5, we notice that Item difficulty for Turkish test-takers is 0.09 indicating that very few respondents – 10 in fact – have answered it correctly even though a total of 110 people responded. Only three students left this item unanswered. Although only 10 Arab language test-takers answered this correctly, the Item difficulty for this group was 0.31, which is considerably higher than those for the Turkish language. Since 26 students out of the 128 responding students in the English FCI got it correct, the item difficulty for this group is 0.2.

In question 11, we observe that the item difficulty for Turkish, English and Arabic students is close, with item difficulty values of 0.15, 0.13 and 0.13, respectively. Also in item 13, we observe that even though all the Turkish students responded, only 8 answered it correctly. This item was also found difficult by English and Arabic students as well with item difficulty values of 0.13 and 0.19, respectively.

In item 25, we find three Arabic, 11 Turkish and 16 English students answering this item correctly even though all Arabic and Turkish students have attempted it and 119 out of 131 English students also attempted it.

In question 26, we observe that those answered correctly of Turkish and English students are ten and item difficulty is very close together at 0.09 and 0.08, respectively. While Arabic students found this question as hard, just one student answered this item correctly.

Finally, in question 30, Turkish and English students found this question to be very hard. Item difficulty for these two is 0.05 and 0.13, respectively. For Arabic students the item difficulty was a little better but still difficult with a value of 0.19.

Table 29. shows the item analysis according to language

Question No	Turkish students				English students				Arabic students			
	No Correct	N. of respond	Item Diff.	Disc. Index	No .of Correct	N. of respond	Item Diff.	Disc. Index	No .of Correct	N. of respond	Item Diff.	Disc. Index
Q.1	50	112	0.44	0.54	72	126	0.55	0.47	12	32	0.38	0.53
Q.2	28	107	0.25	0.27	45	121	0.34	0.46	4	30	0.13	0.1
Q.3	36	110	0.32	0.2	40	125	0.31	0.33	14	32	0.44	0.35
Q.4	43	110	0.38	0.49	45	129	0.34	0.68	6	32	0.19	0.43
Q.5	10	110	0.09	0.08	26	128	0.2	0.13	10	31	0.31	0.36
Q.6	63	112	0.56	0.41	69	127	0.53	0.59	16	31	0.5	0.53
Q.7	59	112	0.52	0.21	62	125	0.47	0.48	15	32	0.47	0.02
Q.8	43	113	0.38	0.18	37	129	0.28	0.23	8	31	0.25	0.13
Q.9	32	111	0.28	0.34	22	122	0.17	0.19	8	32	0.25	0.06
Q.10	30	111	0.27	0.39	50	125	0.38	0.48	5	31	0.16	0.1
Q.11	17	112	0.15	0	21	116	0.16	0.13	4	31	0.13	0.21
Q.12	66	113	0.58	0.47	64	120	0.49	0.61	9	32	0.28	0.5
Q.13	8	113	0.07	0.16	17	162	0.13	0.17	6	31	0.19	0.25
Q.14	35	113	0.31	0.54	38	127	0.29	0.35	8	31	0.25	0.5
Q.15	32	109	0.28	0.28	40	125	0.31	0.44	10	32	0.31	0.21
Q.16	46	111	0.41	0.3	39	126	0.3	0.27	12	31	0.38	0.46
Q.17	14	111	0.12	0.07	21	125	0.16	0.11	5	31	0.16	0.1
Q.18	18	110	0.16	0.03	41	126	0.31	0.37	8	31	0.25	0.36
Q.19	26	110	0.23	0.25	18	125	0.14	0.11	0	32	0	0
Q.20	19	109	0.17	0.18	10	123	0.08	0.09	2	32	0.06	0.14
Q.21	21	109	0.19	0.32	17	125	0.13	0.19	6	32	0.19	0.29
Q.22	34	109	0.3	0.43	51	123	0.39	0.36	9	31	0.28	0.17
Q.23	20	110	0.18	0.23	20	122	0.15	0.06	4	30	0.13	0.29
Q.24	54	110	0.48	0.37	63	119	0.48	0.46	10	30	0.31	0.5
Q.25	11	107	0.1	0.08	16	119	0.12	0.04	3	31	0.09	0.03
Q.26	10	109	0.09	0.17	10	118	0.08	0.07	1	31	0.03	-0.11
Q.27	45	108	0.4	0.38	45	117	0.34	0.35	10	31	0.31	0.13
Q.28	26	105	0.23	0.38	24	115	0.18	0.33	5	31	0.16	0.21
Q.29	53	108	0.47	0.47	60	115	0.46	0.43	7	30	0.22	0.43
Q.30	6	107	0.05	0.07	17	114	0.13	0.12	6	30	0.19	0.21

