Can Confidence Assessment Enhance Traditional Multiple-Choice Testing?

Josef Kolbitsch[1], Martin Ebner[1], Walther Nagler[1], Nicolai Scerbakov[2]

[1]Computing and Information Services, Graz University of Technology
[2]Institute for Information Systems and Computer Media, Graz University of Technology

Keywords: *e-learning, computer aided assessment, multiple-choice, confidence*

Abstract:

This paper presents the results of an experiment with multiple-choice testing including confidence assessment. In a course at Graz University of Technology (TU Graz) 432 students did a multiple-choice test (MCT) on the university's online learning management system. For 172 students the test had been added a confidence parameter for each question, which allowed the students to state their confidence in their answers. The remaining 260 students doing a traditional MCT served as a control group. The results show that there is a relationship between the confidence parameter and the percentage of incorrect answers. Moreover the findings detail that the use of the confidence parameter leads to slightly poorer results.

1 Introduction

Multiple-choice tests have a long tradition in education. Although MCTs are often hard to construct big courses rely on this kind of true-false testing, because marking is uncomplicated

Several publications detail the shortcomings of MCTs. There are two distinct problems related to them — misinformation, misconceptions and guesswork [6] [7]. Especially in paper-and-pencil MCTs students often are not informed which questions were answered incorrectly and what the correct answers would have been.

Moreover students sometimes accomplish their results by a combination of partial knowledge and guesswork. Hence with conventional MCTs it cannot be ensured that the students' knowledge is appropriately reflected [1], [3].

1.1 Confidence Assessment

Previous research in the area of MCTs has mainly focused on the writing of distinct questions and corresponding clear answers. Only little work has been done on countering the drawbacks mentioned above. An approach to decrease guesswork is the use of confidence assessment [4], [8], which means that students do not only have to choose an answer to a multiple-choice question but also have to indicate how confident they feel about their answer given.

At Graz University of Technology a research group decided to investigate confidence assessment in detail. For this purpose the MCT module of the “TU Graz TeachCenter” (TUGTC), an online learning management system developed by the Institute for Information Systems and Computer Media (IICM) [2] [5], has been extended to support confidence assessment. This means that each question of a MCT is complemented with an additional parameter “confidence” to allow students to estimate and state the grade of correctness of their answers.

Moreover, after the completion of the online MCT the correct results for every question of the test are displayed to the student minimizing the likelihood of misinformation.

2 Approach and Methodology

2.1 Lecture Design

In October 2007 the course “Application of Operating and Information Systems” was offered for the first time at TU Graz. The course is recommended to all students of the university though it is not obligatory nor is the MCT reflecting it. Prior knowledge is not required. In winter term 2007 720 students attended the lecture and 432 of them did the final exam—an online MCT with or without confidence assessment.

2.2 Methodology

2.2.1 Test Design

100 questions were designed for that special MCT. The design of the multiple-choice items does not only include assigning a number of points for a correct question. For confidence assessment also the number of (potentially negative) points for an incorrectly answered question and the possible values for the confidence parameter had to be defined.

For this first investigation the same range of points (number of points for a correct/incorrect answer) was applied to all multiple-choice questions. Furthermore the possible values for the confidence parameter were limited to a small number of options. The resulting score *r* for a question was calculated using the following formula:

$$r = \begin{cases} p \cdot \frac{c}{100} \ \ldots \text{ correct answer} \\ -p \cdot \frac{c}{200} \ \ldots \text{ incorrect answer} \end{cases}$$

Formula 1

r … resulting score for this question

p … maximum score for this question

c … confidence defined by learner where $c = \frac{100}{c_p} \cdot u$ and $u \in \{0,1,\ldots,c_p\}$

c_p … number of predefined confidence intervals – 1

The combination of $p = 10$ and $c_p = 3$ was employed for all questions. This means that the confidence parameter was presented to students as a range from 0 % to 100 % in 3 steps of 33% each (see Table 1). The maximum score for a correct answer is 10 points, the minimum score for an incorrect answer is -5 points.

Confidence	Interpretation	Points if correct	Points if incorrect
100 (%)	I know it	10	-5
66 (%)	I am not sure, but I think I know it	7	-3
33 (%)	I am not sure, but I assume something	3	-2
0 (%)	I did not know it and I have to guess	0	0

Table 1: Interpretation and usage of the confidence parameter

2.2.2 Test Groups

The 432 students registered for the final exam were split into two groups. The first group – control group (260 students) – did a traditional MCT without confidence assessment. The second group – experimental group (172 students) – did the MCT with confidence assessment. The students had not been informed about the difference between the two groups before.

