WORKING PAPER 9/2025 (ECONOMICS)

Cutoff Point in Multiple Choice Examinations using Negative Marking or Number of Correct Scoring - An Analysis of Statistical Power

Niklas Karlsson and Anders Lunander

ISSN 1403-0586

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

Cutoff Point in Multiple Choice Examinations using Negative Marking or Number of Correct Scoring

An Analysis of Statistical Power.

Niklas Karlsson[*]

School of Business at Örebro University

Anders Lunander[**]

School of Business at Örebro University

August 2025

Abstract

Given the presence of a cutoff score in a multiple-choice questions test, a challenge for the test maker is to choose a scoring method maximizing the probability of a passing score for those with adequate knowledge given a prescribed risk of passing those with insufficient understanding. Within the environment of a true-false choice test, we analyze the statistical power of the standard method - one point if the correct answer is marked and zero otherwise – with that of the negative marking method - no answer results in zero points, a correct answer generates one point, and an incorrect answer is penalized by one point. Our comparison of power between the two methods indicates that the power is about equal when test taker exhibits a small variance in terms of her degree of confidence across the questions. For larger variance, the negative marking method is superior to the standard method. However, the more the test taker fails to capture her level of confidence, i.e., mis-calibration of knowledge, the lower statistical power of the negative marking. Which method has the highest power depends on the magnitude of mis-calibration. Underrating does not affect the power of NM as much as overrating.

Keywords: multiple-choice questions, negative marking, test of statistical power

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

Multiple-choice questions tests and true-false tests play a significant role in the examination of courses in, for example, academia. The main advantages with these types of tests – especially for larger groups of students - are that they are relatively easy to implement, and the submitted answers can be automated assessed, giving the test-takers instant results (see Brown 2001 for a discussion on the merits of multiple choice versus descriptive examinations).

The structure of multiple-choice and true-false tests varies, as well as the scoring method applied, see Lesage, Valcke and Sabbe (2013) and Kurz (1999) for overviews. In this paper we consider the format of true-false questions, where the most common scoring method is the number correct scoring rule. It implies that the test-taker is awarded one point if the correct option of the two possible is chosen and zero otherwise (Kurz, 1999). The scoring rule encourages the test-maker to always give an answer. This dominant strategy of always answering has both positive and negative consequences. On the one hand, the strategy is straightforward and should be easily understood by all test-takers. On the other hand, the test-maker cannot definitively discern whether a correct answer results from a fortunate guess or if the test-taker genuinely knows the answer. (Bar-Hillel et al., 2005 and Zapechelnyuk, 2015). This implies that a well-informed test taker could underperform due to unfortunate circumstances, while, conversely, a test taker with limited knowledge might exceed expectations simply by chance. Burton (2001) and Kubinger et al. (2010) quantifies the impact of luck in this scenario.

There are scoring methods designed to mitigate test score irregularities due to guessing, usually called formula scoring rules. One such rule is proposed by Holzinger (1924), who advocates a scoring penalty for wrong answers. In the true-false format he suggested that no answer would result in zero points, a correct answer would be awarded by one point, and an incorrect answer would be penalized with one point. The total score is thus the number of correct answers minus the wrong ones. This rule has the feature of giving an expected score of zero when the test-taker guesses an answer at random. Henceforth we will denote this rule by negative marking. A test-taker who prefers not to guess randomly is likely to refrain from answering questions she is unsure about. This approach can be advantageous for the test-maker as well, as unanswered questions can indicate a lack of knowledge. Nonetheless, formula scoring has faced its share of criticism. Budescu and Bar-Hill (1993) along with Bar-Hill et al. (2005) assert that formula scoring fails to address the issue of guessing. They argue that a risk-neutral test-taker who can

eliminate certain options but is uncertain about the remaining ones should still take a chance and make a guess.

