WORKING PAPER 11/2025 (ECONOMICS)

# A Continuous Scoring Function for Confidence-Based Marking using Multiple Choice Questions

**Niklas Karlsson and Anders Lunander**

ISSN 1403-0586

Örebro University School of Business
SE-701 82 Örebro, Sweden

# A Continuous Scoring Function for Confidence-Based Marking using Multiple Choice Questions

Niklas Karlsson[*]

*School of Business at Örebro University*

Anders Lunander[**]

*School of Business at Örebro University*

October 2025

**Abstract**

In most multiple-choice tests using confidence based marking (CBM), a discrete certainty scale is applied, often with three or four probability intervals of equal length. In this paper we derive a continuous certainty scale for CBM which we think circumvents the alleged complexity that would be inherent in a continuous scale. In our approach, the examinee, given a correct answer, is awarded the same number of points as her reported degree of confidence that her chosen alternative is the correct answer, i.e., the examinee's uncertainty is directly reflected in terms of the number of points achieved if the answer is correct. We test our continuous scoring scheme in an examination in basic statistics at our university. The results indicate that most students are quite good at assessing their confidence levels, but students tend on average to overrate their confidence for high levels of stated confidence and underrate their confidence for low levels of stated confidence.

Keywords: Multiple choice questions, confidence-based marking, scoring function

JEL: A22, C12

[*] Örebro University, School of Business, 701 82 Örebro, Sweden
e-mail: niklas.karlsson@oru.se
[**] Örebro University, School of Business, 701 82 Örebro, Sweden
e-mail: anders.lunander@oru.se (corresponding author)

# 1. Introduction

One testing format that increases the information obtained from responses on multiple-choice (MC) tests is certainty-based (confidence-based) marking (CBM). Given the alternative that the examinee chooses as her answer, she is asked to state how certain she is that the chosen answer is the correct answer. Depending on the examinee's stated certainty, she will be rewarded with different points if the answer is correct and potentially penalized if the answer is incorrect. If no alternative is chosen, the examinee obtains zero points. Although the method has a history dating back almost hundred years (see Esternacht, 1972 for a review), CBM is used to a lesser extent than the standard setting of multiple choice questions – in which just the number of correct answers is counted (Kanzow et al., 2023). Nevertheless, CBM is applied today in a variety of educational programs, e.g. medicine, law, foreign language, engineering (see Remesal et al., 2024 for a list of previous studies). A point of departure when formulating the scoring function is that the examinee has an incentive to honestly reveal her certainty, that is, the examinee's own personal probability, that the selected alternative is the correct answer, equals her reported certainty. The scoring function consists of two parts, where the first part specifies the achieved points in case the chosen alternative is correct, whereas the second part specifies the penalty if the examinee selects a wrong answer. Most of the literature analyzing the outcome from CBM tests, reports the use of a scoring matrix, in which the examinee faces a predetermined number of certainty or confidence levels which correspond to a finite number of probability intervals, in general four to five discrete intervals (e.g. Garder-Medwin, 2007; Barr & Burke, 2013; Remesal et al., 2024). For an incorrect answer, the penalty is often set to zero if the examinee reports the lowest level of confidence. The penalty then increases with a higher reported level of confidence, given that the answer is incorrect. The outcome from a row of empirical tests of CBM shows a high correlation between the number of correct answers and the achieved CBM score (e.g., Barr & Burke, 2013; Smrkolj et al., 2022; Wu et al., 2022).

To our best knowledge, very few studies have applied a continuous scoring scheme. In an empirical study comparing different measures of partial knowledge in multiple-choice tests, Ben-Simon et al. (1997) apply a continuous confidence marking scheme – ranging from "complete confidence" to "random guess" - where examinees were awarded or penalized with the same points as their attached confidence to the selected response. However, such a scoring scheme is not compatible with decision theory, i.e., it does not imply honest confidence revelation is the best strategy. In a theoretical contribution, Boldt (1971) derives a continuous scoring function which is quadratic in the reported confidence level where the examinee's

expected score is monotonically increasing as confidence increases. The expected score is maximized when the examinee's reported confidence coincides with the true probability of giving a correct answer. In order to make his method more tractable, Boldt transforms the recorded responses into a discrete rating system.

In this paper we derive and analyze theoretical implications of a continuous score function for CBM which is similar to that by Boldt (1971). The important difference is that in our approach, the examinee, given a correct answer, is awarded the same number of points as her reported degree of confidence that her chosen alternative is the correct answer. Hence, our method is transparent in the sense that the examinee's uncertainty is directly reflected in terms of the number of points achieved if the answer is correct. The penalty for an incorrect answer follows from two conditions, the first-order condition of the expected score being maximized when the examinee reports her true probability and also the condition that the expected score is to equal zero provided the examinee takes a chance on an answer when she is actually totally ignorant, yet honest about her ignorance. We compare the properties of our scoring scheme with that of the standard negative marking method, a commonly applied scoring scheme in examinations with multiple choice questions. In this scheme a correct answer is rewarded by one point, while an incorrect answer is punished by $\frac{1}{k-1}$ points, $k$ being the number of alternative answers. No answer gives zero points. This is a method directly comparable to the method developed in this paper.

We also report on the results from an examination in basic statistics at our university in which we apply our continuous scoring scheme. In addition, a method is developed to decompose total score on the exam into two parts, degree of knowledge and self-awareness.

We organize the rest of the paper as follows. In Section 2 the new scoring system is derived. Section 3 provides a theoretical comparison of characteristics between the new system and traditional scoring systems. In section 4 the practical implementation of the new system is described and analyzed. Section 5 ends the paper with conclusions and a discussion.

## 2. A Continuous Scoring Function

In this section a system for scoring multiple choice questions is modeled and analyzed theoretically. The system is modeled to give the examinee incentives to reveal her confidence in the answer, in terms of her perceived probability of her answer being correct.

### 2.1 Framework and derivation

Consider a multiple-choice question with $k \geq 2$ alternatives of which only one is correct. The examinee has the option of not answering the question by choosing no alternative, resulting in zero points on the question. The other option is to give an answer by picking one of the alternatives while reporting her level of confidence. Let $x_k$ be this reported level of confidence, where $\frac{1}{k} \leq x_k \leq 1$. The lower limit corresponds to the probability that a wild guess, because knowledge is completely lacking, will give a correct answer. If the answer is correct, she gets $x_k$ points, while a wrong answer yields $f_k(x_k)$ points. The reward by $x_k$ points is justified by the fact that the value of being correct increases by the level of significance. Also, to discourage the examinee from reporting an overestimate of the perceived level, $f_k(x_k)$ is expected to be decreasing. Thus, for an incorrect answer, the larger value of $x_k$, the larger point deduction the examinee receives.

We also define $p_k$ to be the true probability of giving a correct answer. To grasp the concept of $p_k$ we define $q_k$ as the perceived probability of the examinee's answer being correct, possibly different from the reported confidence $x_k$. The definition of $p_k$ is conditioned on $q_k$ in the following sense. Suppose there is a bank of many questions, from which the examiner draws a random sample of questions, constituting an exam. If the examinee were faced with all questions in the bank, she is assumed to have a perceived probability of her answer being correct to each question in this bank. We think of $p_k$ as the proportion of correct answers to questions where the examinee has a perceived level of confidence equal to $q_k$.

Moreover, we define the random variable $Y_k$, the achieved number of points on the question given an answer. Making use of our introduced notation, the probability function is illustrated in Table 1.