4.4 Evaluation of CGPA at EMU

We have found that 34.6% of the respondents have grades below C−, 21.4% gave grades of C+, C or better (success). Also, we can observe that English and Arabic students receive two grades (A, A-) 9.2%; 9.4% while Turkish students get 2.7%. Arabic students receive 21.9% for grades (B+, B, B-), then English students receive 16.8% but Turkish students get 7.1%. At level (C+, C) get Turkish students 25.7%, and since English students receive 19.1%, then Arabic students get 12.5%.

Table 30. shows the student evaluation for CGPA according to the language

		English N=131	percent	Turkish N=113	percent	Arabic N=32	percent	Total n=272	percent
A,A-	Count	12	9.2	3	2.7	3	9.4	18	6.8
	Expected	7.9		7.9		2.2		18.0	
B+,B,B-	Count	22	16.8	8	7.1	7	21.9	37	13.7
	Expected	16.3		16.2		4.5		37.0	
C+,C	Count	25	19.1	29	25.7	4	12.5	58	21.4
	Expected	25.6		25.4		7.0		58.0	
C-,D+	Count	32	24.4	52	46	10	31.3	94	34.6
	Expected	41.5		41.1		11.4		94.0	
D,D-	Count	8	6.1	10	8.84	2	6.25	20	7.4
	Expected	8.8		8.8		2.4		20.0	
F,NG	Count	21	16.1	17	15.1	6	18.75	44	16.2
	Expected	19.4		19.3		5.3		44.0	
Total	Count	120		119		33		272	

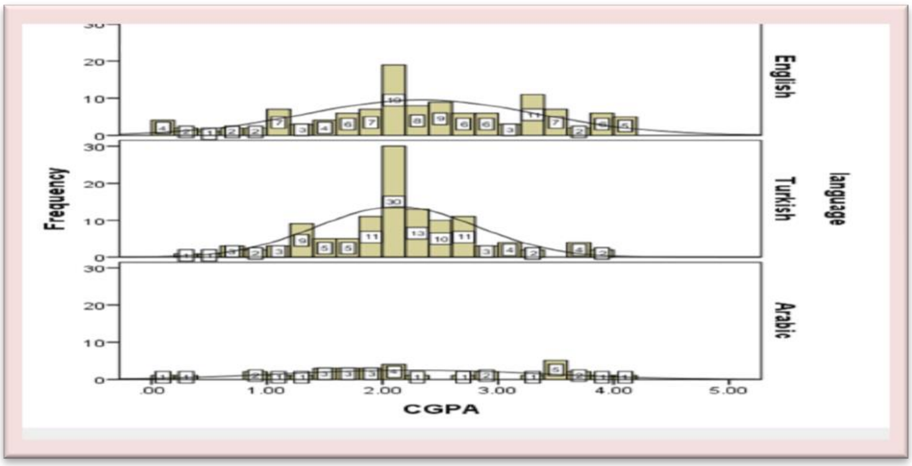


Figure 13. shows the distribution of CGPA according to language

4.5 The score of our results

We found the mean score as 27.8% for our complete sample of 282 respondents, with a mean score of 26.6% for undergraduates and 30.3% for master students. Hence, we plot our results on Hake's graphs [9] which are shown in Figures 14, 15 and 16.

Unfortunately, the results show that the level of understanding of the concepts is very poor, as we observe in each figure that EMU appears at the bottom end of each curve.