The use of formula scoring can unintentionally promote guesswork, and when combined with the varying attitudes test takers have toward risk, it can lead to further complications. Those who are more cautious and hesitant to take risks may find themselves at a disadvantage, as their reluctance to guess could result in lower expected scores due to a higher rate of unanswered questions (Budescu and Bo, 2015; Burton, 2005; Choppin, 1988). For test makers, this cautious approach introduces bias, complicating their efforts to accurately assess a test taker's knowledge based on their responses (Akyol et al., 2022; Muijtjens, van Mameren, Hoogenboom, Evers, and van der Vleuten, 1999).

Espinosa and Gardeazabal, (2010) determine the best penalty amount for incorrect answers on multiple choice assessments by analyzing student behavior through a model rooted in item response theory. Their research indicates that the ideal penalty levels are quite high for students who act perfectly rationally but also remains significant for those who are not entirely rational. This means that, while the penalty tends to be unfavorable to risk-averse students, the impact of this is minor in relation to the measurement error it helps to avoid.

Another factor that might affect this bias is the degree of mis-calibration, i.e., the size of the discrepancies between confidence ratings and true hit rates (e.g., Lichtenstein and Fischhoff, 1977). To illustrate this point, let us consider a risk-averse test taker who chooses to answer only those questions she feels more than 60% confident about. In addition, suppose she underestimates her own abilities, in the sense that she would answer correctly about 70% of the questions that she believes she is 60% sure of. Consequently, the test maker is likely to underestimate the test-taker's knowledge even more than would be the case if she were merely risk-averse and perfectly calibrated in her judgments. By building on the framework established by Espinosa and Gardeazabal (2010) and incorporating factors like mis-calibration, Budescu and Bo (2015) reached a contrasting finding: incorporating penalties in the scoring for multiple choice assessments is harmful.

The test taker's strategy whether to guess or not to answer a question when being less than 100% sure, is also affected by the presence of a predetermined threshold score. A reasonable assumption is that a test taker primarily does what she can to achieve the threshold score rather than trying to maximize the expected score (Budescu and Bar-Hillel, 1993). For instance, if the passing score is set at 75%, and a test-taker knows that answering 80% of the questions she is

most sure about is likely sufficient for passing, the test taker may have little motivation to attempt the remaining questions. In fact, taking unnecessary risks with guesses could jeopardize her ability to pass the test altogether. An assessment of test takers' knowledge in such an environment is likely to give a misleading result. However, given the presence of a threshold score in a multiple-choice test, the test maker's objective when choosing scoring method - number correct scoring or negative marking - may not be to use the method generating the best picture of knowledge. Instead, the choice is focused on selecting the method maximizing the probability of a passing score for those with sufficient knowledge given a prescribed risk of passing those with insufficient understanding.

In this paper we address this question by providing a theoretical comparison of statistical power of two methods, number correct scoring (NCS) and negative marking (NM), used for true or false questions. Although power functions are calculated in Lord (1953) to discriminate between different multiple-choice tests with respect to the distribution of the item difficulty, NM is not considered in that paper and to our knowledge a formal comparison of power between the NM and NCS is yet not to be found in the literature. That said, there exists a row of studies examining test reliability based on empirical findings on actual test results in the context of multiple choice and true/false tests (e.g., Burton, 2002; Harden, Brown, Biran, Dallas Ross, Wakeford, 1976; Muijtjens, Van Mameren, Hoogenboom, Evers, Van Der Vleuten, 1999; Rowley GL, Traub RE., 1977). Recommendations for scoring methods are not uniform; they vary based on the specific data set being analyzed and the reliability measures employed.

Our definition of knowledge follows the approach found in Burton (2001 and 2002), where the exam questions take the shape of a random selection drawn from a vast pool of potential questions. To illustrate this concept, he introduces a knowledge parameter, defined as the proportion of questions in the question bank that the test taker maker would have answered correctly had he attempted them all. In our work we define this knowledge parameter as θ. The two systems are compared in terms of power to correctly detect an examinee with a sufficiently high value of θ. Motivated by the work of Budescu and Bo (2015), we also analyze the effect of various degrees of mis-calibration.