Both tests consist of 10 questions automatically and randomly selected from the pool of 100 questions. For the traditional MCT (control group) +10 points were awarded for a correct answer and 0 points for an incorrect one. Hence the maximum total score in the control group was 100 points and the minimum total 0 points. (see Table 2).

For the test with confidence assessment +10 points were awarded for a correct answer with a confidence stated 100 % and -5 points were awarded for a wrong answer with a confidence stated 100 %. This leads to a points range of 150 points (see Table 2).

In order to be able to confirm that the two groups can be compared and that the approach to confidence assessment is valid the experimental group had been analyzed in different ways:

1. Experimental group without Confidence
   The total scores of the MCT with confidence assessment were also calculated using the conventional mechanism employed for the control group; that means that the confidence parameter is not used and correct answers are awarded with +10 points, and incorrect answers with 0 points. This makes it possible to verify that the two groups of students are equivalent in terms of knowledge and capabilities. The range of points is the same as of the Control Group: 100 points (see Table 2).
2. Experimental Group scaled
   The total scores of the MCT with confidence assessment were scaled down from range of points of 150 points to 100 points. This ensures to compaire the Experimental Group with Confidence directly with both the Experimental Group without Confidence and the Control Group. The total scores were scaled according to the following formula:

$$score_{conf,scaled} = \left[score_{conf} + \left(range_{conf} - range_{no\_conf}\right)\right] \cdot \frac{range_{no\_conf}}{range_{conf}}$$

Formula 2

This results in four distinct "data sets" that will be used throughout the remaining sections of this paper:

Type of Test	Minimum Score	Maximum Score	Range of Points
Experimental Group with Confidence	-50	100	150
Experimental Group without Confidence	0	100	100
Experimental Group scaled	0	100	100
Control Group	0	100	100

Table 2: Points ranges and maximum scores for the tests.

The total scores of all students' exams are represented as percentage of the maximum score. Furthermore it must be pointed out that students were not informed about the influence of the confidence parameter on their results and the fact that the default value of the confidence parameter was set 100 %.

2.2.3 Research Questions

The following research questions are addressed in this investigation:

1. Is there a difference between the results of the MCT with confidence assessment and the results of the Control Group?
2. Is there a difference between the scaled results of MCT with confidence assessment and the results of the Control Group?
3. Is there a relationship between the confidence parameter and the correctness of the answers?

3 Result

3.1 General findings

In general the results are quite satisfying. More than 60 % of all students achieved a total score of more than 80 % of the maximum score. Less than 10 % had a total score of 50 % or less.

The total scores and the distribution of the Control Group and of the Experimental Group without Confidence are very similar (see Figures 2 and 3). This finding confirms that the two groups are equivalent and can be compared. A relevant difference can be found in the 91-100 % range: While almost 35 % of the Control Group are part of this cohort only between 27 and 30 % of the Experimental Group with Confidence can be found in this range. This issue will have to be considered when answering the research questions.

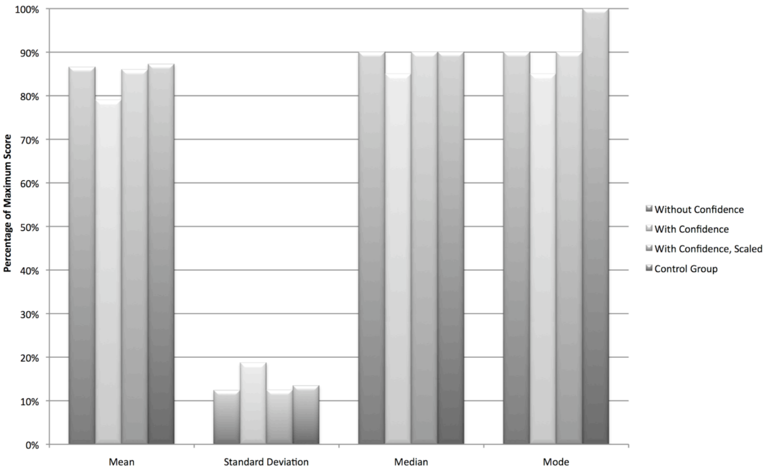

Figure 1: Characteristic statistical values.

