Confidence-Based Assessment of Two-Alternative Format Tests

AHMED A. BELAL (1) & DIALA F.AMMAR (2)

(1)Computer Engineering and Informatics- Beirut Arab University, Lebanon
(2)Psychology department, Lebanese American University, Lebanon
abelal@bau.edu.lb

Abstract

The effect of guessing in multiple choice tests is usually reduced by using a large number of test items and increasing the number of choices per item. The paper addresses the important and practical issue of grading Two-Alternative format tests, that is, decision questions of the true/false type, where the answer is only one of two choices .This format is not only easy to construct, but could reflect real life situations such as when a judge has to decide between a GUILTY or a NOT GUILTY verdict, or when a computer science student is asked whether a given computational problem is NP-hard or not. In binary-choice tests where the number of choices per test item is restricted to two, the degree of certainty that the student has in answering the test item becomes an important factor in assessing his/her score.

The standard procedure for incorporating certainty/confidence levels in grading multiple-choice tests is to have each student attach one of several, usually four, confidence levels to each question separately. In this paper we suggest a different confidence-based procedure for grading binary-choice tests. Assuming that humans tend to be more comfortable in assessing their confidence level to different things in a relative manner rather than in an absolute manner, the concept of relative uncertainty is used. Each student is asked to rank the test items relative to each other according to his/her confidence level in answering the item correctly. The test is graded according to the rank sequence produced. There is no penalty for wrong answers. A step-size is used to determine changes in the student's confidence level based on the number of incorrect answers made. A reduction function is used to determine the scores to correct answers at the different confidence levels. By varying the step-size and the reduction function, many different grading strategies can be obtained.

Key words: confidence-based assessment, binary-choice tests, true/false exams.

I. Introduction

Multiple choice tests are being extensively used as a means of objectively testing large numbers of students by standardizing the grading procedure . Standard procedures for testing include using a large number of test items, increasing the number of choices per item and setting a penalty for a wrong answer.

Computer-Aided Assessment techniques are also used in higher education as a means to keep down the time and effort invested in grading [1].

Automatic grading is not easily done with free answer exams, although some efforts are being made in this area [2].

As reported in [1] some students may not be very comfortable changing from a free answer test format to a multiple choice test, mostly because ,with the former type of test, the student can get a partial credit for his answer and there is no penalty for a wrong answer.

There are many ways to construct a multiple choice exam. In normal multiple choice questions ,each question has a stem and several possible answers of which only one is correct, the other choices are merely distractors. Designing good distractors for a given question is not a trivial matter.

Another is multiple-select exams, where more than one choice may be a correct answer. To get a question correct ,the student has to mark all of its correct answers, but there is no penalty for an incorrect answer. Improving on this format is the multiple T/F exam. The student is given several choices and has to mark each choice as either true or false. This allows for arranging questions with no correct answers. Still no partial credit is given for incomplete correct answers.

Finally the simple format of individual T/F questions can be adopted, with questions grouped by topics to help the student concentration.

Studies have shown [3] that the simple format of individual T/F exams are comparable to open ended questions with respect to the score ranking.

The main disadvantage to the individual T/F format is their weak resistance to guessing, where a correct answer may be chosen based on intuition or chance instead of knowledge.

Some opinions accept guessing and feel that the students can guess with no penalty to make it fair [4]. Others disagree and believe that students should be penalized for a wrong answer to reduce guessing.

In an attempt to eliminate the effect from random guesses on the average, a popular grade assignment scheme [1] is +x points for a correct answer,0 points for an unanswered item and -x points for a wrong answer, where x is a positive number. A pure random guesser is expected to score a total of 0 points.

These kinds of grading assignments with true/false type questions fail to discriminate correctly between the students even when a heavy penalty is given for incorrect answers.

A grading system based on the student degree of certainty, here after called "confidence level", was suggested to provide better discrimination in students' grades [5,6].

In this system the student specifies his/her confidence level for each test item. Four confidence levels are suggested. The higher the level, the more the score for a correct answer and the more the penalty for a wrong answer as shown below.

Confidence Level	**Chance of correct answer**	**Score for correct answer**	**Score for incorrect answer**
0	0 - 25	0	0
1	26 - 50	+3	-1
2	51 - 75	+4	-2
3	76 - 100	+5	-5

Computer programs that generate multiple choice tests and support this grading system were reported [7, 8].