The rest of the paper is organized as follows. Section 2 presents the framework underlying the further analysis. In section 3, the design of the comparison of the two methods, NCS and NM, in terms of power, i.e., the probability of passing an exam given different values of the

knowledge parameter, is presented. The result of the comparison is provided in section 4. Finally, section 5 contains conclusions and a discussion.

2. General Framework

2.1 Two Probability Distributions

We consider a large number of N true or false questions in a bank, from which the examiner draws a random sample of n questions making up a test for a test taker to undergo. It is assumed that if the test taker was given the opportunity to see all N questions in advance, she would be able to indicate a perceived probability of giving the correct answer to each of them. Let w_i be this perceived probability on the i:th question. Furthermore, suppose these N values can be arranged in M groups, where $M \ll N$, such that within a certain group the values of w_i are the same, while no values are the same for separate groups. Let $p_1, \dots, p_M$ be the possible M different values and $g(p) = Pr(P = p)$ be the probability mass function of the associated random variable P, the test taker's perceived probability of giving the correct answer on a randomly drawn question from the bank. We also assume that the smallest possible value on P is 0.5, representing a coin flip as a result of no knowledge whatsoever of the answer to the question, while the largest possible value is 1. Consider Table 1 below, showing an example of $g(p)$ where $M = 3$.

Table 1. Example of $g(p)$

p	Pr $(P = p)$
0.5	0.2
0.75	0.2
1	0.6

Here, the test taker believes she knows 60 percent of the answers with certainty and 20 percent of the answers with a probability of 0.75, while she has no knowledge whatsoever of the answer to 20 percent of the questions.

If the test taker was to give an answer to all the questions in the bank, to each one of the M groups of questions, there is a fraction of correct answers. For the $j{:}\,th$ group this fraction is denoted by τ_j. If $p_j = \tau_j, \ \forall j \in \{1, \dots, M\}$ the test taker perfectly assesses her ability to give correct answers, we say that she is perfectly calibrated.

Suppose there are K unique such fractions, $\pi_1, \dots, \pi_K$, where $K \leq M$. To the experiment of randomly drawing a question from the bank and observing which one of the K groups the question belongs to, we can associate a random variable Π taking the values $\pi_1, \dots, \pi_K$ with corresponding probabilities

$$Pr(\Pi = \pi_k) = \sum_{j:\tau_j=\pi_k} Pr(P = p_j),\ k = 1, \dots, K.$$

The probability distribution is denoted by $f(\pi)$.

The two probability distributions $g(p)$ and $f(\pi)$ are identical when the examinee perfectly assesses her ability to give correct answers.[1] Based on Table 1 this case is shown in Table 2.

Table 2. The distribution $f(\pi)$ being identical to the distribution $g(p)$

π	Pr ($\Pi = \pi$)
0.5	0.2
0.75	0.2
1	0.6

The interpretation is as follows. If the test taker was given the opportunity to answer all of the N questions, she would in fact give a correct answer to all those questions she would identify as being 100 percent confident on, making up 60 percent of the questions in the bank. Moreover, for those 20 percent of the questions she would identify as being 75 percent sure of, she would have a hit rate of 75 percent. Finally, she is correct in identifying those 20 percent of the questions, where she has simply no knowledge and would give correct answers to 50 percent of those questions.

Instead, suppose the test taker is overrating her ability based on $g(p)$ in Table 1 in the following way. For the group of questions that she claims to be 75 percent confident on, the hit rate is only 65 %. Even worse, we see the same hit rate for the group of questions she would consider safe bets. In addition, she answers only correctly on 45 percent of those questions she believes she does not have a clue about, which might result from choosing an answer option in a systematically unfavorable way. The resulting probability distribution $f(\pi)$ is shown in Table 3. Besides, this is an example where $K < M$ ($K = 2$ and $M = 3$).