The characteristic statistical values in Figure 1 illustrate the mean, median and mode values as well as the standard deviation of the four "data sets". The key finding that can be derived from this chart is that the four data sets have similar statistical properties. Mean and median values are almost identical, which indicates that the total scores of the Experimental Group scaled, the Experimental Group without Confidence and the Control Group are similar.

However there is also a notable exception. The Experimental Group with Confidence yielded total scores whose mean, median and mode values were between 5 and 10 % lower than the total scores of all other data sets. This is a very strong indicator that the use of the confidence parameter results in lower scores and in slightly worse grades (see also below).

Another finding is that the mode value of the Control Group is higher than the one of the other data sets.

3.2 Research Question 1

Is there a difference between the results of the MCT with confidence assessment and the results of the Control Group?

Figure 2 depicts the distribution of results in 10 % steps. It can be seen that the percentage of bad results (total scores of less than 40 %) is higher for the Experimental Group with Confidence than for the Control Group. While the total scores of the 60-70 % range occur more frequently in the Experimental Group with Confidence, total scores of the 70-80 % range occur significantly less often. This let assume that compared to the Control Group the results in the 70-80 % range frequently deteriorated by about 10 %. Scores in both the 80-90 % and the 90-100 % ranges occur less often in the Experimental Group with Confidence than in the Control Group. The Control Group is slightly "better"; it produced more results in the 90-100 % range than the Experimental Group without Confidence. The Experimental Group with Confidence yielded even less results in this range.

Analyzing the distribution of results in 20 % steps (Figure 3) it is obvious that these graphs are very similar for the four data sets. Total scores of 40 % or less occur more frequently in the Experimental Group with Confidence than in the Control Group, whereas total scores of more than 60 % occur more often in the Control Group.

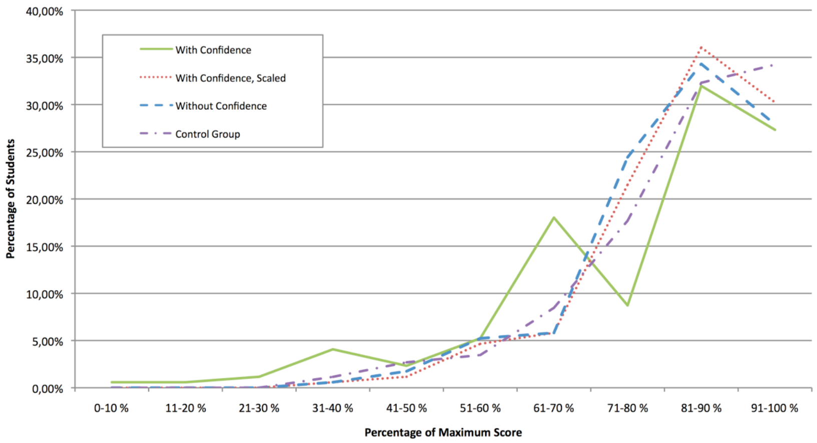

Figure 2: Distribution of total scores achieved by students; in 10 % steps.

The conclusion is that the use of confidence assessment in MCTs has an impact on the result of the MCT. Worse results (less than 40 %) occur more often, whereas good results (more than 70 %) occur significantly less often. Therefore it can be said that total scores for MCTs with confidence parameter are lower.

3.3 Research Question 2

Is there a difference between the scaled results of MCT with confidence assessment and the results of the Control Group?

Figure 2 shows that the results for the Experimental Group scaled are in almost all cases better than the results of the Control Group. The results of the Experimental Group scaled occur less often in the range below 50 %, while they occur more often in the range above 70 %. This finding is basically confirmed by Figure 3.