Although the grading system with degrees of confidence level is widely accepted [9, 10], studies have shown that it is not easy for the student to correctly measure his/her confidence level to a given item.

Multiple choice questions as well as yes/no questions do not provide sufficient information on how much information they have learned.

One unresolved question in the literature relates to the relationship between knowledge and perception of this particular knowledge. Answering correctly a test item relates to cognitive accuracy whereas perception of knowledge refers to confidence [10]. Two dominant approaches have attempted to explain the effect of confidence on performance. The ecological approach [11] believes that the type of questions on the test contribute to confidence levels. In other words, environmental factors out of the individual's locus of control affect the match between cognitive accuracy and perceptual knowledge. In contrast, the heuristics and biases approach believes that internal subjective factors such as negative feelings or previous experience contribute to a mismatch between what we know and what we think we know.

Previous work has attempted to apply the expectancy-value theory of motivation to test performance [12, 13, and 14] .This theory states that student's performance is highly connected to the importance of the task and the prospect of success. Researchers have included other variables that could affect performance such as emotional attributes, motivation, test anxiety, personal traits, perceived effort to effectively complete diverse items on the test (such as amount of time to study). Others attributes could include confidence level when answering questions on exams combined with one's positive or negative belief of personal test-taking abilities [19]

Previous work has demonstrated individual differences in confidence (e.g., [20, 21, and 22]. Literature regarding personal characteristics (such as affective factors) is relatively new and inconsistent. Some studies have demonstrated that confidence ratings increase with learning abilities. For example, [15] tested undergraduate students in an introductory psychology course and rated their confidence levels before and after the test. Findings indicated a positive correlation between test performance and confidence levels. Similar findings were reported when undergraduates where asked to rate their confidence levels when answering multiple choice tests especially with students who had higher memory aptitude [16]. In contrast, low performance students have been reported to overestimate their confidence levels in performance judgments. Overestimations of confidence have been correlated with test difficulty [17]. Poor performers tend to overestimate their abilities and in turn may not allocate enough studying time before taking the test. Other variables have been found to affect confidence. For example, [18] indicated that student's confidence levels and performance dropped when test items were placed randomly. Overall, most individuals tend to have biased perceptions and tend to overestimate their performance (e.g., [23, 24] There is extensive research confirming the effect of study skills and abilities, anxiety etc. and performance but information regarding affective factors such as confidence is still scarce. According to [19] affective factors such as self-perception and cognitive abilities (such as grade point average) are strong predictors of test performance.

Understanding the effect of confidence on test performance would ultimately provide a more comprehensive approach to interpreting test scores.

Current procedures for incorporating confidence levels in the test score, is to ask the student to assign a confidence level to each answered item .Studies have shown that students tend to invariably misestimate their confidence levels. Some students tend to always overestimate their level of confidence while other conservative students almost always, when in doubt, underestimate their confidence level. Another drawback to the current system of grading is the penalty assigned for incorrect answers.

In this paper we suggest a novel system where students are asked to assign a confidence level to each test item relative to the rest of the items as opposed to an absolute confidence level assigned to each item separately. We believe this scoring technique is more transparent to the students and is more dynamic, allowing for many different grading strategies.

II. Relative Confidence Level

In a binary-choice test with K items, the student is asked to rank the K items according to his/her relative confidence level among the K items, by numbering them 1,2,3,….,K in such a manner that if C(i),C(j) are the confidence levels to the questions ranked i , j respectively then $C(j) \leq C(i)$ for j > I, that is the item with rank 1 is the one with the highest relative confidence level.

In order to make the task of ranking the questions easy, the value of K should be kept small, for example 10, by breaking the whole test into several modules each with K questions.

III. Scoring

The number of points the student gets for a correct answer to a question should decrease when the confidence level decreases. If 10 points are awarded for a correct answer to the question with rank 1, that is the question with highest relative confidence level, a reduction function R(C) is used to determine the value of the score when the confidence level C decreases.