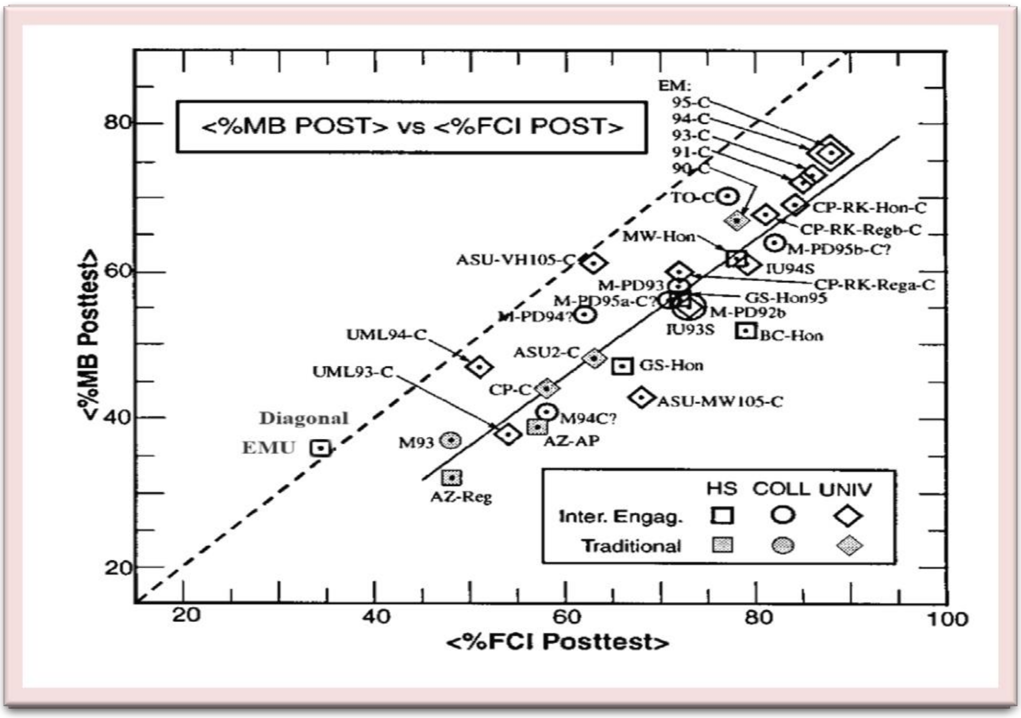


Figure 14. compares our results for undergraduate students with Hake 1997

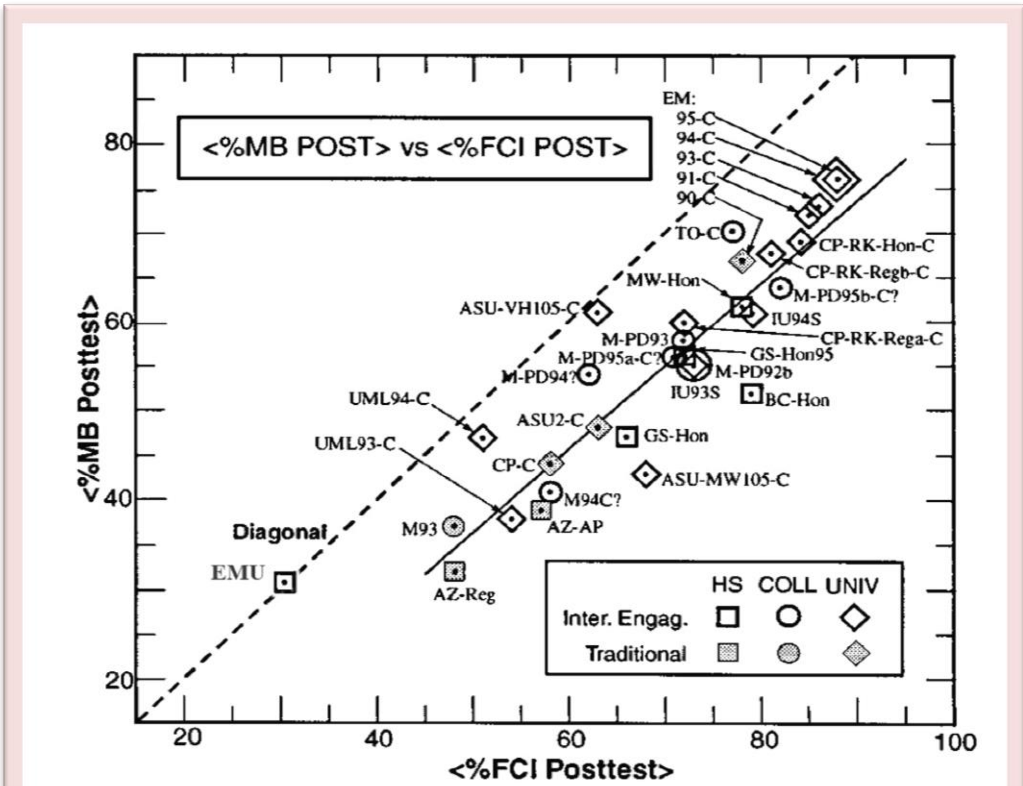


Figure 15. shows our results for master students compare with results Hake 1997

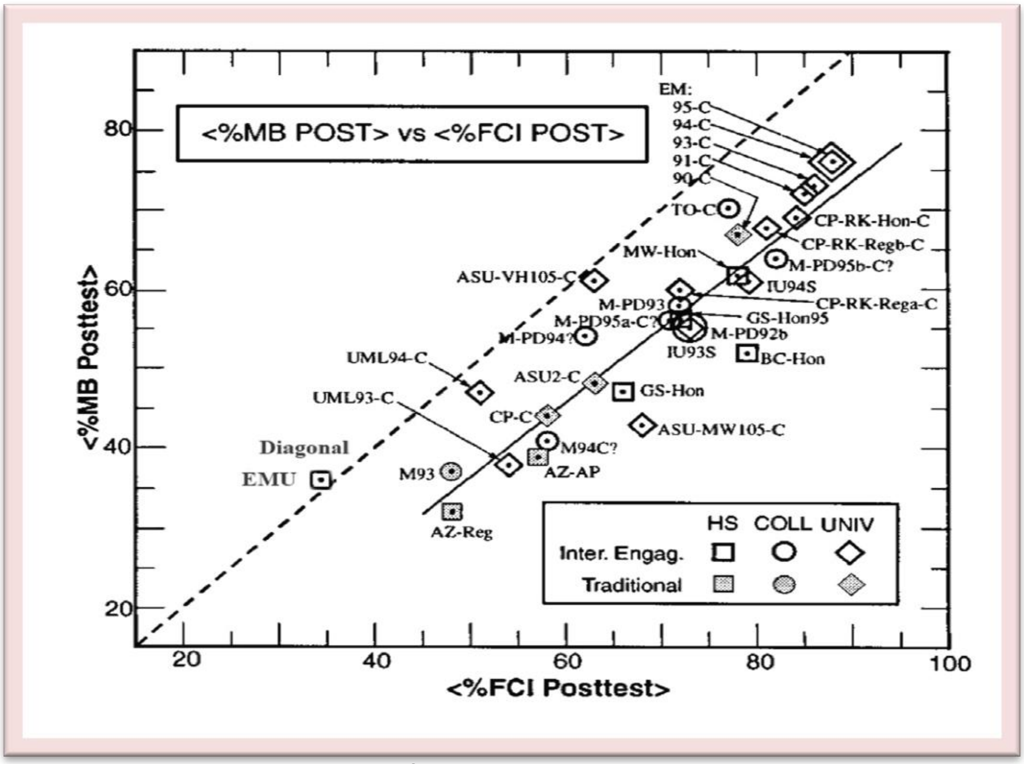


Figure 16. shows our results for all samples N=282 compare with results Hake 1997

4.8 Summary

In this chapter, we have presented and analyzed the data. Although in most instances we were unable to find relationships between the various grouping parameters and test score or differences in the test scores of the respondents, when grouped by test-language, year of study, age or registered program, CGPA, and gender. There were, however, differences when we considered the data on an item or dimension basis. So, one of our future goals is to look in to this issue in more detail.

Chapter5

CONCLUSION

5.1 Implications

By using the Force Concept Inventory as a basis for measuring student understanding of physics concepts learned during one year, we found a mean test-score of 27.8% for a sample of 282 respondents. When separated into two parts as undergraduate students, and master students we find a mean score for Master's student as 30. 3%, while the mean score for undergraduate students is 26.6 %. Whether taken separately or as a whole, the reality is that there is very low understanding of Newtonian force concept among the sampled students.