*Table 1. The probability function of* $Y_k$

| $y_k$ | $P(Y_k = y_k)$ |
|---|---|
| $f_k(x_k)$ | $1 - p_k$ |
| $x_k$ | $p_k$ |

Now, the expected number of achieved points, $E(Y_k)$, can be written as

$$E(Y_k) = f_k(x_k)\,(1 - p_k) + x_k p_k.$$

While $p_k$ is fixed at the time of the examination, the value of $x_k$, the reported confidence is not. It is chosen by the examinee, a choice that will affect the expected number of points. Since it is valuable information for the examiner to know the examinee's perceived probability $q_k$ , for $x_k$ to come close to $q_k$, we define $f_k(x_k)$ to be a function such that $E(Y_k)$ is maximized for $x_k = p_k$. We note that, although the examinee faced with this criterion has strong incentives to indicate the true perception about $p_k$, this does not necessarily mean $x_k$ will be close to the true probability $p_k$. The perception might be wrong, the examinee might for example overrate her confidence level. Nevertheless, to find such a function, that rewards examinees whose reported confidence $x_k$ is close to the true probability $p_k$, the first order condition is derived. We get

$$\left(\frac{dE(Y_k)}{dx_k}\bigg|_{x_k = p_k}\right) = f_k'(p_k)\,(1 - p_k) + p_k = 0.$$

Solving for $f'(p_k)$, we have

$$f_k'(p_k) = -\frac{p_k}{1 - p_k}.$$

Thus, the derivative of the function we are looking for is given by

$$f_k'(x_k) = -\frac{x_k}{1 - x_k},$$

and we can solve for the function $f_k(x_k)$ by integration. We get

$$f_k(x_k) = \int f_k'(x_k)dx_k = -\int \frac{x_k}{1 - x_k}\,dx_k = x_k + \ln(1 - x_k) + C_k.$$

To solve for the constant $C_k$, we consider an examinee with simply no knowledge about the question asked. If the examinee after all chooses to answer the question, it is reasonable for $E(Y_k)$ to equal zero

given an awareness and honesty of total ignorance of the question, i.e., given $x_k = p_k = \frac{1}{k}$. We get the restriction

$$\left(E(Y_k)\middle| x_k = p_k = \frac{1}{k}\right) = f_k\left(\frac{1}{k}\right)\cdot\left(1-\frac{1}{k}\right)+\frac{1}{k}\cdot\frac{1}{k} = 0.$$

Solving for $f\left(\frac{1}{k}\right)$, we have

$$f_k\left(\frac{1}{k}\right) = -\frac{1}{k(k-1)}.$$

Now,

$$f_k\left(\frac{1}{k}\right) = \frac{1}{k}+\ln\left(1-\frac{1}{k}\right)+C_k = -\frac{1}{k(k-1)},$$

and $C_k$ is given by

$$C_k = \ln\left(\frac{k}{k-1}\right)-\frac{1}{k-1}.$$

As can been seen above, $f_k(x_k)$ is not being defined for $x_k = 1$. i.e., a 100 percent reported confidence in the answer being correct. Note that $\lim_{x_k\to 1}(x_k+\ln(1-x_k)+C_k) = -\infty$. A system where some examinees, indicating a confidence of 100 percent in the answer being correct, are punished with an infinitely large point deduction if their answer is incorrect, would of course be difficult to implement. To address the problem, to begin with, we believe it is reasonable to limit the accuracy with which the examinee may indicate her perceived probability to two decimal places. Now, we are looking for values $f_k(1)$ such that an examinee being 99 percent confident, still would indicate $x_k = 0.99$ rather than $x_k = 1.00$ to maximize her perceived expected number of points on the question. The requirement

$$(E(Y_k)|x_k = p_k = 0.99) > (E(Y_k)|x_k = 1, p_k = 0.99)$$

can be written as

$$f_k(0.99)\cdot\ (1-0.99)+0.99\cdot 0.99 > f_k(1)(1-0.99)+1\cdot 0.99.$$

Solving for $f_k(1)$, we have

$$f_k(1) < f_k(0.99) - 0.99.$$

To keep it simple, we define $f_k(1)$ to be the largest (negative) integer satisfying the above inequality. Thus, provided $f_k(0.99) - 0.99$ is not an integer[1], we have

$$f_k(1) = floor(f_k(0.99) - 0.99).$$

To summarize,

$$f_k(x_k) = \begin{cases} x_k + \ln(1 - x_k) + \ln\left(\frac{k}{k-1}\right) - \frac{1}{k-1}, & x_k = a, a + 0.01, a + 0.02, \dots, 0.99, \\ floor(f_k(0.99) - 0.99), & x_k = 1, \end{cases}$$

where $a = \frac{ceiling\left(\frac{100}{k}\right)}{100}$.[2]

Also, a maximum is assured since

$\left(\frac{d^2E(Y_k)}{dx_k{}^2}\Big|_{x_k = p_k}\right) = -\frac{1}{1-p_k} < 0$ for $p_k < 1$.

This fulfilled condition finalizes the derivation part. Next, the properties of the proposed system will be examined.

## 2.2 Properties

Consider Figure 1 conditioning on the case where $k = 4$. It shows possible outcomes on the achieved number of points $Y_4$ given the reported confidence $x_4$. The upper graph shows number of points achieved if the answer is correct for different values of $x_4$, simply the identity function, while the lower graph depicts $f_4(x_4)$, the negative number of points achieved given an incorrect answer as a function of $x_4$. Conditional on an incorrect answer, there is an increasing negative effect of $x_4$ on $y_4$, with the lowest possible value of $-5$ attained at $x_4 = 1$.

[1] For feasible values of $k$ the value $f_k(0.99) - 0.99$ is not an integer.

[2] For some $k$, for example $k = 3$, $x = \frac{1}{k}$ is not allowed when limiting the accuracy of $x_k$ to two decimal places.

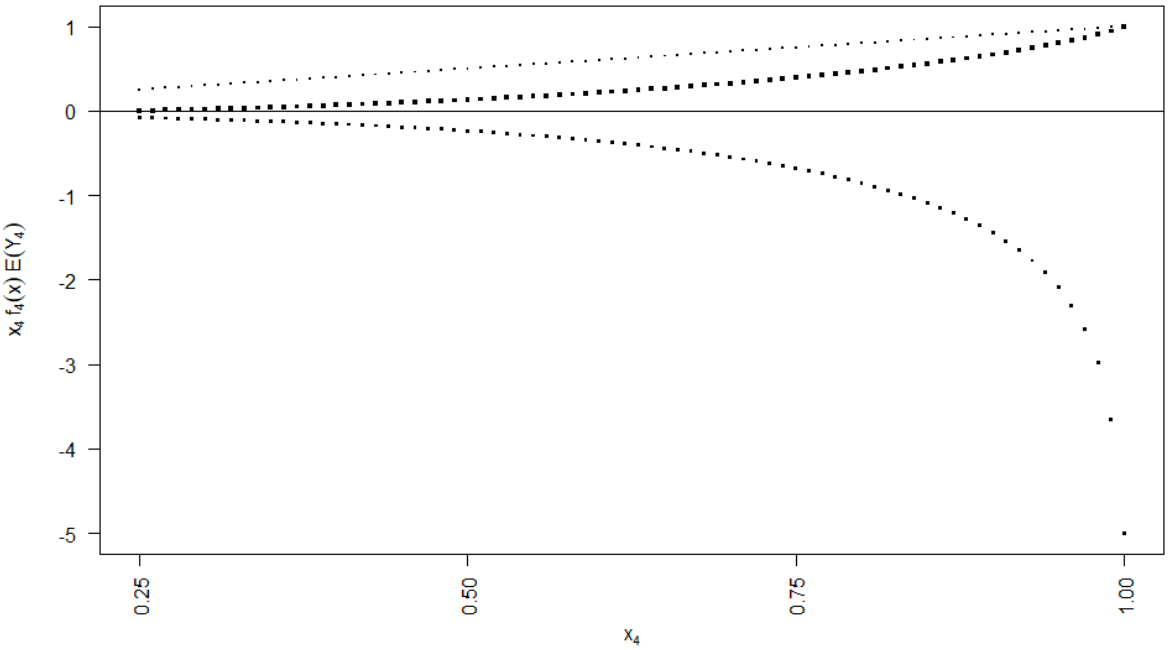


*Figure 1. Possible outcomes* $y_4$ *and expected number of achieved points* $E(Y_4|\, x_4 = p_4)$ *for different values of reported confidence level* $x_4$*. Achieved points given a correct answer are represented by small dots, points given an incorrect answer* $f_4(x_4)$ *by medium-sized dots and expected number of points by large dots.*

Similar patterns hold for other values of $k = 4$ and are therefore not depicted graphically, although in Table 2 values of $f_k(x_k)$ and $E(Y_k|\, x_k = p_k)$, respectively, are shown for selected values of $x_k$ for $k = 2, 3, 4$.