There is no penalty for incorrect answers, they are merely used to indicate changes in the student's confidence level. The step-size S is the number of incorrect answers to reduce the confidence level by one step. Both the reduction function R(C) and the step-size S can be varied to achieve many different grading strategies. For example in a 10 question module with +10 points assigned to the highest confidence level, a reduction function which decreases the score by 2 points for every decrease in the confidence level will produce the score assignments shown

CONFIDENCE LEVEL	C1	C2	C3	C4	C5	C6
SCORE	10	8	6	4	2	0

If a "1" indicates a correct answer to a test item and a "0" indicates an incorrect answer, then a student's answer to the 10 questions can be represented by a 10-bit vector with the most significant bit representing the answer to the question with rank 1. Using the reduction function described above, the score for the answer vector "1101001101" will be

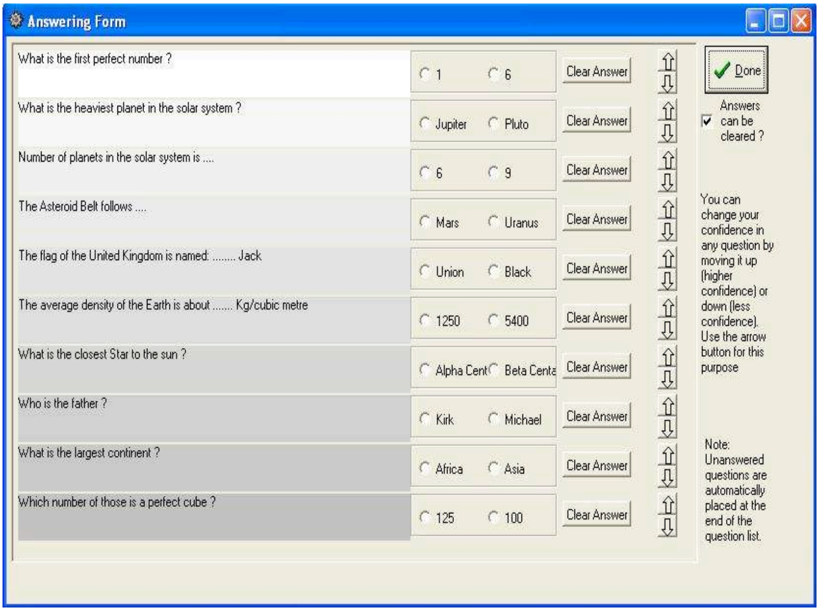
With step-size S = 1, Score = 10+10+8+4+4+2 = 38
With step-size S = 2, Score = 10+10+10+8+8+6 = 52

IV. Grading

Once the K-bit answer vector is determined, the score can be calculated using the reduction function and the step-size defined for the test. The test administrator can then check the scores and may simply change the reduction function and/or the step size to get other score distributions. To avoid any errors made by having the students mark the rank of each question with a number from 1 to K, the tests were administered on computers with an interface that allowed the student to rank the questions by actually moving them up and down relative to each other, with the higher ranking question being physically on top of a lower ranking one.

V. Implementation

Several tests were administered to junior and senior students of the computer science department at the university of Alexandria in Egypt. All tests were of the true/false type. Both the grading system based on the relative ranking and the one based on independent ranking were used. The student felt more comfortable with the relative ranking system, but had some time trouble ranking the items when their number exceeded 20. With modules of 10 items no problems in ranking were mentioned.



What is the first perfect number ? ○ 1 ○ 6

What is the heaviest planet in the solar system ? ○ Jupiter ○ Pluto

Number of planets in the solar system is .... ○ 6 ○ 9

The Asteroid Belt follows .... ○ Mars ○ Uranus

The flag of the United Kingdom is named: .... Jack ○ Union ○ Black

The average density of the Earth is about .... Kg/cubic metre ○ 1250 ○ 5400

What is the closest Star to the sun ? ○ Alpha Cent ○ Beta Centa

Who is the father ? ○ Kirk ○ Michael

What is the largest continent ? ○ Africa ○ Asia

Which number of those is a perfect cube ? ○ 125 ○ 100

☑ Answers can be cleared ?

You can change your confidence in any question by moving it up (higher confidence) or down (less confidence). Use the arrow button for this purpose

Note: Unanswered questions are automatically placed at the end of the question list.

