Journal of Education and Vocational Research
Vol. 4, No. 4, pp. 109-118, Apr 2013 (ISSN 2221-2590)

Less Subjectivity in Setting Cut Scores: A Novel Approach

Jean Pierre Atanas
The Petroleum Institute, Abu Dhabi, United Arab Emirates
jpierre@pi.ac.ae

Abstract: Recently, standard-setting cut scores and assessment techniques became of major concerns for many organizational institutions worldwide. A cut score separates one performance level from another. It differentiates between those who pass and those who fail. They may vary according to the recommendations of policy makers and stakeholders. Passing scores were suggested by many methods on numerous types of tests: certification tests and educational tests. Most of these standard setting methods rely on panelists' subjectivity in ordering items by level of difficulty. This paper presents a simple approach to assessments by minimizing considerably panelists' subjectivity. Items are classified in levels of difficulties rather than in an increasing order in most of the standard methods. This novel approach respond to three main criteria: practicality, wide range of applicability and maximum agreement with the empirical data. Provisional and operational cut scores were derived and discussed.

Keywords: *Educational assessments, cut scores, standard-setting, assessment techniques, performance level, modular arithmetic*

1. Introduction

In the last decade, college entrance exams went through a series of major changes in dealing with scoring issues, standard-setting, cut scores and assessment techniques. Many colleges worldwide still struggle to define appropriate criteria for admission to their programs, others have done away with popular tests for admission applications. Although some colleges and universities are falling away with entrance exams, others still heavily rely on them as tools of assessment for new students' applications. According to the College Board, entrance exams are a way of assessment for academic readiness of all prospective students and their ability to pursue a given academic program in higher educational institutions (Evangelauf, 1990; Herbert, 2010; Zagier, 2011). This affirmation, concerning entrance exams, serves as a predictor for student pursuit and success in a given academic program. Well-designed entrance exams could be a good indicator for students' future performance and hence help their admission to appropriate academic programs (Akin, 2012;Konečný, 2012). In addition, well planned entrance exams could be a preventive measure for students' failure, later on, in these academic programs. By opposition, some argue that entrance exams are not determinant factors of student's abilities and skills necessary for their success (Samuels, 2008). For example, a student interested in pursuing education in a creative field like arts or music might have scored weakly in a multiple choice test but could have a fertile imagination and creativity. One could agree with this view, if students are oriented towards arts in general, or any related field that demands some creativity and artistic skills. Even in this case, well-designed multiple choice tests or standardized tests exist to assess artistic intelligences, creativities and skills that a student might possess(Manning, 2008).Another example would be of students whose first language is not English and hence could have failed in scoring well in multiple choice tests because of the language barrier or language difficulties, while their knowledge of the subject matter may be considerably good. An alternative solution of that could be to administer entrance exams to students both in English and in their native language. People proponent to the traditional entrance exams argue that too much subjectivity destroys the objective fairness which they feel puts all students on equal footing. While the debate continues, students from all over the world are still being asked to take the required tests. Students do have choices. They often are able to decide which entrance exam to take and to which colleges they wish to apply.

The mission of the PI as defined by the committee board and stakeholders is "to provide high quality engineering and science professionals through a continued commitment to excellence in its undergraduate

and graduate academic programs alongside fundamental and applied research serving the Oil, Gas and Energy sectors' need for talent, solutions and advanced technical innovations that contribute to the UAE society and economy". Stakeholders provided the academic program for foundation with sponsorship and funds to enforce the mission and the vision of the PI. With the new cohort of the academic year 2011, the management committee at the Petroleum Institute decided to administer the multiple choice exams in both languages Arabic and English. The reason behind is to be more accurate in students' discrimination towards the appropriate academic programs by eliminating the language factor. Their concern was to test students' readiness and ability to understand the major scientific concepts by minimizing all other factors that might interfere with their performance. Admission of the Petroleum Institute could not rely on high school students' reports and scores. Undoubtedly high school scores are neither good indicators nor a uniform measure of students' performance. Schools in the United Arab Emirates have different academic standards, curricula and grading scales. The ministry of higher education in the United Arab Emirates has fortunately set a common exam CEPA (i.e., Common Educational Proficiency Assessment) for all high school students desiring to apply for higher education institutions (NAPO 2011). Uniform measure of students' performance is well achieved with these planned common exams and resolves issues related to the diversity of standards in schools of the United Arab Emirates. All high school students desiring to apply for higher education institutions must sit for the CEPA test, otherwise, their application to the higher education institutions will not be approved. The CEPA exam consists of two major parts.