From Table 2, for values of $x_k$ corresponding to no knowledge of which alternative being correct, we also note a decrease in point deduction for incorrect answers with $k$. We have $f_2(0.5) = -0.5$, $f_3(0.34) = -0.17$ and $f_4(0.25) = -0.08$.

*Table 2. A selection of points* $\left(x_k, f_k(x_k)\right)$ *and* $E(Y_k|\, x_k = p_k)$*, respectively, for different values of* $x_k$ *for* $k = 2, 3, 4$.

| $x_k$ | $f_2(x_2)$ | $E(Y_2\|\, x_2 = p_2)$ | $f_3(x_3)$ | $E(Y_3\|\, x_3 = p_3)$ | $f_4(x_4)$ | $E(Y_4\|\, x_4 = p_4)$ |
|---|---|---|---|---|---|---|
| 0.25 | | | | | -0.08 | 0.00 |
| 0.30 | | | | | -0.10 | 0.02 |
| 0.34 | | | -0.17 | 0.00 | -0.12 | 0.04 |
| 0.40 | | | -0.21 | 0.04 | -0.16 | 0.07 |
| 0.45 | | | -0.24 | 0.07 | -0.19 | 0.10 |
| 0.50 | -0.50 | 0.00 | -0.29 | 0.11 | -0.24 | 0.13 |
| 0.55 | -0.56 | 0.05 | -0.34 | 0.15 | -0.29 | 0.17 |
| 0.60 | -0.62 | 0.11 | -0.41 | 0.20 | -0.36 | 0.22 |
| 0.65 | -0.71 | 0.18 | -0.49 | 0.25 | -0.45 | 0.27 |
| 0.70 | -0.81 | 0.25 | -0.60 | 0.31 | -0.55 | 0.33 |
| 0.75 | -0.94 | 0.33 | -0.73 | 0.38 | -0.68 | 0.39 |
| 0.80 | -1.12 | 0.42 | -0.90 | 0.46 | -0.86 | 0.47 |
| 0.85 | -1.35 | 0.52 | -1.14 | 0.55 | -1.09 | 0.56 |
| 0.90 | -1.71 | 0.64 | -1.50 | 0.66 | -1.45 | 0.67 |
| 0.95 | -2.35 | 0.78 | -2.14 | 0.80 | -2.09 | 0.80 |
| 0.98 | -3.23 | 0.90 | -3.03 | 0.90 | -2.98 | 0.90 |
| 0.99 | -3.92 | 0.94 | -3.71 | 0.94 | -3.66 | 0.94 |
| 1.00 | -5.00 | 1.00 | -5.00 | 1.00 | -5.00 | 1.00 |

Recall that to get a close estimate of the perceived probability through the reported confidence, our system rewards examinees where this reported

value is close to the true probability of a correct answer. Therefore, it is of interest to examine the difference in expected achieved points for various degrees of deviation from the true probability. In Figure 2 the expected achieved points as a function of $p_4$ is illustrated for various degrees of underestimation of one's ability to answer correctly. The dashed line corresponds to the case where the indicated perceived probability $x_4$ coincides with the true probability $p_4$, while the other four cases from the top down correspond to growing underestimation, here $x_4 - p_4 = -0.05, -0.1, -0.2, -0.3$.

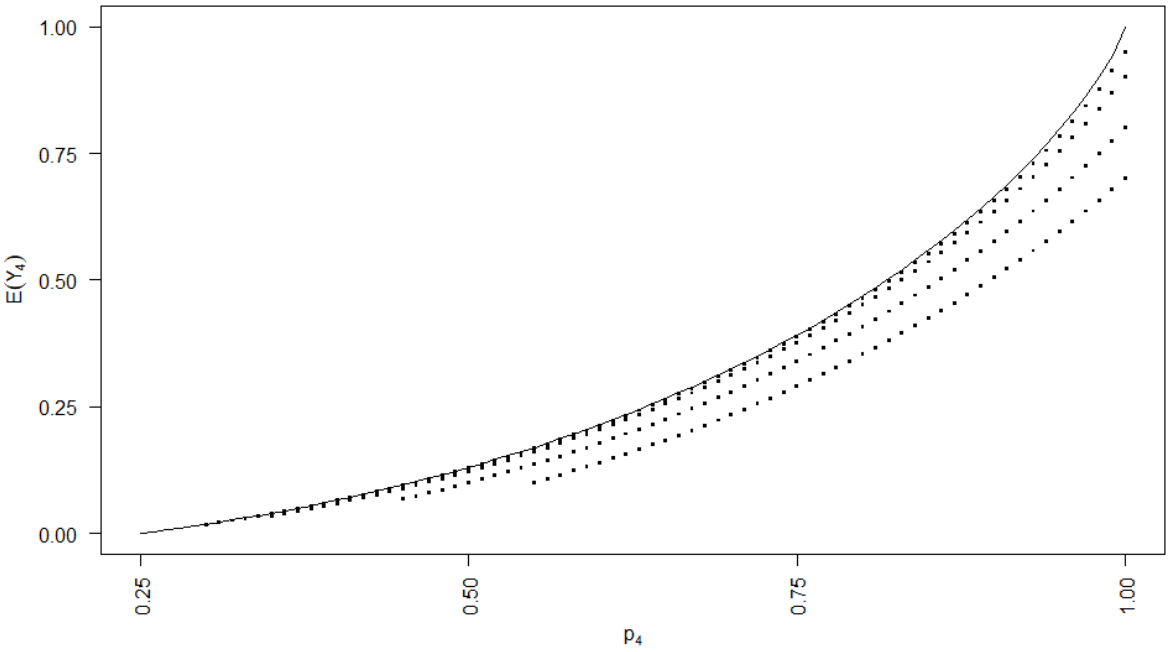


*Figure 2.* $E(Y_4|\, x_4)$ *as a function of* $p_4$ *for different negative deviations of* $x_4$ *from* $p_4$*. The dashed line corresponds to the case where* $x_4 = p_4$*, while the other four cases from the top down correspond to* $x_4 - p_4 = -0.05, -0.1, -0.2, -0.3$.

The ideal case in terms of maximizing the expected number of achieved points, i.e., $x_4 = p_4$, serves as the benchmark when comparing the effect of underestimating one's ability. For a given value of $p_4$, the expected number of points decreases as the underestimation increases. For example, for $p_4 = 0.9$ the expected number of points is 0.67 provided $x_4 = 0.9$ as well, while only 0.63 if $x_4 = 0.8$ and 0.50 if $x_4 = 0.6$. This effect increases with $p_4$. As $p_4 = 1$, the effect of underestimation on the expected number of points is simply the amount of the underestimate itself.

Figure 2 also reveals that given a certain amount of underestimation, the effect of $p_4$ on the expected number of points is positive and increasing, where the degree of convexity is the largest for the benchmark case where $x_4 = p_4$. Although not shown, similar patterns hold for other values than $k = 4$. For a certain value of $p_4$ it makes it possible to find the higher value of $p_4$ that would compensate for a certain underestimation of one's ability. For example, $x_4 = p_4 = 0.80$ gives $E(Y_4|\, x_4 = p_4 = 0.80) = 0.47$. To get the same number of expected points if the ability is underestimated by 0.3 units, it requires a value of $p_4 = 0.96$, i.e., an increase of $p_4$ with 0.16 units.