[1] Note that being perfectly calibrated is only a sufficient condition for the probability distributions being equal, not necessary. To illustrate, the two distributions in Table 1 and Table 2 would also be equal if $p_1 = \tau_2 = 0.5, p_2 = \tau_1 = 0.75$ and $p_3 = \tau_3 = 1$.

Table 3. An example where the test taker overstates her ability based on Table 1

π	$\Pr(\Pi = \pi)$
0.45	0.2
0.65	0.8

The distinction between perceived probability and true probability in this context is motivated by empirical findings on misconception about ability (Lichtenstein and Fischhoff, 1977). For further analysis on power to correctly detect an examinee with a sufficiently high value of θ, the distributions $g(p)$ and $f(\pi)$ are both conceptually vital.

2.2 The Knowledge Parameter

Based on the distribution $f(\pi)$ defined in the previous section we operationalize knowledge in the field in which the student is examined as the expectation of the random variable Π. Henceforth, this expectation is denoted by θ. Thus,

$$\theta = E(\Pi) = \sum_{\pi} \pi \Pr(\Pi = \pi)$$

where we label θ as the knowledge parameter. Possible values for θ are within the interval $[0, 1]$ and coincide with the proportion of correct answers the examinee would have had if she were to answer all questions in the bank.[2]

3. Comparison of Statistical Power

Next, we consider the introduced testing situation with a passing predetermined cutoff. In such a situation, a reasonable goal for the test taker is to maximize the probability of passing the exam, i.e., to reach at least the passing cutoff level with as high a probability as possible, conditional on her knowledge of the questions comprising the exam. For the test maker the goal could be to choose the scoring system that maximizes the probability of a passing score for those with adequate knowledge given a prescribed risk of passing those with insufficient understanding.

[2] θ might be smaller than 0.5 if the examinee is systematically misinformed.

If NCS is applied there is a dominant strategy to achieve this goal, simply to give answers to all questions. However, with NM the strategy varies, where the test taker's risk behavior depends on the situation.

To illustrate what is to be taken into account in the NM case, consider an exam comprising 20 true-false questions with a cutoff of 13 points, that is, at least 13 points are required to pass the exam. As a first scenario consider a test taker 100 percent confident of the correct answer to 5 of these questions and 50 % confident, i.e., totally ignorant, of the remaining 15 questions. Here, risk loving behavior is required to maximize the probability of passing the exam. It can be shown that all questions, except one of the questions that she doesn't know the answer to at all (50 % confident), should be answered.[3] In the second scenario we assume a test taker who is confident with 100 percent of the correct answer to 13 of these questions and 80 % confident with the rest. A risk-averse behavior does the trick here to maximize the probability of passing, since only those 13 questions she is completely confident with should be answered, although answering all questions would increase her expected total score on the exam.

In the latter scenario, if the examinee chooses the correct strategy to maximize the probability of passing, the test maker would most likely have severely underestimated the examinee's knowledge parameter θ. However, for this kind of testing situation, the objective for the test maker is not necessarily to estimate θ as precisely as possible. Instead, we assume the goal is to choose a scoring method that would result in a large probability for an examinee with a large (small) value of θ to pass (fail) the exam.

To formalize the discussion, we consider the examination as a hypothesis testing situation, where the examiner is to decide, given the information on the result of the exam consisting of n randomly drawn questions from the bank of N questions, whether the examinee has enough knowledge. We would like to test

$$H_0: \theta \leq \theta^*$$

against

$$H_1: \theta > \theta^*.$$

Thus, the test taker is considered to have enough knowledge to pass the course if and only if the value of θ is larger than a certain amount θ^*. At significance level α, the examiner decides

[3] In Appendix the probability model used for this calculation and all other probability calculations in this section are presented. The model is based on the so-called Poisson binomial distribution.

to pass the examinee if and only if the score on the exam is at least a certain number of points, t_0. The total scores for the NCS design and NM design, respectively, serve as competing test statistics.