Since the relationship between courses and year is weak, the significant difference is 0.020 between PHYS101 and year. Since the significant difference is 0.464 in MATH151, and in CHEM101 the sig. is 0.065.

5.2 Comparison

In this study, the Force Concept Inventory was administered to 278 undergraduate respondents and 10 master students. The FCI-Test was administered towards the end of the spring semester of 2012-13 academic years, so in that sense we may consider it as post-test. The respondents were allowed 30 minutes on answering the FCI-Test. We found the mean score to be 27.8% for the whole sample of 288 students. Time available for the FCI-Test is an important variable. Others have allocated different times for the test. For example, Hestenes, Wells and Swackhamer, 1992, allocated 50 minutes in some of the high-schools (Arizona Reg., Wells Reg., Chicago Reg., Arizona Hon., Swackhamer Hon., Arizona AP, Swackhamer AP), and 40 minutes in some others (Van Heuvelen 105, Wells 105, Arizona State Reg., Harvard Reg., Harvard Honors).

In the same paper, the authors calculated the mean score and standard deviation in High school and University which are given in table 5.1. Our result of 27.8 % mean score and standard deviation of 3.85 are very low when compared with their scores.

Table 31 shows that compare our results with results in 1992

Class	**FCI-Post test (in 1992)**		**Our results (N=282)**	
High School	Mean score %	S.D	Mean score %	S.D
Arizona Reg.	48	16	27.8%	3.850
Wells Reg.	64	20		
Chicago Reg.	42			
Arizona Hon.	56	19		
Wells Hon.	78	15		
Swachhamer Hon.	66			
Arizona AP	57	18		
Swachhamer AP	85			
University				
Van Heuvelen 105	63	18		
Wells 105	68			
Arizona State Reg.	63	18		
Harvard Reg.	77	15		

Halloun and Hestenes in 1985, found the average FCI-Test post-test score as 42% while our result is 27.8%.

In another study, reported in the thesis “Study of Epistemological Beliefs, Attitudes towards Learning and Conceptual Understanding of Newtonian Force Concept among Physics Education Undergraduates” by S. S. Kiong in 2010 at University

Technology Malaysia, students spent about an hour answering the Force Concept inventory items [28]. In that study, the mean score obtained was 24.47% for a sample of 68 respondents.

Furthermore, in the same thesis, he stated his findings that the final year undergraduates mean score in the FCI was 27.60% with a standard deviation of 11.41% and the mean score for first years was 18.75% with a standard deviation of 8.24%. These results are much worse than ours, since first year mean score is 27.03% with a standard deviation of 12.16% while our final year mean score is 25.27% with a standard deviation of 11.54%.

Also they found in 2010, the mean score for male students was 23.63% and (N=26), and for female students the mean score was 25.00% (N=42). When these means were tested for difference using the t-test, a significance of 0.626 was obtained, indicating that there is no statistical difference between the two means. In our study, we found a mean score for males as 26.4% (N=239), and for female a mean score of 28.17% (N=33). And likewise w also found that there is no statistically significant difference between these two means.

Another comparison of our results is with that of S V Sharma and K C Sharma [29], who reported for items 5, 9, 15, 16, 17, 20, 21, 22, 24 and 26, a score rate of 32% (68% incorrect). In our study the cumulative correct score for the same questions is 23.3% (76.7% incorrect).

Steinberg and Sabella looked at students' performance in the FCI test and in their own exam, and they found a certain amount of correlation between comparable questions in the two tests [27]. More specifically, students doing well on a particular question also did well on items in the FCI that were similar to the exam questions. However, there were also discrepancies and variations for some students as well as for some comparable questions. An example of a closely similar question in the FCI and the exam was item 13. The question is about two objects that remain in contact and accelerate uniformly for the whole motion. The authors report that about half the students answered this question correctly in both the FCI and the exam. Twenty one of the students giving a correct response to the exam question also gave a clear explanation for their thought process. Suprisingly however, only six of these students gave a correct response to question 13 in the FCI, in line with their correct explanation in the exam. In the case of our sample, only 11.4% of the students (N=282) answered question 13 correctly.