The first is the CEPA-English designed to test students' English proficiency, with duration of two hours and contains three sections: Grammar and vocabulary, reading and writing. The second, CEPA-Math, with duration of ninety minutes, designed to test students' ability to perform basic mathematical operations in algebra, arithmetic, geometry, data analysis and probability. While CEPA-English constitute an important requirement for students to apply to any higher education institution in the United Arab Emirates, a high score allows them direct entry to academic programs. On the other hand, CEPA-Math is used by admission for appropriate placement of the students in different academic programs or for direct entry to freshman. However, CEPA doesn't test students in Science. It was left for higher education institutions, based on CEPA-Math scores, to decide along with their entrance exams whether the student will be granted entrance to freshman or other foundation programs. Knowledge skills allow students to be separated into the appropriate levels of the AUP program (i.e., Advanced University Program).The AUP program is similar to foundation programs in other institutions of the gulf region, with the exception that it offers a wide range of tailored course curricula in English and Sciences. Learners in foundation are offered some university credits for courses in mathematics, chemistry, computation and physics. Entrance exams performance determines whether or not students will complete between one and five modules of language study within the allotted year. English language and computing courses have been designed to help students acquire the language, the technical tools, and the analytical skills required to meet entrance requirements fixed by the college of engineering. They will be expected also to take remedial courses in Mathematics and Science during the academic year before joining the freshman year. Students are assessed on a regular basis. A multiple choice test has to be designed and implemented accordingly. Student's progress is monitored by means of a data mining technique that brings preventive measures at early stages to the learning process (Atanas, 2012).

2. Methodology

Motivated by these requirements, admission has to advise an entrance exam allowing students to proceed through the AUP program in less time and to install an accurate assessment tools that discriminate students accordingly. The consensus of the working team is to focus on readiness, knowledge and understanding of the basic concepts in science. Students should be able to discern from problem statements relevant information necessary for understanding the concept being tested. They should also interpret a simple graph into textual statement and utilize information from these graphs to aid in the analysis of physical phenomena prior to join any freshman or sophomore courses. To target these skills, standards have been defined in all topics in physics. Table 1 illustrates indicators for each topic.

Table 1: Basic, proficient and advanced provisional average cut scores

Motion and kinematics (S1)	S1. Student knows how to read a graph and knows also how to do a graphical representation of motion S1. a. Students know how to solve problems that involve constant speed and average speed. S1. b. how to discern between position, distance and displacement. S1. c. Students know how to solve problems that involve instantaneous acceleration and average acceleration. S1. d. Students know and recognize free fall problems and derive the initial conditions for solving it. They should recognize that free fall is a particular case of a projectile motion. S1. e. Students know the difference between accelerated, decelerated and uniform motion, with the conditions on the acceleration and velocity for each kind of motion. S1. f. Students know how to solve problems graphically.
Motion and Forces (S2)	S2. Newton's laws predict the motion of all objects on earth. The basics to understand this concept include: S2. a. Students know how to solve problems that involve acceleration. S2. b. Students know that when forces are balanced, the net force is zero. The object continues to move at a constant speed or stays at rest (no acceleration occurs as a result). S2. c. Students know how to apply the law F = ma to solve problems that involve constant forces (Newton's second law). S2. d. Students know that forces can act on objects, and in their turn objects can react to forces applied on them. (Newton's third law). S2. e. Students know how to show a good free body diagram. S2. f. Students know applying a force to an object perpendicular to the direction of its motion causes the object to change direction but not speed. S2. g. Students know circular motion requires the constant application of a force toward the center of the circle. S2. h. Students know what friction is and how it acts on an object in motion. S2. i. Students the difference between static and kinetic friction.
Conservation of Energy and Momentum (S3)	S3. The laws of conservation of energy and momentum provide a way to predict and describe the movement of objects. The basics to understand this concept include: S3. a. Students know how to calculate kinetic energy by using the formula $E = (1/2) mv^2$. S3. b. Students know how to calculate changes in gravitational potential energy near Earth by using the formula (change in potential energy) = mgh (h is the change in the elevation). S3. c. Students know how to solve problems involving conservation of energy in simple systems, such as falling objects. S3. d. Students know how to calculate momentum as the product mv. S3. e. Students know momentum is a separately conserved quantity different from energy. S3. f. Students know an unbalanced force on an object produces a change in its momentum.
Electricity (S4)	S4. The laws of conservation of charges, predicts the motion of charges under coulomb forces. The basics to understand this concept include: S4. a. Students know the nature of an electric charge and the different methods of charging. S4. b. Students know how to calculate coulomb forces on charges and hence predict the motion of charges. S4. c. Students know the definition of an electric field and its effect on other surrounding charges. S4. d. Students know what a capacitor is and how it stores charges and energy. S4. e. Students know about resistance and flow of currents in a circuit. (Apply Ohm's law) S4. f. Students know the role and effect of capacitors in DC circuits.