3.1 Designing a test situation

To compare the two systems in terms of power we think of a test consisting of $n = 20$ randomly drawn true or false questions from a large bank of N questions. In this setting we would like to test

$$H_0: \theta \leq 0.75$$

against

$$H_1: \theta > 0.75$$

at a certain significance level α. Recall that a value of θ equal to 0.75 may arise from infinitely many distributions of Π. Below you find two such distributions.

Table 4. Two probability distributions of Π where $\theta = E(\Pi) = 0.75$

Distribution I		Distribution II	
π	$\Pr(\Pi = \pi)$	π	$\Pr(\Pi = \pi)$
0.5	0.5	0.75	1
1	0.5		

The distribution I means that the test taker is 100 % confident of the answer to a half of the questions in the bank and 50 % confident with the rest where, as previously mentioned, the latter confidence represents no knowledge at all in the true or false setting. A test taker with the distribution II is 75 % confident of the answer to all questions in the bank.

To determine a critical value t_0 to a certain prescribed significance level is not possible, except for a few values of α since the underlying test statistic, the total score on the exam, is discrete. It is also not possible to find two critical values, one for each system, corresponding to a common level of significance. It is essential for the comparison of the two systems to overcome these two problems. Therefore, we are considering a situation which, although not practically possible in a real situation, is theoretically plausible to be able to get an arbitrary desired common level of significance for both systems.

3.1.1. Critical values and randomization probabilities for the NCS method

Let the desired significance level be $\alpha = 0.25$. It can be shown that a test taker with a knowledge parameter θ equal to 0.75 has a probability of 0.415 to score 16 points or higher, where she gives an answer to all 20 questions. Thus, the critical value of 16 corresponds to a significance level of 0.415. It also can be shown that a critical value of 17 corresponds to 0.225. Therefore, 16 is too low a critical value, while 17 is too high for the desired significance level, and a critical value in between is obviously not possible. Now, consider the situation where the test taker meets a critical value of 16 with probability p_{16} and a critical value of 17 with probability $p_{17} = 1 - p_{16}$, henceforth called randomization probabilities, where

$$p_{16} = \frac{0.25-0.225}{0.415-0.225}.$$

The value 0.25 in the numerator corresponds to the desired significance level. It is easily verified that the probability is 0.25 for such an examinee to pass the exam prior to the randomization of the critical value to meet. The law of total probability yields the desired level of significance as

$$\alpha = Power(\theta = 0.75) = p_{16} \cdot 0.415 + (1 - p_{16}) \cdot 0.225 = 0.25.$$

The power for other values of θ than 0.75 is calculated in a similar way. For example, for an examinee with a value of θ equal to 0.85 it can be shown that the probability to score 16 points or higher is 0.830, while the probability to score 17 points or higher is 0.648. Therefore, we get the power as

$$Power(\theta = 0.85) = p_{16} \cdot 0.830 + (1 - p_{16}) \cdot 0.648 = 0.672.$$

This means that for an examinee with a value of θ equal to 0.85 the probability is 0.672 to pass the exam prior to the randomization of the critical value.

3.1.2 Critical values and randomization probabilities for the NM method

Unlike the NCS method, there is no unique power for a given value of θ for the NM method. It appears there are three aspects to consider when relating power to the parameter θ. First, the power for a given θ depends on the distribution of Π. For example, if the significance level is 0.25, conditioning on distribution I in Table 4, the probability of passing the exam is lower than

0.25 for a test taker with the distribution II than with the distribution I. It can be shown that the probability is lower than 0.25 for any other distribution as well. Therefore, we will base our calculation of critical values and randomization probabilities for the NM method on distribution I in Table 4.

Second, if the two probability distributions $g(p)$ and $f(\pi)$ differ, the power is smaller compared to the case where the distributions are identical. Therefore, we will also base our calculation of critical values and randomization probabilities for the NM system on $g(p)$ and $f(\pi)$ being identical.