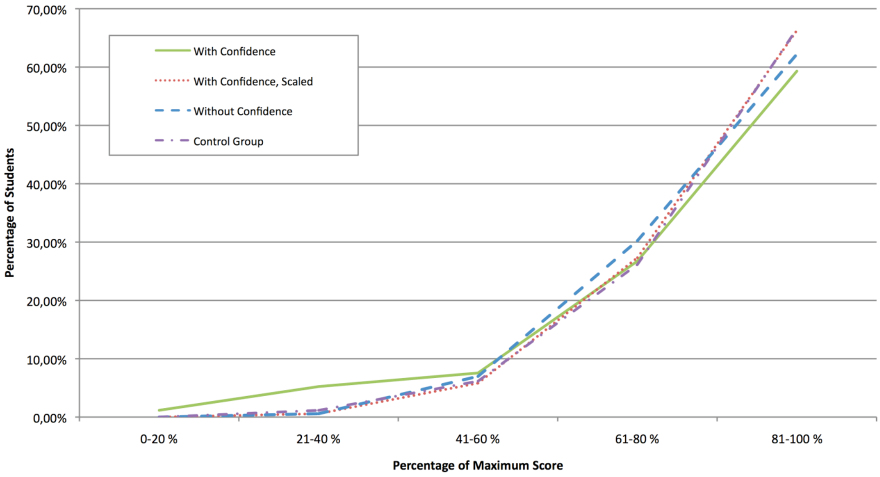

Figure 3: Distribution of total scores achieved by students; in 20 % steps.

The reason for the better results can be found in the function used for scaling the results (see Formula 2). As explained above the function scales the total scores from a [-50,+100] points range to a [0,100] points range. This means that -50 points were scaled to 0 points. When the equivalent function is applied to the scores of individual questions rather than the total scores, -5 points (100 % confidence, incorrect answer) were scaled to 0 points. Furthermore 0 points

(0 % confidence, wrong answer) were scaled to +2 points. This means that bad as well as good results from the Experimental Group with Confidence are “boosted” through the scaling function.

Therefore it can be concluded that the results of the Experimental Group scaled differs from both the results of the Control Group and the unscaled results of the Experimental Group with Confidence.

3.4 Research Question 3

Is there a relationship between the confidence parameter and the correctness of the answers?

Figure 4 shows the possible values of the confidence parameter on the horizontal axis and the percentage of all answers (solid line) as well as the percentage of incorrect answers (dotted line) on the vertical axis. The solid line in Figure 4 illustrates that for more than 90 % of all questions in all MCTs with confidence assessment students set the confidence parameter to 100 %. Only a small percentage of all questions were assigned a confidence of 66 %, 33 % or 0 %.

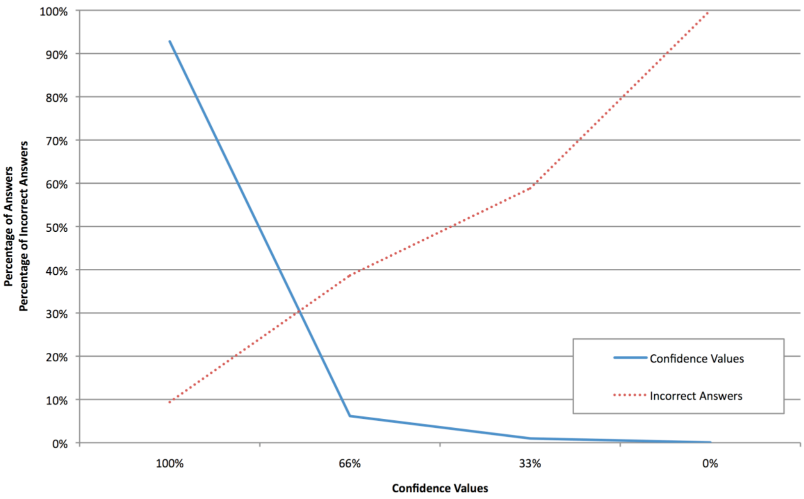

Figure 4: Distribution of confidence values and percentage of correct answers.

The dotted line displays that approximately 9 % of all questions that had been given a parameter of 100 % were answered incorrectly. For questions with a confidence parameter of 66 % wrong answers were provided in almost 40 %; questions with a confidence of 0 % were always wrong!

This is a very strong indication that the confidence parameter is proportional to the correctness of the questions.

4 Discussion

The traditional MCT yields more excellent results (total score of more than 90 %). The results of the Control Group lie roughly between the results of the Experimental Group with Confidence and the results of the Experimental Group scaled.

The results of the Experimental Group with Confidence are slightly worse than those of the Control Group, which is most likely due to the use of the confidence parameter. The nature of the scaling function (Formula 2) determines that the results of the Experimental Group scaled are slightly better than those of the Control Group.