Figure 3 shows the expected achieved points as a function of $p_4$ for various degrees of overestimation of one's ability to answer correctly. As in Figure 2 the dashed line corresponds to the case where $x_4 = p_4$, while the other four cases from the top down correspond to increasing overestimation, here $x_4 - p_4 = 0.05, 0.1, 0.2, 0.3$.

As for the case of underestimation, for a given value of $p_4$, the expected number of points decreases with degree of overestimation. For example, for $p_4 = 0.8$ the expected number of points are 0.47 provided $x_4 = 0.8$ as well, while only 0.43 if $x_4 = 0.9$ and $-0.20$ if $x_4 = 1$. Again, this negative effect increases with $p_4$.

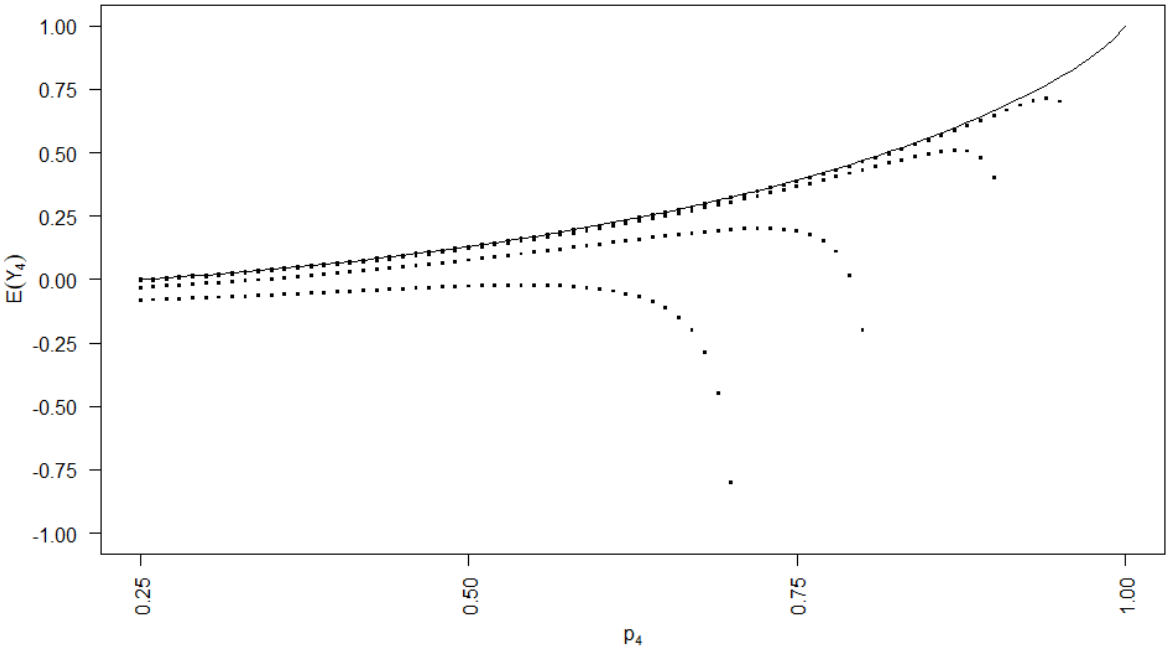


*Figure 3.* $E(Y_4|\, x_4)$ *as a function of* $p_4$ *for different positive deviations of* $x_4$ *from* $p_4$*. The dashed line corresponds to the case where* $x_4 = p_4$*, while the other four cases from the top down correspond to* $x_4 - p_4 = 0.05, 0.1, 0.2, 0.3$.

Contrary to the case of underestimation, the drop in expected number of points could be severe for large deviations. This is especially true for the case where $x_4$ is close to or equal to 1. To illustrate, consider an examinee indicating a perceived probability of 100 %, i.e., $x_4 = 1$, while the true probability of answering the question correctly is $p_4 = 0.7$. The negative effect of this misjudgment of ability, in terms of loss in expected points, is 1.13. This follows from $E(Y_4|\, x_4 = p_4 = 0.7) = 0.33$ and $E(Y_4|\, x_4 = 1,\ p_4 = 0.7) = -0.8$.

### 2.3 Comparison of properties: the CMB method vis-a-vis the standard negative marking method

The design of the CBM scoring system was derived from two conditions being satisfied. Firstly, the points rewarded given a correct answer should be increasing in the level of confidence and secondly, the system must have a built-in mechanism guaranteeing the student reveals her true level of confidence. The first condition is met by definition. While the second condition might not exactly be fulfilled, in the sense that the reported confidence not always coincides with the true perceived probability, we believe the system guarantees that the former probability is at least a good estimate of the latter one. In designing a system fulfilling the above conditions, we end up with a system having properties different from standard negative marking setting. Three such properties will be discussed.

First, unlike the negative marking method, the point deduction given an incorrect answer depends in our system on the examinee's level of confidence. The more confident the larger point deduction. This property is certainly needed to discourage the examinee from reporting an overestimate of the perceived level. Whether the property per se is desirable or not is hard to tell. There are arguments for both sides. On the one hand, one could argue that the deduction should be constant, regardless of the level of confidence, by claiming lack of knowledge is not relative. An answer that is wrong remains a wrong answer, no matter if the test taker believes she is correct or not. On the other hand, if you find it reasonable that a student, call her A, being for example 90 % confident giving a correct answer, should get more points than a student B, being 60 % confident, you might accept a smaller point deduction for B than for A for the following reason. Suppose, initially, the question contains $k = 2$ alternatives, only. Given an incorrect answer, A has indirectly assigned a low confidence on the correct alternative, only 10 %, while B has assigned a confidence at 40 %. This difference could motivate a smaller point deduction for B. The reasoning works for $k > 2$ as well, although not with the same clarity. Here, you could argue that B should get a smaller point deduction than A, since B has indirectly assigned a level of confidence *up to* 40% for the correct alternative, while A only *up to* 10 %.

Second, the dashed line in Figure 2 reveals for the case $x_4 = p_4$ that the effect on the expected number of points of an increase in the true probability of being correct is positive and increasing. Thus, given awareness of $p_4$, the system rewards a certain increase in $p_4$ from a high level more

than from a low level, in terms of increase in $E(Y_4)$. This result holds for other values than $k = 4$ as well. For the negative marking method the corresponding effect is certainly positive. However, the effect is, unlike our method, constant. The expected number of points corresponding to the extreme values $\frac{1}{k}$ and 1 of the true probability are the same for the negative marking method and our system, 0 and 1, respectively. Now, let $V_k$ be the expected number of points on the question of where negative marking is used. It is easily verified that[3]

$$E(V_k) = -\frac{1}{k-1} + \frac{k}{k-1}p_k,$$

where the marginal effect is $\frac{k}{k-1}$, positive and constant. For ease of comparison, Figure 4 shows both $E(V_k)$ and $E(Y_k)$ as functions of $p_k$ for the case $x_4 = p_4$.