The assessment will be through running a multiple choice based-exam with no free response questions. The entrance exam in Mathematics, Physics and Chemistry was implemented using Blackboard©, a very well-known learning management system. For all the arguments stated above, cut scores should be derived for a successful discrimination of students to the appropriate streams set by the institution at the AUP level. A useful method for deriving cut scores should also be advised. Setting cut scores is a challenge faced by any testing program that uses scores as part of a decision-making process. Knowing that performance levels do not depend on a particular method for setting cut scores, the working team decided to keep it simple by choosing three levels of performance: basic, proficient, and advanced. It was agreed that more than three or four levels will make the differentiation between them difficult to discern. The working team adopted the proficiency level statements defined by the NAEP (i.e., National Assessment of Educational Progress). Although, diverse methods and techniques for setting cut scores exist in the literature with their strength and weaknesses(Lin 2006), however the standard setting method developed in this paper should avoid as much as possible the subjectivity introduced by judges in item's "disordinality"(Hein and Skaggs 2009; Skaggs et al. 2007; Skaggs and Tessema 2001) explained that item "disordinality" refers to the disagreement among judges on the ordering of the items in the booklet. (Karantonis and Sireci 2006) noted that the complexity of the Bookmark method and its task for panelists is still under investigation, and panelists may not comprehend all tasks undertaken. It is necessary to apply judgment when setting cut scores, no purely objective method exists. Nevertheless, one can minimize subjectivity by providing items with equivalent level of difficulty for each category. Items were not classified in an increasing order of difficulty like in the OIB booklet of the bookmark method. Items were developed such that questions in each category are as much as possible equivalent and target basic, proficient and advanced knowledge and skills mentioned above. Working teams from the AUP program were formed from each discipline, and were mobilized to prepare a pool of two hundred question items for that purpose. From the first round, 99% agreement among judges in distributing these items to the three main categories was achieved. Each team decided to pull out from the pool of items a sequence of "x" basic, "y" proficient and "z" advanced questions. For example, the physics team pulled out the following sequence (seven basic, twelve proficient and six advanced questions). Figure 1.a, 1.b and 1.c, show examples of basic, proficient and advanced questions.