Third, when comparing the two test situations in terms of power, we assume the test taker's objective is to maximize the probability of passing the exam, i.e., to get at least a certain number of points on the exam. For an exam using NCS, it is not difficult to make reasonable assumptions about the test taker's behavior consistent with that objective. She will simply give answers to all questions, irrespective of perceived confidence level. When the test taker is facing an exam using NM, it is by no means evident that the test taker has the cognitive skills to act consistently with the objective of maximizing the probability of passing. To illustrate the complexity a test taker may face, suppose that 13 points is the cutoff to pass an exam with 20 questions. Moreover, suppose the student is 90 % confident of all 20 questions. To maximize the probability of passing in this case, the test taker should give answers only to 19 questions, even though the expected number of points increases if she gives answers to all questions. However, for all power calculations in this paper, we will assume the test taker always is capable to pick the optimal strategy of maximizing the probability of passing given the confidence level on the questions, although this might not always be the case in practice. Therefore, the assumption most likely means that the power calculations for the NM test situation provide an upper limit.

Taking these three aspects into account, to find critical values and randomization probabilities for the NM test situation, we condition on a test taker having the distribution I in Table 4, where this distribution also coincides with her distribution of perceived probabilities, $g(p)$. We also stipulate that she has the skills to maximize the probability of passing given her perceived probability of correct answers to the questions.

For a fair comparison of the two methods, we work with the same significance level $\alpha = 0.25$. For the NM method, the test taker, with the characteristics referred to above, has a maximum probability of 0.343 to score 13 points or higher and a maximum probability of 0.241 to score

14 points or higher. Thus, a critical value of 13 corresponds to a significance level of 0.343, while a critical value of 14 corresponds to 0.241. Randomization probabilities are now determined in a similar way to how we determined the corresponding probabilities for the NCS method. We consider the situation where the test taker meets a critical value of 13 with probability p_{13} and a critical value of 14 with probability $p_{14} = 1 - p_{13}$, where

$$p_{13} = \frac{0.25-0.241}{0.343-0.241}.$$

Now, the probability is 0.25 for such a test taker to pass the exam prior to the randomization. We get

$$\alpha = Power(\theta = 0.75) = p_{13} \cdot 0.343 + (1 - p_{13}) \cdot 0.241 = 0.25.$$

The power for test takers with other characteristics is calculated accordingly. For example, suppose a test taker is 100 % confident of the answer to 70 % of the questions in the bank and 50 % confident of the answers on the rest of the questions. This results in a value of θ equal to 0.85. Also, suppose her perceived probabilities coincide with the corresponding true probabilities. Then, it can be shown that the probability to score 13 points or higher is 0.864, while the probability to score 14 points or higher is 0.755. Therefore, the power is given by

$$Power(\theta = 0.85) = p_{13} \cdot 0.864 + (1 - p_{13}) \cdot 0.755 = 0.765.$$

Thus, for this test taker, the NM method is preferred over the NCS method since the power is higher, 0,765 compared to 0,672. However, this is not always the case. In the next section we will examine and compare the two methods' statistical power for different distributions of Π and under different assumptions about the deviation of $g(p)$ from $f(\pi)$.

4. Results

In this section we present the results from our comparison of statistical power. We divide the comparisons into two cases. First, it is assumed that perceived probabilities coincide with true probabilities. Second, we allow for a deviation of perceived and true probabilities.

4.1 Perceived probabilities coincide with true probabilities

Consider Figure 1. Here, it is assumed that the two distributions $g(p)$ and $f(\pi)$ do not differ. The red curve shows the power of the NCS method or different values of θ. For this method, the power for a given value of θ is not affected by the appearance of $f(\pi)$. The power increases in θ, and takes the significance level 0.25 at $\theta = 0.75$. The green curve shows the power for different values of θ, where Π can take on values 0.5 and 1, only. Thus, for this situation the variance of Π is high, given the value of θ. For this case, the NM method outperforms the NCS method. For $\theta > 0.75$, i.e., values of θ representing a knowledge level that we believe is enough, the probability of passing is higher for the NM method, holding the significance level constant at 0.25 for both methods. Also, for low levels of knowledge the probability is lower to pass the test making use of the NM system.