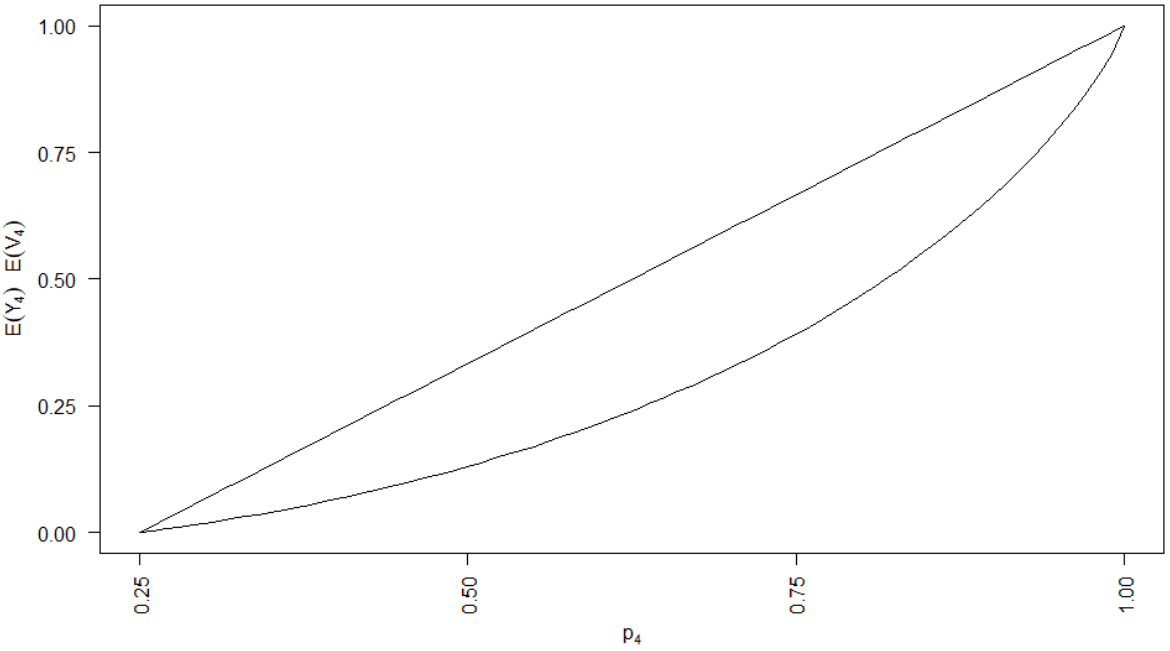


*Figure 4. $E(V_4)$ and $E(Y_4)$ as functions of $p_4$ for the case $x_4 = p_4$.*

[3] By definition, $E(V_k) = -\frac{1}{k-1} \cdot (1 - p_k) + 1 \cdot p_k$, provided the examinee always gives an answer if $p_k > \frac{1}{k}$.

The difference in curvature has some interesting implications. Since the effect on $E(Y_4)$ of an increase in $p_4$ from 0.25 to 0.26 is smaller than the corresponding effect of an increase in $p_4$ from 0.99 to 1.00, we value the latter increase in $p_4$ higher. This is not the case with negative marking, where both increases in $p_4$ are equally valued.

Third, provided that the reported confidence reflects the perceived probability, in our model a student is rewarded if the perceived probability is close to the true probability. This is in contrast with the negative marking method, where the expected number of points on an exam is independent of the perceived probability, given the true probability. To illustrate the implications of our model in this regard, we consider two fictional students, C and D, writing an exam with 20 questions with four alternatives on each question. While the perceived probabilities of C are assumed to coincide with the true probabilities, here equal to 0.90 for all questions, we assume D to underestimate her knowledge. The perceived probabilities here are assumed to be 0.80, and the true ones, as for C, equal to 0.90. Thus, both students have the same knowledge, whereas C has a better perception. Assuming the reported confidence coincides with the true perception about her knowledge, this quality is rewarded. C is expecting 13.4 points on the exam, while D is expecting only 12.7 points. Is it a reasonable difference? Most people would probably agree that it is a good quality to keep track of your own level of knowledge. A person who constantly overestimates her ability in various respects, risks making many mistakes. In the same way, a person who constantly underestimates her ability risks not getting much done, even though she has high capacity.

## 3. An Implementation of a Continuous Scoring Function for CBM

### 3.1 Design of an examination

To launch a confidence-based marking examination with a continuous scoring function, students on a 10-week-long course in Basic Statistics were invited to participate in a lab session. The students took a test based on the CBM method consisting of 20 questions, where the number of alternatives was set to four for each question. To achieve a high attendance at the lab session, all students who participated could add points to the traditional written examination with open questions at the end of the course. The number of points added was determined by how well they performed on the test. The 20 % best performing students were rewarded five points, the second best 20% of students were given four points, and so on. That is, even the 20 % worst performing students got one extra point. The differentiation of points was justified by

motivating the students to do their best on the test. While the questions on the traditional written exam aim to test the student's ability to choose a suitable statistical method to a given situation and to perform that method in a proper way, the questions on this quiz were, to a larger extent, designed to test the student's understanding of statistical concepts and interpretation of results.

Before the two-hour test was taken by the students, the teacher introduced the new scoring system for a period of 15 minutes to make sure the students knew what was expected from them. A table showing $f_4(x_4)$ for $x_4 = 0.25, 0.26, \ldots, 1.00$ was handed out to the students. They also got the information that those figures are calculated in such a way that the expected score on a certain question is maximized if the reported confidence coincides with the true probability of giving the correct answer. The students were also informed about the reasonableness of $0.25$ being a lower limit for the perceived probability.

### 3.2 Data

In total, 126 students participated in the session. It makes up 90 % of all registered students on the course. The students wrote the test by hand. Consider Table 3. It shows the descriptive statistics for five variables based on the 126 students taking the test. Among these questions there is certainly a variation in these variables.

*Table 3. Descriptive statistics for some variables based on the 126 students taking the test*

| *Variable* | $min$ | $q_1$ | $md$ | $q_3$ | $max$ |
|---|---|---|---|---|---|
| Proportion of responded answers | 0.55 | 0.80 | 0.90 | 1.00 | 1.00 |
| Proportion of correct answers | 0.1667 | 0.5500 | 0.6583 | 0.7500 | 0.9500 |
| Average reported level of confidence | 0.3727 | 0.5877 | 0.6722 | 0.7339 | 0.9185 |
| Confidence deviation | -0.280000 | -0.076414 | 0.009412 | 0.119474 | 0.487500 |
| Average score | -2.06 | 0.06895 | 0.2254 | 0.345 | 0.86375 |

Row one shows that half of the students respond to 90 % or more of the questions. From the second row the variation in *Proportion of correct answers* is quite large, where this proportion stretches from 16.7 % to 95.0 %, with a median value of 65.8 %. For the variable A*verage reported level of confidence* the location is about the same, yet a smaller variation.

The variable *Confidence deviation* on row four is defined as the difference between *Average reported level of confidence* and *Proportion of correct answers*. With a median value close to

zero, on average half of the students overestimate their capacity and the rest underestimate their capacity. The maximum value of 0.49 for the variable *Confidence deviation* is especially high, meaning that one student systematically overrates her level of confidence by 49 percentage units. In the lefthand tail of the distribution another student instead underrates her capacity by 28 percentage units on average.

Finally, for the variable *Average score* the median value is 0.225. At first glance, this score may seem low. But on reflection, it is not, but a consequence of the design of our scoring system. To put the score 0.225 into context, consider a student who is perfectly correct about her level of confidence being 61 % on each question. Her expected score is 0.225 on each question. The same holds for a student having a probability of 0.63 of being correct on each question, while overestimating her capacity by 11 percentage units on each question. The extreme minimum value of -2.06 is a result of a student having a low proportion of correct answers in combination with heavy overestimation of capacity.

### 3.3 Assessing confidence levels with individual ability curves

In this section it is described how we estimate the relation between the true probability of giving the correct answer to a question and the reported confidence level at the individual level. We call this relation the individual's ability curve for assessing confidence levels. The estimated ability curves serve two purposes. First, they help us to estimate the distribution among students of the true probability of giving the correct answer given a reported confidence level. Second, they can be used to estimate a particular student's potentially best score achievable if she would have had perfect self-awareness, that is, if her reported confidence level would have coincided with the true probability. This makes it possible to separate the two components making up the total score: knowledge about the course and self-awareness.