Figure 1.a: Example of a basic question with weight one

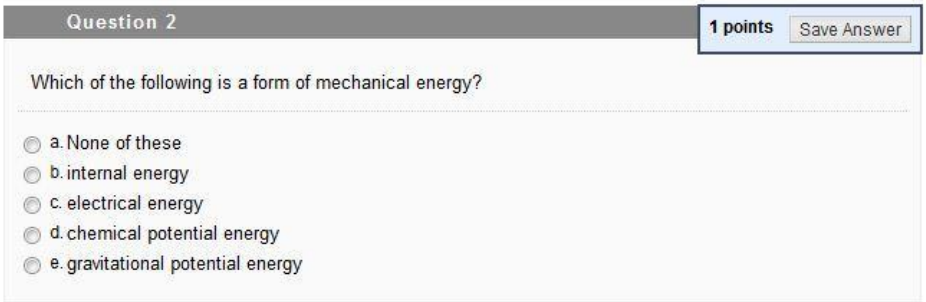


Figure 1.b: Example of a proficient question with weight ten

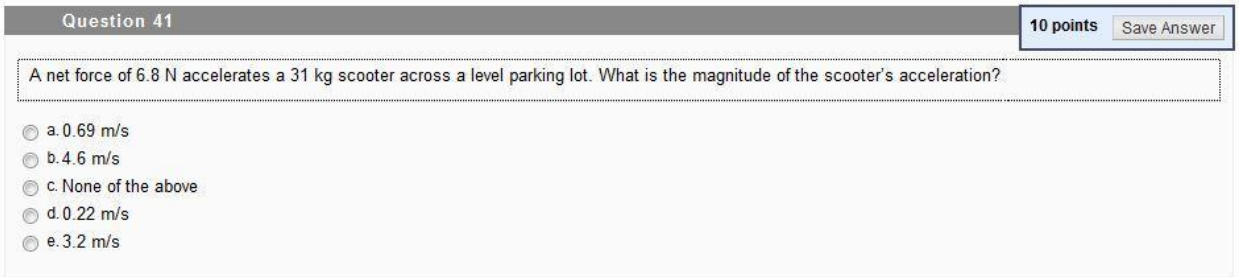
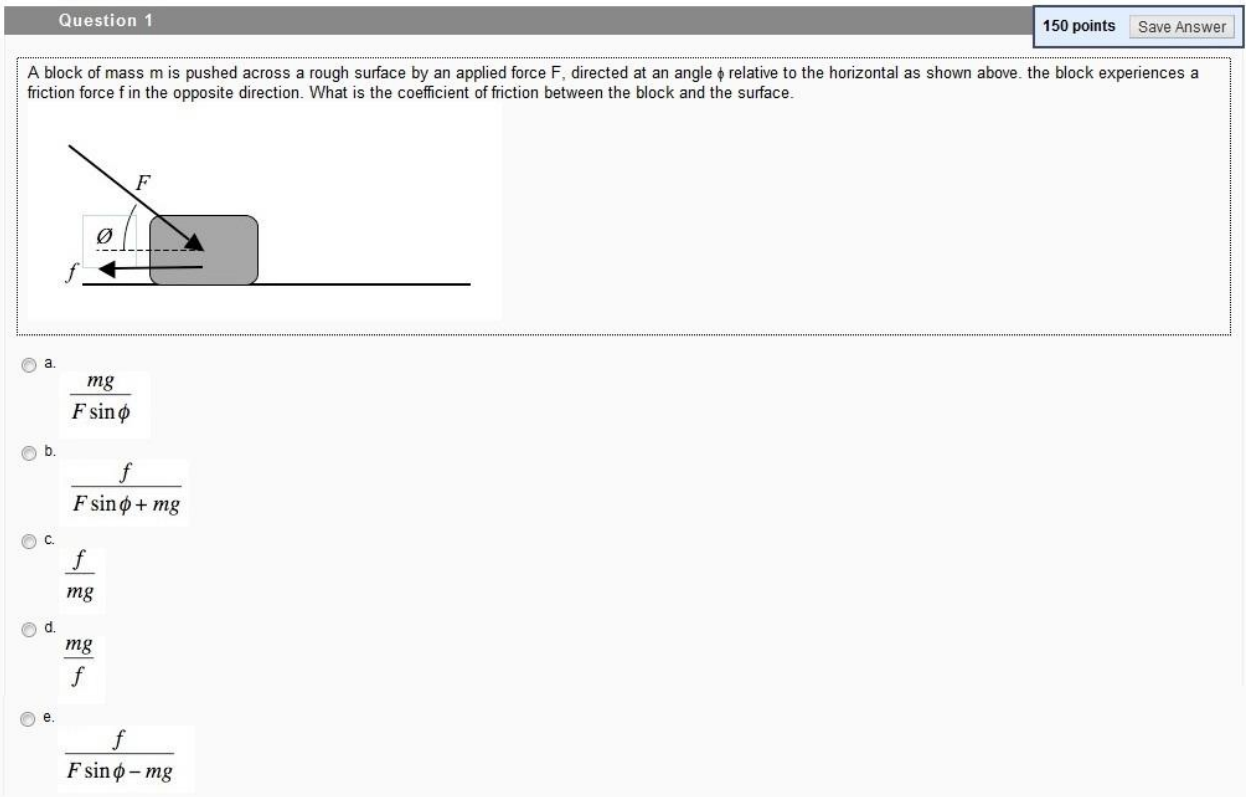


Figure 1.c: Example of an advanced question with weight hundred and fifty



Basic questions were weighted one point, proficient questions ten points and advanced questions hundred and fifty points each. The choice of this weighting for the sequence of questions is set this way to retrieve the number of answered questions for each level, since Blackboard generates total scores only. One could go through the exam of each student one by one and check for the answered questions, but this process is time consuming and requires tremendous hours of work to be completed.