Another example from Steinberg and Sabella is about the forces on an elevator moving with constant velocity in the exam and in item 18 of the FCI. Although the situations in the exam and the test are identical, 90% of the students answered the exam question correctly, while only 54% were correct in the FCI test. For comparison, the correct response to question 18 in our sample is only 24.3%.

5.3 Answering for all objectives

Our first objective as to answer the thesis question: “To what extent do engineering students master and retain an understanding of Newtonian mechanic throughout their university life”. We found, unfortunately, the mastery and retention of concepts in Newtonian mechanics is very weak amongst the engineering students, both undergraduates as the limited sample of Master's students.

The second objective was to determine the level of conceptual understanding in Newtonian force concept among the engineering students. We found poor conceptual understandings of the Newtonian force concepts and also confirmed others findings that students hold many misconceptions [1].

The third objective was to see if there is any significant difference between test language, and the level of conceptual understanding of Newtonian force. We found that there is no significant difference between English, Turkish and Arabic language test-takers. Furthermore we were unable to find any correlation or relation between students score in the FCI - test, and factors such as their academic years and registered program.

Another objective was to see if there are any significant differences among courses (PHYS101, MATH151, and CHEM101). We found that there is a statistically significant difference between three courses ‘PHYS101, MATH151 and CHEM101’ and year, but is very weak relationship.

The fourth objective was to determine if there is any significant difference between test scores of students registered in different programs. The results showed that there are no significant differences in test scores amongst students in different programs.

Another objective was to see if age had any effect on the level of conceptual understanding in Newtonian force concept. The results show that there is no statistically significant difference between age and the correct answer in the FCI-Test.

Final objective was to determine if there is any effect of the academic year of the respondent on his test score. The results showed that year in which a student is in has no significant effect on his FCI test score.

On the whole, we found in our study that the conceptual understanding by our students of the Newtonian force concept to be very low as indicated by the low scores. Furthermore we were unable to relate or associate test scores those variables that we considered to be important such as age, academic year, CGPA, achievement in science courses et cetera. Because of the limitations of this present work, it would be damaging and dangerous to draw sweeping conclusions about the EMU population. However, these results warrant further research in to this area.

5.4 Limitations

The respondents for this study were mainly engineering undergraduates from Faculty of Engineering, at EMU. Hence, the results obtained in the study cannot be generalized to all the students in EMU, because the research involved mainly the engineering students.

Also the sample size for masters students we very small (N=10), such that we cannot obtain any meaningful information or make any generalizations about this group.

The second limitation of this study is the limited participation of female students among Masters Students (N=1) as well as undergraduate students (N=33).

5.5 Extensions

Students' performance in the spring 2012-13 assessment test confirmed that students have or hold misconceptions about Newtonian physics. Future research is required to show why students find these concepts difficult. It's apparent that students tend to find some classes of concepts harder than others. These harder concepts need abstraction and information transfer.

Also, we will extend the study to include master's students and ensure a higher proportion of females so as to measure if there is a gender gap at EMU.

REFERENCES

[1] Halloun, I., Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics. 53,* 1043-1055.

[2] Hestenes, D., Swackhamer, M. (1992). Force Concept Inventory. *The Physics Teacher. 30*, 141-158.

[3] Hendrson, C. (2002). Common concerns about the force concept inventory. *The Physcis Teacher. 40,* 542-547.

[4] Mazure, E. (1997). Peer Instruction: A User's Manual. Upper Saddle River.

[5] Huffman, D., Heller, P. (1995). What does the force concept inventory actually measure?. *The physics Teacher . 33*, 138-143.

[6] Hestense, D., Halloun, I. (1995). Interpreting the Force Concept Inventory. *The Physics Teacher. 63,* 502-506.

[7] Zollman, D., Rebello, N. (2004). The efffect of distracters on student performance on the force concept inventory. *American Journal of Physics. 72,* 116-125.

[8] Hake, R. (1997). Interactive-engagement versus traditional methods:A six-thousand-students survy of mechanics test data for introductory physics

courses. 66, 64-74.

[9] Hake, R. (2002). Lesson's from the physics education reform effort http://www.ecologyandsociety.org/vol5/iss2/art28/.

[10] Foundation Coalition, Concept Inventory Assessment Instruments. Available: http://foundationcoalition.org/home/keycomponents/concept/index.html.