Our method is similar to that used in Wu et al. (2021), but instead of using a Rasch model to estimate the individual's ability as they do, we let the explanatory variable in the logistic regression be the log odds of the individual's reported confidence level. For a responded question, let $Y$ be a random variable taking the value 1 if the answer is correct, 0 otherwise. In addition, let $X$ be the reported confidence level. We think of the 20 questions comprising the exam as a random sample from a large bank of questions. For a given individual, $(X_1, Y_1), \dots, (X_n\ Y_n)$, where $11 \leq n \leq 20$, constitute a random sample where the dependent variable $Y_i \sim Bern(p_i)$, $i = 1, \dots, n$, $P(Y_i = 1) = p_i$ being modelled as a function of $X_i$ as

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 \ln\left(\frac{X_i}{1 - X_i}\right)\right)}{1 + \exp\left(\beta_0 + \beta_1 \ln\left(\frac{X_i}{1 - X_i}\right)\right)}.$$

The log odds of the reported confidence level are used to allow for perfect self-awareness, i.e., $p_i = X_i$, being a special case in the model, corresponding to the parameter restrictions $\beta_0 = 0$ and $\beta_1 = 1$.

We note that the log odds are not defined when $X_i = 1$, i.e., when the student claims to be 100 % percent sure of the answer. This is solved by assigning $X_i$ the value 0.9963 instead of 1 in those cases. The reasoning behind this number goes as follows. Theoretically, from section 2.1 an infinite point deduction should be the result of giving the wrong answer being 100 % confident. As this is not practically feasible, we choose, somewhat arbitrary, the punishment to be the largest negative integer such that an examinee being 99 percent confident, still would indicate a confidence of 0.99 rather than 1.00 to maximize her perceived expected number of points on the question. Upon examination it was found that it is beneficial in terms of expected points to indicate 100 % confidence provided the examinee is at least 99.26 % confident. The assigned value 0.9963 is the midpoint in the interval $[0.9926, 1]$.

We tried to estimate ability curves for all students in the sample by the maximum likelihood method. Estimates were successfully obtained for 124 students.[4]

### 3.4 Distribution of ability curves

Based on the estimated ability curves it is possible to estimate the distribution of the probability of giving a correct answer among the students conditioning on various confidence levels between 0.25 - 1.

Consider Figure 5. It shows the distribution of the probability of giving a correct answer among the students for various confidence levels. The five curves show the 5:th, 25:th, 50:th, 75:th and 95:th percentile, while the straight line represents perfect self-awareness, $P(Y = 1) = x$. The curve representing the median is below the straight line for high levels of confidence and above

[4] For the remaining two students, we failed due to convergence problems, most likely caused by too small a spread in the dependent variable. Instead, for these two students, simple linear probability models were estimated by ordinary least squares.

for low values, meaning that a typical student stating a high level of confidence overestimates her knowledge, while underestimating her knowledge for low levels. The gap between the 95:th percentile and the 5:th percentile is quite high for all confidence levels. This means that there are students who systematically overestimate as well as underestimate their knowledge largely for all levels of confidence.

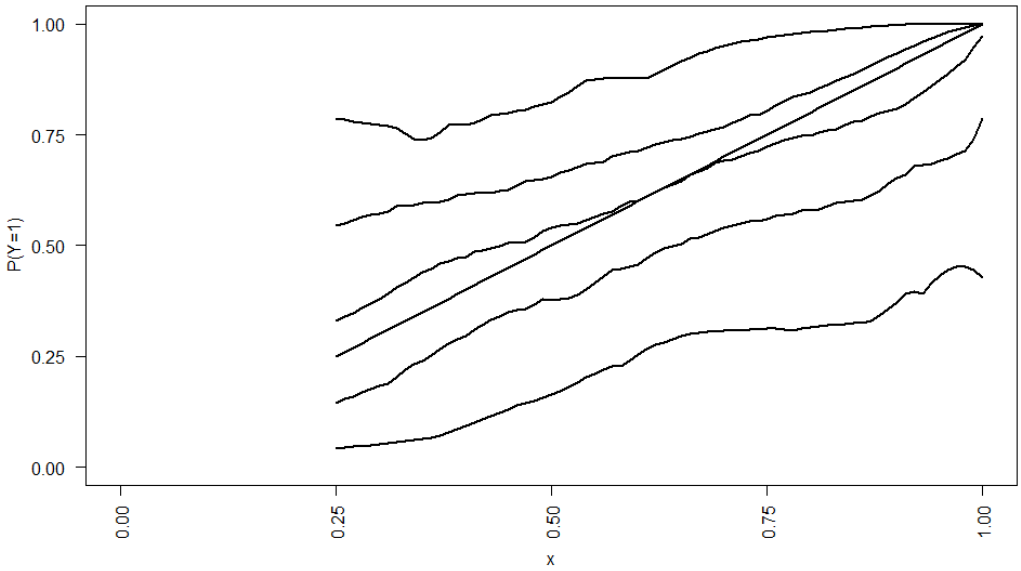


*Figure 5. Distribution of the probability of giving a correct answer among the students shown by the 5:th, 25:th, 50:th, 75:th and 95:th percentile for various confidence levels. The straight line represents perfect self-awareness,* $P(Y = 1) = x$.

In Figure 6 the median of the probability of giving a correct answer among the students, divided by sex is shown. Men are represented by the blue curve and women by the red curve. The figure indicates that men are better to assess their confidence level on average. The difference is especially notable for high values of $x$, where women overestimate their knowledge to a larger extent than men do.

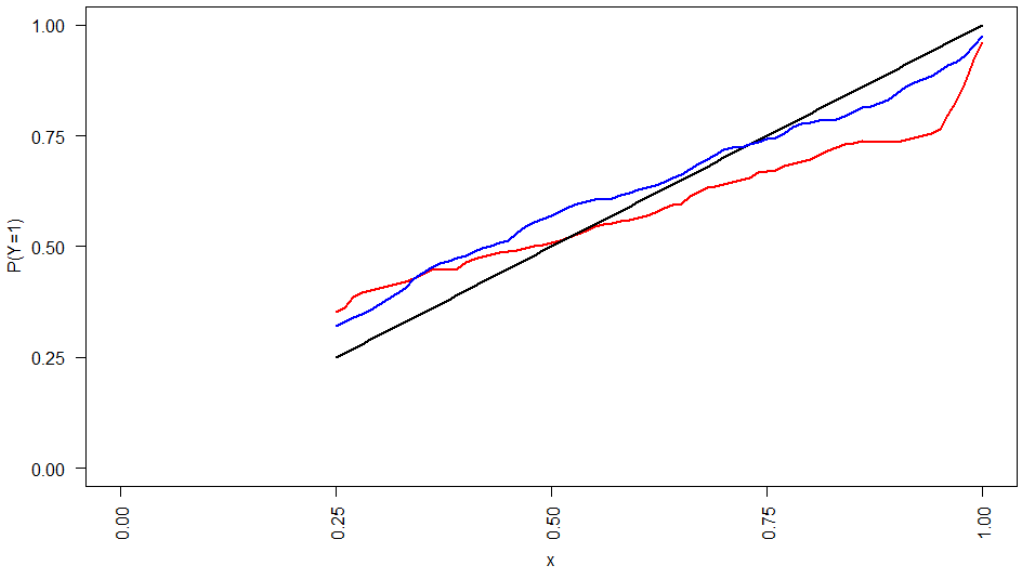


*Figure 6. The median of the probability of giving a correct answer among the students, divided by men (blue curve) and women (red curve). The black line represents perfect self-awareness,* $P(Y = 1) = x$.