Fig.1

Following are the results of a test administered to the senior class of 28 students. The test was of the true/false type and covered general topics in computer science. The test was administered on computers with the interface shown in Fig.1 which allows the students to rank the questions by physically moving them up and down relative to each other. The students were also asked to assign one of four confidence levels to each item separately according to the table:

CONFIDENCE LEVEL	C1	C2	C3	C4
CORRECT ANSWER	10	6	2	0
INCORRECT ANSWER	-10	-4	-1	0

The scores were computed using the confidence levels assigned by the students to each separate question and are shown in the first column, termed absolute, in table 1 below. Several grading functions based on the relative ranking proposed in this paper ,were applied to the ranked answer vectors ,shown in the last column of table 1 and the scores were included in the table.

The first column below gives the scores using the above table. The second and third columns give the scores based on the relative ranking for the two scoring functions (10-2p) and (10-p) respectively, where p is the number of incorrect answers to items of higher relative rank.

The entire test consisted of 10 modules each having 10 items. The table shows the results of the first module. The fourth column is an example of a harsh grading function where a lot of emphasis is placed on how well the student trusts his/her answer to a given question. The score is computed using the function (10-5p) thus the student is only allowed two errors in his test. Any correct answers after committing two errors will gain no score. These types of exams may be of great importance in jobs where the degree of confidence in one's actions is a critical issue, and they lead to better discrimination amongst the scores . On the other hand, column five uses the lenient scoring function with step size S=2 and a reduction value R=1, that is the score for a correct answer is only reduced by one every two errors . Column six uses a moderate scoring function with S=2 and R=2.Table 2 below gives the average and standard deviation for the different scores and different grading strategies. Fig.2a through Fig.2f give column charts of the different strategies, while Fig.3a through Fig.3f give the corresponding bar charts .

Absolute	(10-2p)	(10-p)	(10-5p)	R=1;s=2	R=2;S=2	Answer Vector
16	24	32	10	37	34	1001011000
38	44	47	35	50	50	1101110000
25	40	50	15	56	52	1010111010
24	30	40	10	47	44	0110101100
62	64	72	40	76	72	1111001111
80	90	90	90	90	90	1111111110
62	66	73	45	80	80	1011111110
37	42	46	25	50	50	1011110000
18	32	36	20	39	38	1011010000

51	64	72	40	80	80	0111111110
30	50	60	30	66	62	1110011011
86	78	84	60	90	90	1110111111
46	42	56	20	65	60	1100101111
30	36	38	30	40	40	1101100000
58	64	67	55	70	70	1111011100
49	62	66	50	70	70	1110111100
70	58	64	40	70	70	1011111100
38	42	46	30	50	50	1011110000
16	22	36	5	45	40	0100111010
41	40	45	25	50	50	0111110000
16	24	32	5	37	34	0101101000
0	24	32	0	27	24	0011100000
1	22	26	10	29	28	0110100000
29	30	35	20	39	38	1011001000
4	34	37	30	39	38	1110001000
28	30	40	5	46	42	0101110100
54	62	66	50	68	66	1111100110
40	52	61	30	66	62	1110011101