3. Results

A simple method was used to extract the number of solved questions per student based on modular arithmetic and imposes that weighting sequence must be predetermined and not randomized. If all questions were answered, the total score would be:

$Total_Score = x \times w_1 + y \times w_2 + z \times w_3$, Where $w_1 = 1; w_2 = 10; w_3 = 150$ (Eq. 1)

In this case, for the physics exam, the total score is1027.Once the weight of basic questions " w_1 " is set to one, the choice of proficient questions" w_2 "must be greater than seven which represents the total set of basic questions. The same applies to the proficient and advanced questions. If the weight ten has been chosen for proficient questions, then advanced questions must be weighted more than hundred and twenty each. Hundred fifty is chosen as weight for advanced questions. To extract the number of basic questions, the algorithm to apply on the total score for each student, is straight forward and is given in pseudo-code through the following conditional statements.

To extract the number of solved advanced questions:

If (Total_Score – MODULO (Total_Score, w_3))/ w_3 > 0 Then

Return (Total_Score – MODULO (Total_Score, w_3))/ w_3

Else

Return 0

To extract the number of solved proficient questions:
If (Number_of_advanced_questions > 0) Then

Return INTEGER_PART_OF ((Total_Score – w_3 * Number_of_advanced_questions)/ w_2)
Else
Return 0
To extract the number of solved basic questions:

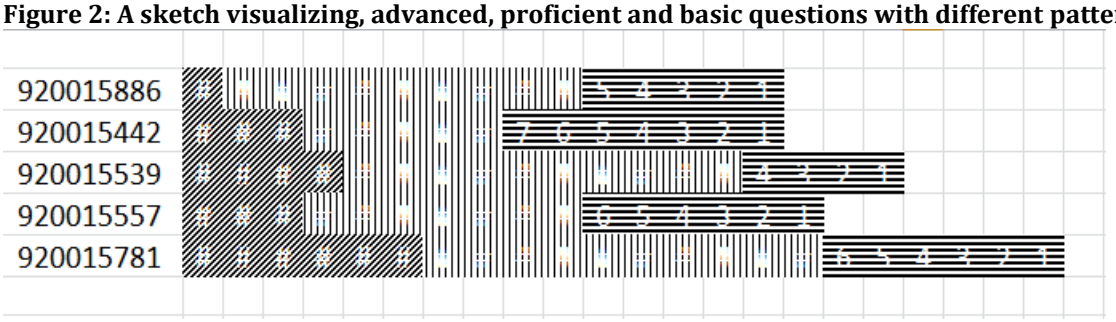
Return(Total_Score– w_3 *Number_of_advanced_questions– w_2 * Number_of_proficient_questions)
Else
Return 0
A run on a random selection of students is shown in table 2.

Table 2: Run on a random selection of students' total score yielding the above sequences

Student ID	Total score (1027)	Advanced 150pt	Proficient 10pt	Basic 1pt
920015886	245	1	9	5
920015442	507	3	5	7
920015539	704	4	10	4
920015557	526	3	7	6
920015781	1006	6	10	6

Scoring manually with identification of solved items is time consuming. The whole process is subject to human errors. A simple macro showing a visual sketch of students' performance has been designed to solve two issues. The first is to facilitate score reading without the hassle of going through understanding their origin and the second is to compare students' performance easily. A sample run is illustrated in figure 2.

Figure 2: A sketch visualizing, advanced, proficient and basic questions with different patterns.