The relationship between the confidence parameter and the percentage of wrong answers depicted in Figure 4 is highly relevant. Figure 4 also points out a further interesting aspect: most students set the confidence parameter to 100 %. Possible interpretations include:

- the students are not familiar with this kind of MCTs
- the test was too easy; so most students were really confident
- in other exams students have to be confident of their statements and answers, although they probably are not sure about their answers
- the students “gamble” and choose answers at their own risk
- because of the high time pressure many students might have neglected choosing a confidence parameter differing from the default value of 100 % and therefore simply have chosen the default value.

5 Conclusion and Outlook

Confidence testing seems to have an effect on the results of MCTs. In general MCTs with confidence assessment lead to slightly worse results. This might be due to the fact that guesswork is discouraged by means of the confidence parameter and the scoring scheme. Moreover there is an obvious relationship between confidence and correct answers.

A further decrease in guesswork can be expected when guessing really “hurts”. A different weighting scheme (e.g., +10 points for a correct answer, -30 points for an incorrect answer at 100 % confidence) that is previously announced will be focus of further experiments. Moreover different ranges of points for individual questions may also be implemented in order to take various levels of difficulty of questions into account. In further experiments students will be informed about the grading scheme before they do the test and the default parameter will be set 0 %.

It can be summarized that by adding a confidence parameter to a MCT the result of the MCT deteriorates. Further research will be necessary to gain further insight and to prove this theory.

References:

[1] Burton, R. F.: *Quantifying the Effects of Chance in Multiple Choice and True/False Tests: Question Selection and Guessing of Answers,* Assessment & Evaluation in Higher Education, Volume 26, Number 1, pp. 41-50, February 2001.

[2] Ebner, M.; Scerbakov, N.; Maurer, H.: *New Features for eLearning in Higher Education for Civil Engineering*, Journal of Universal Science and Technology of Learning, Volume 1, Number 1, pp. 93-106, 2006. Available online at http://www.justl.org/justl_0_0/new_features_for_elearning, Accessed June 29th, 2008.

[3] Levine, M. V., Rubin, D. B.: *Measuring the Appropriateness of Multiple-Choice Test Scores, Journal of Educational Statistics,* Volume 4, Number 4, pp. 269-290, Winter 1979. Original version dated December 1976 available online at http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/3a/16/3e.pdf, Accessed June 29th, 2008.

[4] Davies, P.: *There's no Confidence in Multiple-Choice Testing, ……,* Proceedings of the 6th International Computer-Assisted Assessment (CAA) Conference, Loghborough, 2002. Available online at http://hdl.handle.net/2134/1875, Accessed June 29th, 2008.

[5] Helic, D., Maurer, H., Scerbakov, N.: *Knowledge transfer process in a modern WBT system,* Journal of Network and Computer Applications, Volume 27, Issue 3, pp. 163-190, August 2004.

[6] Hutchinson, T. P.: *Ability, Partial Information, Guessing: Statistical Modeling Applied to Multiple-Choice Tests,* Rumsby Scientific Publishing, Adelaide, 1991.

[7] Roediger, H. L. III, Marsh, E. J.: *The Positive and Negative Consequences of Multiple-Choice Testing,* Journal of Experimental Psychology: Learning, Memory, and Cognition, Volume 31, Number 5, pp. 1155-1159, 2005. Available online at http://psych.wustl.edu/memory/Roddy%20article%20PDF's/Roediger&Marsh2005.pdf, Accessed June 29th, 2008.

[8] Gardner-Medwin, A. R.: *Confidence Assessment in the Teaching of Basic Science,* Association for Learning Technology Journal (ALT-J), Volume 3, pp. 80-85, 1995. Available online at http://www.ucl.ac.uk/~ucgbarg/tea/altj.htm, Accessed June 29th, 2008.

Author(s):

Josef Kolbitsch, Dipl.-Ing. Dr.techn.

Business Solutions Unit
Computing and Information Services
Graz University of Technology
Steyrergasse 30, A-8010 Graz

josef.kolbitsch@tugraz.at

Martin Ebner, Dipl.-Ing. Dr.techn.

Social Learning Unit
Computing and Information Services
Graz University of Technology
Steyrergasse 30, A-8010 Graz

martin.ebner@tugraz.at

Walther Nagler, Mag.

Social Learning Unit
Computing and Information Services
Graz University of Technology
Steyrergasse 30, A-8010 Graz

walther.nagler@tugraz.at

Nicolai Scerbakov, Prof. Dipl.-Ing. Dr.techn.

Institute for Information Systems and Computer Media
Graz University of Technology
Inffeldgasse 16, A-8010 Graz

nsherbak@iicm.edu