Table 1

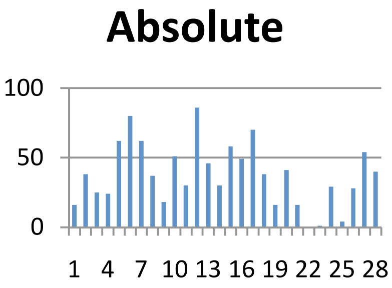


Fig 2.a

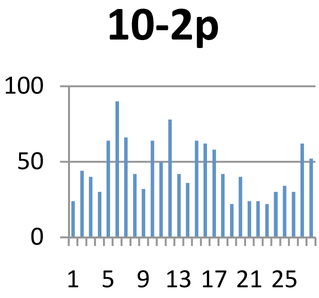


Fig 2.b

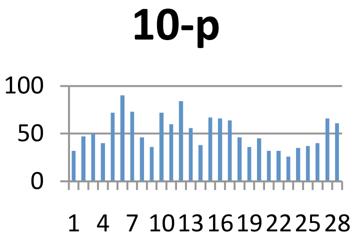


Fig 2.c

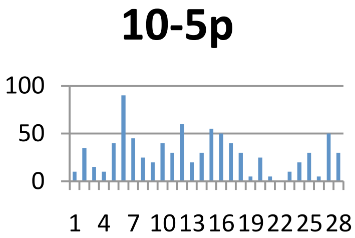


Fig 2.d

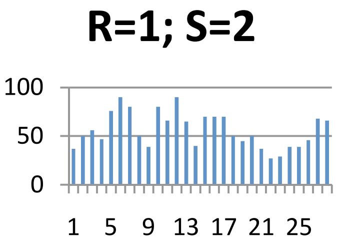


Fig 2.e

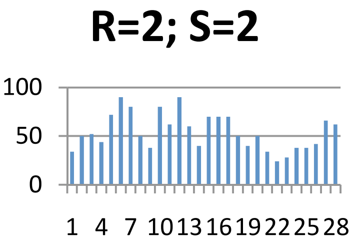


Fig 2.f

Strategy	Average	Standard Deviation
Absolute	37.46428571	22.57084672
10-2P	45.28571429	18.21854396
10-P	51.75	17.3901313
10-5P	29.46429	20.19989
R=1 S=2	56.14286	17.98912
R=2 S=2	54.42857	18.44597

Table 2

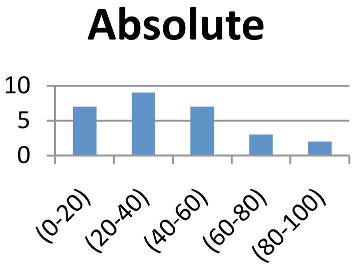


Fig 3.a

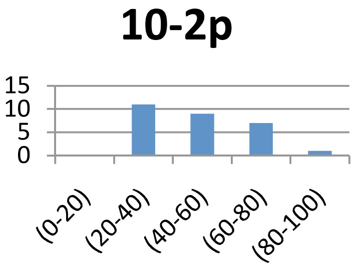


Fig 3.b

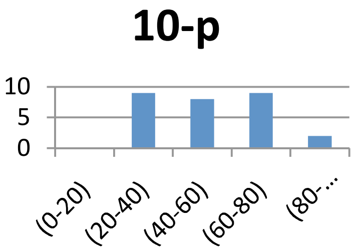


Fig 3.c

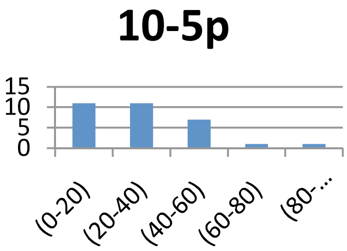


Fig 3.d

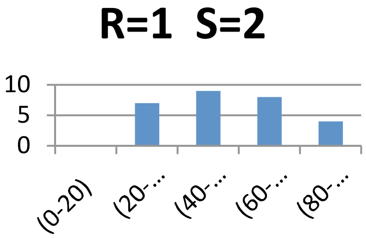


Fig 3.e

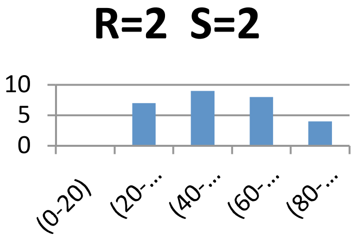


Fig 3.f

CONCLUSION

The concept of confidence levels widely accepted in grading multiple-choice tests is modified to the benefit of the student when the test is of the binary-choice type. The student's level of confidence in answering an item is used to rank the test items relative to each other, a confidence level step-size and a reduction function can be chosen in a variety of ways to provide different grading strategies. The proposed scheme has the advantages of being more dynamic, and more transparent to students than other used schemes.

.

REFERENCES

[1] B.Weiss, G.Gridling, C.Trodhandl, W.Elmenreich, ‘Embedded Systems Exams with True/False questions :A case study ‘, Research Report 5/2006,T U Wein,Institut fur Technische Informatik,2006.

[2] P.Thomas, ‘The Evaluation of Electronic Marking of Examinations’,8th Annual Conference on Innovation and Technology in Computer Science Education,2003.

[3] L.W.Lackey and J.W.Lackey, “ Influence of true/false tests and first language on engineering students test scores “,Journal of Engineering Education, 2002.

[4] M.Bar-Hillel,D.Budescu,and Y.Attali ‘ Scoring and Keying multiple choice tests :A case study in irrationality’,Mind and Society,1(1),2005.