The dark pattern to the left represents the number of solved advanced items. Light pattern, at the middle of figure 2, corresponds to the proficient items and those at the right end represent the basic ones. For about three hundred and fifty students, performance reading will be easily accessible by the visualization chart and hence comparing students' performance is immediately accessible and straight forward. In this visual representation, details of achievement in all questions are seen on one chart with a glimpse of an eye. Looking closely, one can discern possibilities of guessing. The sequence(5 advanced, 8 proficient and 0 basic questions), for instance, is not likely to occur. An interpretation of that could be a pure guessing, since cheating is made impossible for that entrance exam. Another controversial sequence would be (5 advanced, 0 proficient and 2basic questions). Here too, guessing was made from the set of advanced questions, since basic and proficient questions were mostly answered incorrectly. An algorithm was designed to point out controversial issues like the above outlined sequences. Students will be separated into the appropriate academic programs according to criteria designed by each team. For the physics team, three main streams were defined. The pre-physics stream named "AUP1" which consists mainly of fundamental topics in mechanics and electricity (i.e., kinematics, Newton's second law, the conservation of energy and basics of electricity) with emphasis on learning skills related to reading and interpreting graphs. The second stream

named“AUP2” is reserved for students having issues or lack basic knowledge of electricity but demonstrated good understanding of concepts in mechanics. It consists of introductory lessons in electricity with emphasis on critical learning and hands-on activities.

The third stream is the advanced physics program where students will be prepared to pass the College Board Advanced placement physics C exam (i.e., AP physics C) at the end of the academic year and will be granted direct entry to freshman. These students will be exempted from physics freshman courses once they join the college of Arts and Sciences. Rules of inference were developed in order to dispatch students to the appropriate streams. The result of students’ classification per gender for the three different streams is shown in table 3.Students were monitored two weeks after courses had begun to test the validity of the discrimination process. Less than 2% of the students were misclassified in the appropriate streams. The second step is to determine cut scores for the three levels of performance. The operational cut score will be determined based on the judgments of the working team within the NAEP’s performance level framework. The team agreed upon the following criteria: basic performers should prove their ability to solve 75% of the basic items, 33%of the proficiency category and none of the advanced items. Proficient performers should demonstrate an ability to solve more than 85% of the basic items, 75% of the proficiency category and 33%of the advanced items. Finally, advanced performers should answer at least85% or above of the basic items, 85% of the proficiency category and at least 67% of the advanced items. These statements were translated into rules of inference that applies to answered items. Results of the classification are summarized in table 3.

Table 3: Percentage of students’ separation into the three different streams

Gender	AP	AUP1	AUP2
Males	10	60	30
Females	18	55	27

It is worth mentioning that some examinees do not fall necessarily into the three predefined categories. As mentioned earlier, these cases are controversial and require more investigations. It is not expected that an examinee could solve more than 85% of the advanced category but wouldn’t be unable to solve33%of the basic items. Although, these cases were very rare in particular for advanced and proficient categories, some were detected from the basic category. According to table 4, no guessing for advanced students was found. Only five proficient students guessed from the advanced items category.

Table 4: Basic, proficient and advanced students with percentage of guessing in each category

Gender	Count	Basic (guess %)	Proficient (guess %)	Advanced (guess %)
Males	249	223(21)	23(17)	3(0)
Females	72	59(22)	10(10)	3(0)

However, for basic students, the guessing reached 22%. In reality, less than 10% of basic students were guessing, the remaining12 % admitted making mistakes while solving questions related to the proficient and advanced items categories. All these findings were validated during an interview with the students. Despite the guessing performed by some students, their classification into the appropriate categories hasn’t been much affected. In contrast with operational cut scores, provisional cut scores could be derived from analyzing data of solved items per category and per gender. Histograms of such findings are shown in figures 3 and 4.