### 3.5 Decomposition of total score

We decompose the total score into two components, reachable total score and self-awareness score. We define reachable total score as the total score that the student would have reached if she could have perfectly assessed her confidence level, i.e., if $P(Y_i = 1) = x_i$ for all $i = 1, \dots, n$. A student's value of the reachable total score is calculated through use of the student's estimated ability curve described in Section 3.4. To illustrate, suppose that according to the estimated ability curve, $P(Y = 1|X = 0.85) = 0.70$, i.e., the student overestimates her confidence level by 15 percentage units when stating a level of confidence equal to 85 %. If this student on a particular question states 85 % level of confidence, the contribution to the total score is 0.85 if the student is correct and -1.09 if not, according to Table 2. If the student had had perfect assessment, i.e., stating a confidence level of 0.70 instead of 0.85, the possible contributions are 0.70 and -0.55, considered to be possible contributions to the so-called reachable score. So, given the student's estimated ability curve, her $n$ indicated confidence levels and answers, correct or not correct, we get $n$ contributions, whose sum makes up the total reachable score.[5]

The self-awareness score is simply defined as the difference between the reachable total score and the total score. For example, a student having a total score of 6.2 and a reachable total score

[5] For some students and questions, we get $P(Y = 1|X = x) < 0.25$. In these cases, the contribution is set to 0, because with this information it is assumed that the student would not have answered to the question.

of 8.5 has a self-awareness score equal to -2.3, meaning that the student loses in total 2.3 points because of imperfect assessment skills.

Table 4 shows descriptive statistics for the variables total score, reachable total score and self-awareness score among all students.[6] The first line is simply the last line in Table 2 multiplied by 20, the number of questions. The median self-awareness score is moderately -1.08, while both the minimum value and the maximum value are salient. One of the students loses 41.67 points by bad assessments, while another student has perfect assessment skills with a self-awareness score equal to 0.00.

*Table 4. Descriptive statistics for the variables total score, reachable total score and self-awareness score among all students.*

| *Variable* | $min$ | $q_1$ | $md$ | $q_3$ | $max$ |
|---|---|---|---|---|---|
| Total score | -41.248 | 1.379 | 4.508 | 6.907 | 17.275 |
| Reachable total score | 0.378 | 3.829 | 5.771 | 8.040 | 17.275 |
| Self-awareness score | -41.67 | -2.70 | -1.08 | -0.32 | 0.00 |

In Figure 7 the variable reachable total score is plotted against the variable total score, where blue circles represent men, and red circles refer to women. Not surprisingly, we see a positive relationship between these variables with a Spearman's rank correlation equal to 0.90 (0.91 for men and 0.86 for women). The black line in the figure represents perfect assessment, where the total score coincides with the reachable total score. The figure reveals that a not negligible number of students has severe misperception about their confidence levels. At the same time, quite a large proportion of the students are close to this border. This observation is also mirrored in Table 4; 25 % of the students has a loss from imperfect assessment less than or equal to only 0.32 points.

[6] Note that it is not possible to get the figures on line 1 by summing the figures in line 2 and 3 column wise, because the figures in a particular column might refer to different students.

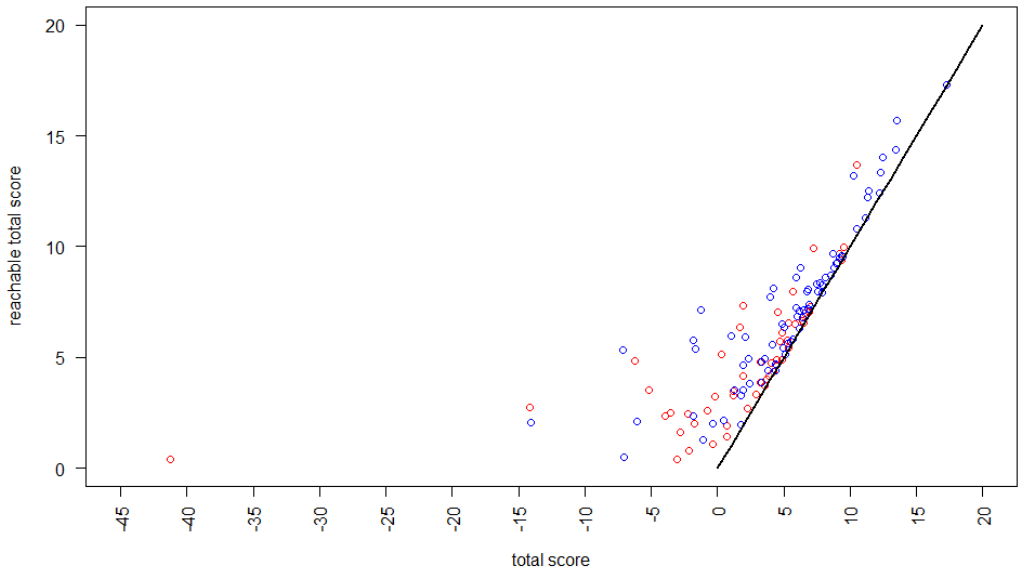


*Figure 7. Plot of reachable total score on total score. blue circles represent men, and red circles refer to women. The black line in the figure represents perfect assessment.*

Figure 8 shows a scatterplot of rank values of reachable total score against rank values of self-awareness score, where both ranking values are in ascending order. The linear relationship is not strong, yet we have a significantly positive correlation with the Pearson correlation coefficient being 0.327 (p-value equal to 0.00018, while no difference between the sexes was supported). Thus, on average, the higher self-awareness of your knowledge, the more knowledge.

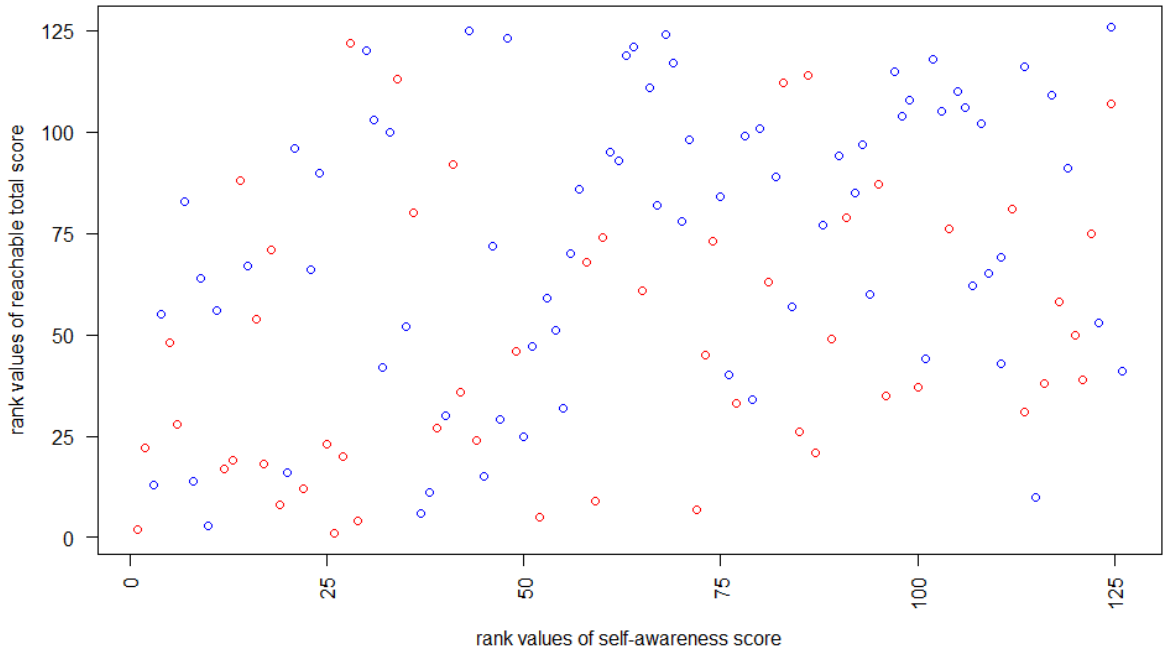


*Figure 8. Plot of ranking values of reachable total score on ranking values on self-awareness score. Blue circles represent men, and red circles refer to women.*

To further facilitate the comparison between men and women, in Table 5 we provide descriptive statistics for the variables total score, reachable total score and self-awareness score divided by men and women. From the table it is evident that men, on average, have performed better than women on the test in all respects, having distributions located more to the right for all three variables.