[11] Pavelich, M., Jenkins, B., & Birk, J., Bauer, R., Krause, S. (2004). Development of a chemistry. American Society of Engineering Education, American.

[12] Notaros, B. (2002). Concept inventory assessment instruments for electromagnetics. IEEE Antennas and Propagation Society International Symposium.

[13] Martin, J., Mitchell, J., & Newell, T. (2003). Development of a concept inventory for fluid. Frontiers in Education.

[14] Steif, P. (2004). Initial Data From A Statics Concept Inventory. American Society for Engineering Education Annual Conference & Exposition.

[15] Jacobi, A., Martin, J., & Mitchell, J., Newell, T. (2003). A Concept Inventory for Heat. Frontiers in Education Conference.

[16] Freeman, S., O'Connor, E., & Parks, J., Cunningham, M., Hurley, D., Haak, D. (2007). Prescribed Active Learning Increases Performance in Introductory Biology. *Life Sciences Education. 6,* 132-139.

[17] Lai, L. (2009). Students Performance in Introductory Physics:A Report on Trends in Physics125,126,140,and 240.

[18] Lemke, J. (1982). Talking physics.The Physics Education. 17, 263-267.

[19] Williams, T. (1999). Semantics in teaching introductory physics. *American J.Physics. 67*, 670-680.

[20] Brookes, D., Etkina, E. (2005). Do our words really matter? Case studies from quantum mechanics.

[21] http://en.wikipedia.org/wiki/Student's_t-test.

[22] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

[23] Kelley, T. (1939). The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology. 30, 17-24.

[24] http://en.wikipedia.org/wiki/One-way_analysis_of_variance.

[25] http://en.wikipedia.org/wiki/Linear_regression

[26] http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test.

[27] Steinberg, R., & Sabella, M. (1997). Performance on multiple- Choice diagnostics and complementary exam problem. *Physics Teacher*. *35*, 150-155.

[28] Kiong, S. (2010). Epistemological beliefs, Attitudes and Conceptual understanding towards learning physics among physics education undergraduates. University Technology Malaysia.

[29] Sharma, S. & Sharma, K. (2007). Concepts of force and frictional force: the influence of preconceptions on learning across different levels. *Physics Education. 42*, 516-518.

APPENDICES

Appendix A : Coded in SPSS

To use SPSS Program, we should apply coded during analysis data. Like age, we divided into three categories. Firstly, from 1995 to 1998, this gave number 1. Secondly, from 1991 to 1994, this gave number 2, and finally from 1990 to the oldest, this gave number 3.

Table 1- Code of age in SPSS

Code of Age in SPSS	
Age	Code
1995-1998	1
1991-1994	2
1990 to oldest	3

Table 2- Code of language in SPSS

Code of Language in SPSS	
language	Code
English students	1
Turkish students	2
Arabic students	3

Appendix B : Normal Distributions

We used Kolmogorov- Smirnov test to show that three courses "PHYS101, MATH151, and CHEM101" normal distribution that explain it in point courses (4.3).

Table -1 Normal distribution for PHYS101

One-Sample Kolmogorov-Smirnov Test		
		PHYS101
N		264
Normal Parameters[a,b]	Mean	1.6883
	Std. Deviation	1.33376
Most Extreme Differences	Absolute	.141
	Positive	.141
	Negative	-.103
Kolmogorov-Smirnov Z		2.291
Asymp. Sig. (2-tailed)		.000

Table -2 Normal distribution for MATH151

One-Sample Kolmogorov-Smirnov Test		
		MATH151
N		264
Normal Parameters[a,b]	Mean	2.0716
	Std. Deviation	1.32604
Most Extreme Differences	Absolute	.117
	Positive	.117
	Negative	-.099
Kolmogorov-Smirnov Z		1.908
Asymp. Sig. (2-tailed)		.001

Table -3 Normal distribution for CHEM101

One-Sample Kolmogorov-Smirnov Test		
		CHEM101
N		181
Normal Parameters[a,b]	Mean	1.9017
	Std. Deviation	1.19905
Most Extreme Differences	Absolute	.094
	Positive	.094
	Negative	-.063
Kolmogorov-Smirnov Z		1.270
Asymp. Sig. (2-tailed)		.079