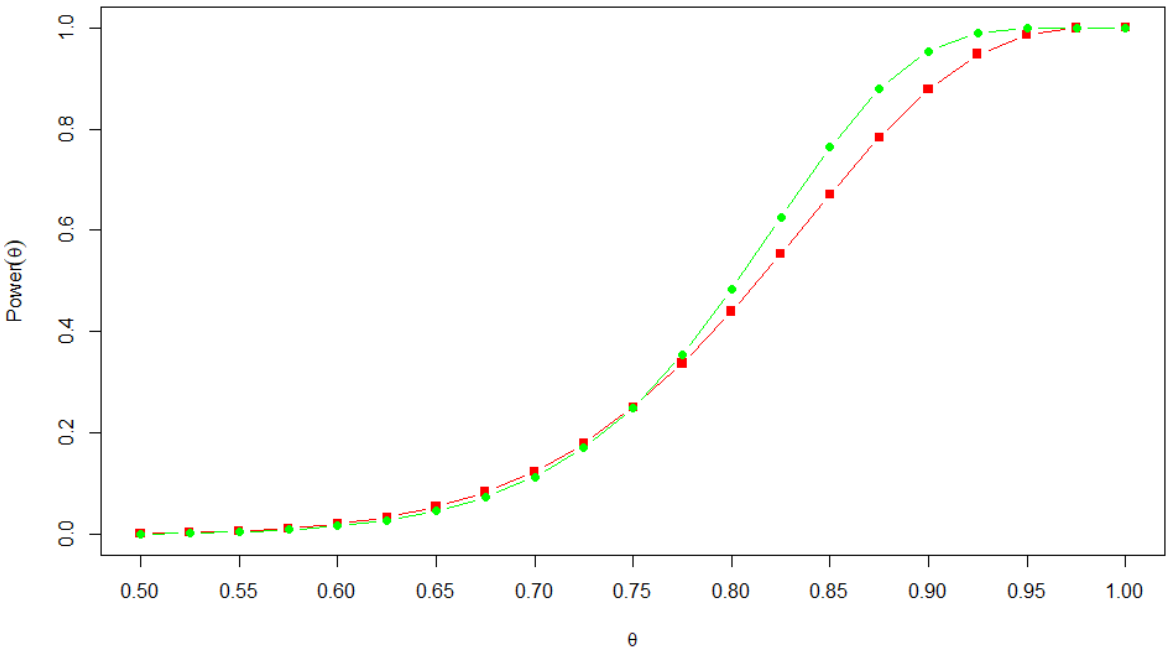

Figure 1. Power comparisons of NCS (red line) with NM (green line) based on an exam with 20 true or false questions sampled from distributions of Π which only assumes values of 0.5 and 1. Perceived probabilities coincide with true ones. The significance level is set at 0.25.

The difference in power is especially high for values of θ around 0.85-0.90. This result can be intuitively explained as follows. A test taker with a value of θ equal to 0.875, where the difference seems to peak, has complete knowledge of 75 % of the questions in the bank. This means that the expected number of such questions the test taker will face on the exam is 15 and, using the NM method, it is likely that the randomized number of such questions will exceed the

critical values, 13 or 14. Thus, it is likely that the test taker will pass the exam, without depending on sheer chance for the answers to questions she is unfamiliar with, as she will refrain from answering to those types of questions to increase her probability of passing. For the NCS method the critical values are higher, 16 or 17 depending on the randomization. This means that the number of questions the test taker will know for sure on the exam, is most likely below the critical value. The risk is not negligible that the achieved score on the questions the test taker will guess at, is not high enough for her to pass the exam.