*Table 5. Descriptive statistics for the variables total score, reachable total score and self-awareness score divided by men and women.*

| *Variable* | $min$ | $q_1$ | $md$ | $q_3$ | $max$ |
|---|---|---|---|---|---|
| Total score (men) | -14.043 | 2.370 | 5.783 | 8.421 | 17.275 |
| Total score (women) | -41.248 | -0.249 | 3.409 | 5.353 | 10.485 |
| Reachable total score (men) | 0.500 | 4.846 | 6.973 | 8.948 | 17.275 |
| Reachable total score (women) | 0.378 | 2.732 | 4.568 | 6.556 | 13.697 |
| Self-awareness score (men) | -16.10 | -2.33 | -0.95 | -0.28 | 0.00 |
| Self-awareness score (women) | -41.67 | -3.40 | -1.24 | -0.40 | 0.00 |

Finally, a comparison with the classical scoring system using negative marking is made. For each student, the total score that the student would have had if the test had used the negative marking system instead is calculated. For those questions that the student has given an answer, the student is rewarded by 1 point if the answer is correct and punished by 1/3 points if the answer is incorrect. Also, just like in our system, the student gets 0 points if the question is unanswered. In this way we get 126 observations on the variable total score with negative marking. The Spearman's correlations between this variable and the variables total score and reachable total score are 0.83 and 0.93, respectively. The reason why the latter correlation is higher is explained by the downsized importance of self-awareness about confidence levels using the negative marking system. Moreover, the explanation for the latter correlation not being even closer to 1 is found in Figure 4, showing that knowledge, defined as expected number of points, is defined in two different ways for the systems.

## 4. Conclusions and Discussion

The CBM scoring scheme derived in this paper and implemented in an examination has the property of being strict monotonically increasing in the examinee's reported confidence. In contrast to various discrete certainty scales applied in previous empirical CBM studies, where the starting values for rewards and penalties are determined arbitrarily, albeit surely educationally meaningful, our scoring scheme rewards the examinee with the same score as her reported level of confidence, given a correct answer. The penalty, in case of an incorrect answer, is based on the reported confidence level such that the expected score on the question is maximized when the reported level coincides with the true probability of a correct answer.

The theoretical properties of our continuous scoring model are in line with many of those using a discrete confidence scale. The effect of the true probability of correctly answering the question on the expected score is positive and deviations of reported confidence from true probability have a negative impact on the expected score, especially for high positive deviations with a reported confidence close to or equal to 100 %.

Part of the analysis of the data obtained from the implementation rests on a method we develop to split the total score into two components, reachable total score and self-awareness score, where reachable total score is defined as the total score the student would have reached if she could have perfectly assessed her confidence level.

We observe that most students are quite good at assessing their confidence levels, a result found in previous studies (e.g., Barr & Burke, 2013; Foster, 2016; Wu et al., 2021). Also, a small significantly positive rank correlation is observed between reachable total score and self-awareness score, suggesting that those who know more also know more about what they know, although the relationship is not very strong. These results align with the work of Lichtenstein and Fischhoff (1977).

Our findings, that students on average tend to overrate their confidence for high levels of stated confidence and underrate their confidence for low levels of confidence is not consistent with the cumulative prospect theory and deviates from what is observed in Wu et al. (2021). It should though be pointed out that, in contrast to the response data analyzed in their study, there was no cutoff point set in our exam. Instead, the students were awarded points depending on their performance relatively others. Hence, the higher score obtained in the test, the higher the probability of ending up in the right tail of the distribution. The presence of a cutoff point in a CBM test, might very well change the objective of the examinee. The goal is now not to maximize expected score, but instead to try to obtain enough score to pass the test, given the attitudes towards risk. An examinee with little knowledge might overrate her confidence levels to maximize the probability of passing. To what extent a cutoff affects examinees' risk behavior in CBM tests is a question for further research.

Moreover, the result that men were better calibrated than women while women were more over-confident than men, contrast findings in the literature (e.g., McMurran et al., 2023). The deviating result might be due to the women in our study typically scoring lower than the men, which means the findings could have been different had we taken this into account.

There might be different motives for an examiner to use the proposed scoring system. One motive is that the examiner finds the way the scoring system measures the examinee's knowledge attractive. If this is the case, it has two main implications. First, the examiner prefers the effect of the true probability to correctly answer the question on the expected score being not only positive but also increasing. This contrasts with the negative marking model having a positive constant effect. Second, the examiner also finds the scoring system's property of rewarding self-assessment a plausible one.

Another motive for the examiner is to obtain valuable information about the distribution of confidence among the examinees on various questions in the exam, information that can be

useful in future teaching. However, if this is the main motive, the examiner should also be aware of the properties the system has on scoring at the individual level.

In addition, by using the proposed scoring system, it is possible at an individual level to estimate how well the students can assess their level of knowledge. Here is especially the method used to estimate ability curves useful. While the estimated ability curves are based on a maximum of 20 observations only, and it might be risky to generalize to other topics, we still believe that presenting these curves to students can provide some valuable insights to take with them.

In the course evaluation many students were positive about having had the possibility of trying the proposed system and that the design of the system was an instructive application of theories and concepts that they learned during the basic statistics course. While the experiment turned out well and the reactions from the students were positive, it is not obvious that this system would work for other student groups. We must keep in mind that the student group on which the experiment was conducted was unusually favorable in this context, since they were well familiar with statistical concepts such as probability and expected value, which the scoring system is based on.

## References


Barr, D. A., & Burke, J. R. (2013). Using confidence-based marking in a laboratory setting: A tool for student self-assessment and learning. *Journal of Chiropractic Education*, *27*(1), 21-26.

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*(1), 65-88.

Boldt, R. F. (1971). A SIMPLE CONFIDENCE TESTING FORMAT 1. *ETS Research Bulletin Series*, *1971*(2), i-18.

Echternacht, G. J. (1972). The use of confidence testing in objective tests. *Review of Educational Research*, *42*(2), 217-236.

Foster, C. (2016). Confidence and competence with mathematical procedures. *Educational Studies in Mathematics*, *91*(2), 271-288.

Gardner-Medwin, T. (2019). 12 Certainty-based marking. *Innovative Assessment in Higher Education: A Handbook for Academic Practitioners*, 141.

Gardner-Medwin, T., & Curtin, N. (2007, May). Certainty-based marking (CBM) for reflective learning and proper knowledge assessment. In *REAP International Online Conference on Assessment Design for Learner Responsibility*, 29-31.

Kanzow, A. F., Schmidt, D., & Kanzow, P. (2023). Scoring single-response multiple-choice items: scoping review and comparison of different scoring methods. *JMIR Medical Education*, *9*, e44084.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*. 20(2), 159-183.

Mcmurran, M., Weisbart, D., & Atit, K. (2023). The relationship between students' gender and their confidence in the correctness of their solutions to complex and difficult mathematics problems. *Learning and Individual Differences*, *107*

Remesal, A., García-Mínguez, P., Domínguez, J., & José Corral, M. (2024). Certainty-Based Self-Assessment in Higher Education: A Strategy for All? *International Journal of Advanced Corporate Learning*, *17*(3).

Smrkolj, Š., Bančov, E., & Smrkolj, V. (2022). The reliability and medical students' appreciation of certainty-based marking. *International journal of environmental research and public health*, *19*(3), 1706.

Wu, Q., Vanerum, M., Agten, A., Christiansen, A., Vandenabeele, F., Rigo, J. M., & Janssen, R. (2021). Certainty-based marking on multiple-choice items: Psychometrics meets decision theory. *Psychometrika*, *86*(2), 518-543.