Figure 3a: Females Provisional cut scores for the category items 'distribution

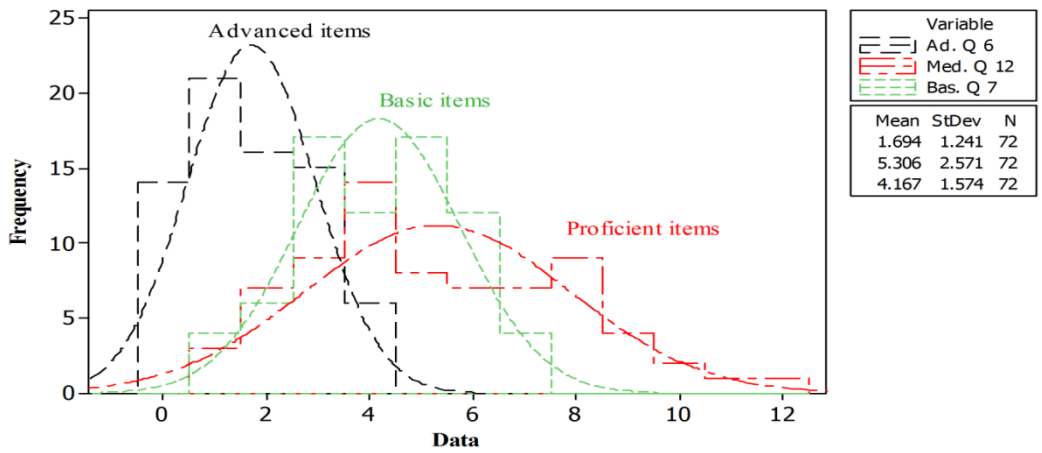
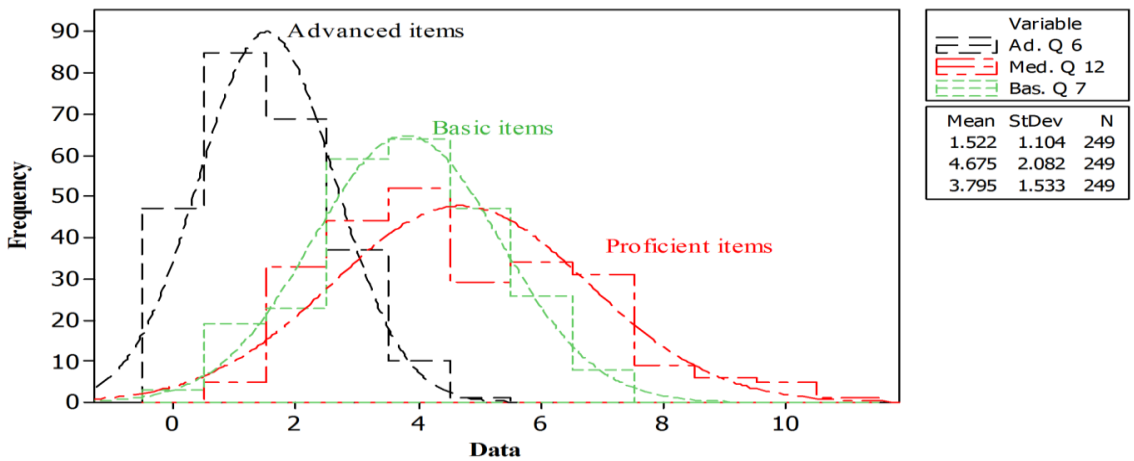


Figure 3b: Males Provisional cut scores for the category items 'distribution



Normal fit for categorical items' distribution yield the mean and standard deviation which could serve to define cut scores relative to the average level of performance for the same cohort. Table 5 summarizes these findings.

Table 5: provisional average cut score and (SD) in percent derived from figures 3 and 4.

Gender	Basic(SD)	Proficient(SD)	Advanced(SD)
Males	55(18.4)	38(17.3)	25(21.9)
Females	60(20.6)	44(21.4)	28(22.4)

Using the null hypothesis for each item's distribution and according to Shapiro–Wilk normality test, analysis showed that all items are normally distributed. Assuming an alpha level of 0.05, p-value in all cases was greater than 0.100. The Ray-Joiner correlation coefficient ranges from 0.992 to 0.997 for all data. Hence, the null hypothesis is not rejected justifying the use of a normal fit on the item's distributions. The provisional average cut scores provide valuable information about examinees' levels of performance. Provisional cut scores are relative measurements, and will definitely change with the intake of each academic year. Percentages, in table 5,represent items' average which yield cut scores for examinees' performance level. In this relative perspective, advanced performer is required to have solved more than 25% or 28 % of the

advanced items for the same cohort. Same reasoning applies on the proficient and basic performers. According to (Norcini, 1997)cut scores have to be set using a method that produces absolute and not relative assessment references. Picking a pass mark of 50% then making sure the distribution maps onto this is not a defensible approach in professional training as it produces relative standards and examination results that cannot be compared to an agreed standard or compared year over year. Therefore, the provisional cut score derived from data distributions won't be applied, in practice, for the 2011 cohort. However, it provides valuable information about students' readiness and other educational factors. In fact, the provisional cut score will serve as variable indicator for students' performance for the academic year around the operational cut score that serves as a reference.