[5] D. Leclereq, “Banque de Questions et Indice de Certitude (I)”, Educ, Tribune Libre, no. 149, pp. 49-58, 1975.

[6] D. Leclereq, “Banque de Questions et Indice de Certitude (II)”, Educ, Tribune Libre, no. 150, pp. 71-79, 1975.

[7] R. H. Posteraro, D. E. Blackwell and L. Huddleston, “Techscore : A program for tabulating the results of multiple choice questions and correcting multiple choice examinations”, Comput. Biol. Med., vol. 16, pp. 259-265, 1986

[8] P. lira, M. Bronfman and J. Eyzaguirre, “Multitest II : A Progra for the Generation, Correction, and Analysis of Multiple Choice Tests”, IEEE Trans. Education, vol. 33, no. 4, pp. 320-325, 1990.

[9] A.Gardner-Medwin and M.Gahan,’ Formative and summative confidence-based assessment’,7th Annual Computer Assisted Assessment Conference 2003.

[10] S.Lichentenstein, B.Fischoff and L.D.Philips, “Calibration of probabilities: The state of the art to 1980”.In D.Kahneman, P.Slovic, & A.Tversky (Ed.), Judgments

under uncertainty: Heuristics and biases (pp.306-334). Hillsdale,NJ:Erlbaum, 1982.

[11]G.Gigerenzer, U.Hoffrage, & H.Kleinbolting, "Probabilistic mental models: A Brunswikian model of confidence", Psychological Review, 98, pp506-528, 1991

[12] . Pintrich, P. R. (1988). A process-oriented view of student motivation and cognition. In J. S. Stark & L. Mets (Eds.), Improving teaching and learning through research. New directions for institutional research, 57 (pp. 55-70). San Francisco: Jossey-Bass.

[13] Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In C. Ames & M. Maehr (Eds.), Advances in motivation and achievement: Vol. 6. Motivation enhancing environments (pp. 117-160). Greenwich, CT: JAI Press

[14] L.F.Wolf, J.K. Smith and M.E. Birnbaum, M.E, " Consequence of performance, test motivation and mentally taxing items". *Applied Measurement in Education,*1997, 8, 341-352

[15] K. Sjostrom & A.Marks, " Pretest and Posttest Confidence Ratings in Test Performance by Low-, Medium- and High-Scoring Students" ,Teaching of Psychology, Vol. 21, 1994

[16] J.J. Shaughnessy**,** "Confidence-judgment accuracy as a predictor of test performance". *Journal of Research in Personality*, 1979, 13, 505-514.

[17] Roedel, T., Schraw. G., & Plake, B. S. (1994). Validation of a Measure of Learning and 12 Performance Goal Orientations. Educational and Psychological Measurement, 54, 1013-1021.

[18] F.Savitz, **"**"Effects of Easy Examination Questions Placed at the Beginning oScience Multiple-Choice Examinations." *Journal of Instructional Psychology*, 1985, 12(l), 6-10

[19] J.L., Smith, "Understanding the processes of stereotype threat: A review of meditational variables and new performance goal directions." *Educational Psychology Review***,** 2004, 16, 177–206.

[20] Stankov & R.D. Roberts ,G. Pallier, R. Wilkinson,V. Danthir,S. Kleitman, G. Knezevic, L. "The Role of Individual Differences in the Accuracy of Confidence Judgments", *The Journal of General Psychology*, 2002, 129(3), 257–299

[21] L.Stankov and J.Crawford (1997). “Self-confidence and performance on cognitive tests”. *Intelligence,* 1977, 25, 93–109.

[22] Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology, 19, 460-475.

[23] Fischof, B., P.Slovic. and S.Lichtenstein. “Knowing with Certainty: The Appropriatness of Extreme Confidence.” *Journal of Experimental Psycholog:Human Perception and Performance*, 1977, 3(4),pp.552-64.

[24] S.Lichentenstein, B.Fischoff and L.D.Philips, “Calibration of probabilities: The state of the art to 1980”. In D.Kahneman, P.Slovic, & A.Tversky (Ed.), Judgments under uncertainty: Heuristics and biases (pp.306-334). Hillsdale, NJ:Erlbaum, 1982.

.