4. Conclusion

This study helped in addressing issues related to students' weaknesses by detecting them earlier at admission level. Specially designed placement exams, as detailed in the manuscript, were developed with the help of stakeholders to provide a quantitative and accurate measurement of how much concepts are being grasped and understood by students in some schools of the UAE and what could be done to address these issues. It shows how useful sometimes students streaming into similar levels, lacking additional education background, students were put in an" incubational phase" where they are prepared for freshman. The study could serve for many as a fire alarm to detect student's level of performance early at admission stage. It is known that proving the correctness of a cut score is impossible. Therefore, it is crucial to follow a process that is appropriate and defensible. Ultimately, cut scores are based on the opinions of a group of people. The best we can do is choose the people wisely, train them well in an appropriate method, minimize subjectivity to the minimum, evaluate the results, and be willing to start over again if the expected benefits of using the cut scores are outweighed by the negative consequences. Sometimes, students may score well above average on a commercially developed standardized test or on a "world-class" test like AP, SAT or IB, and still be labeled a "failure" or in "need of improvement". This method distinguishes itself in its simplicity, reliability and in minimizing subjectivity of decision takers. The method doesn't need to sort questions or seeks to approximate the probability of answering an item by a student. Most of us have experienced the stress associated with taking a test. But what if your future is directly tied to whether or not you pass? Add the fact that the definition of "pass" is largely a judgment call. It is clear that mislabeling a student in this situation has emotional and psychological consequences beyond the prospect of graduation or promotion decisions. Given these circumstances, it is imperative that we continue to scrutinize performance levels to help insure their proper use.

References

Akin, M. S. (2012). An analysis of the competion for university entrance test. *Journal of Developing Areas*, 46(1), 55-70.

Atanas, J. (2012). Measuring Students' Performance with Data Mining. *Journal of Education and Vocational Research*, 3(5), 132-137.

Evangelauf, J. (1990). College Board to Revise Entrance Exam; Says New Version Will Be More Useful. (cover story). *Chronicle of Higher Education*, 37(10), A1-A34.

Hein, S. F. & Skaggs, G. E. (2009). A Qualitative Investigation of Panelists' Experiences of Standard setting Using Two Variations of the Bookmark Method. *Applied Measurement in Education*, 22(3), 207-228.

Herbert, M. (2010). ACT Sets New Goals for College Readiness. *District Administration*, 46(1), 10-10.

Karantonis, A. & Sireci, S. G. (2006). The Bookmark Standard setting Method: A Literature Review. *Educational Measurement: Issues and Practice*, 25(1), 4-12.

Konečný, T. J. J. N. (2012). Alternative models of entrance exams and access to higher education: the case of the Czech Republic. *Higher Education*, 63(2), 219-235.

Lin, J. (2006). The Bookmark Procedure for Setting Cut-Scores and Finalizing Performance Standards: Strengths and Weaknesses. *Alberta Journal of Educational Research*, 52(1), 36-52.

Manning, D. (2008). Identifying students' mathematical skills from a multiple-choice diagnostic test using an iterative technique to minimise false positives. *Computers & Education*, 51(3).

NAPO. (2011). Interpreting CEPA Scores. NAPO technical report.

Norcini, J. S. J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, 10, 39-59.

Samuels, C. A. (2008). N.Y.C. Entrance Exam Questioned. *Education Week*, 28(10), 5-5.

Skaggs, G., Hein, S. F. & Awuor, R. (2007). Setting Passing Scores on Passage-Based Tests: A Comparison of Traditional and Single-Passage Bookmark Methods. *Applied Measurement in Education*, 20(4), 405-426.

Skaggs, G. & Tessema, A. (2001). Item disordinality with the bookmark standard setting procedure. Paper presented at the Annual Meeting of the National Council on Measurement in Education,Seattle, WA.

Zagier, A. S. (2011). ACT Scores Show College Readiness Problems Persist. *Community College Week*, 24(2), 3-3.