Turning to Figure 2, the red curve is the same as in Figure 2 showing the power for the NCS-method. The green curve shows the power of the NM method for the case when all questions in the bank have the same confidence level, for a given value of θ. For example, a value of θ equal to 0.75 means that the test taker is 75 % confident on all questions. Unlike the situation displayed in Figure 1, here the variance of Π is low, in this case zero, given different values of θ. The two power curves do not differ much in this case. The reason is that the test taker, also with the NM method, will answer pretty much all questions to maximize the probability of passing.[4] The reason why the power of the NM method is even somewhat lower than the NCS method in this case is, as discussed above, that the critical values and the randomization probabilities are based on the case displayed in Figure 1, where the test taker knows a fraction of the questions with certainty.

[4] She will answer 19 questions if the critical value is 13, all 20 questions otherwise.

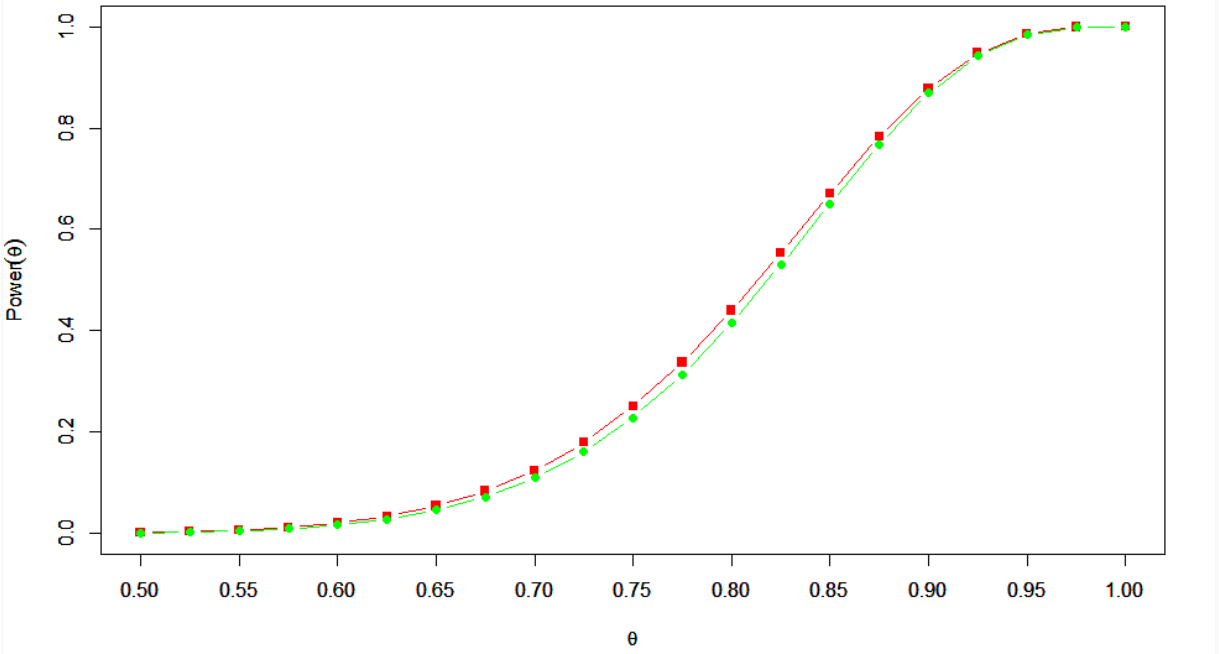

Figure 2. Power comparisons of NCS (red line) with NM (green line) based on an exam with 20 true or false questions sampled from distributions of Π which each assumes one value, only. Perceived probabilities coincide with true ones. The significance level is set at 0.25.

4.2 Perceived probabilities deviate from true probabilities

Consider Figure 3 showing the effect on power of overrating one's knowledge. All curves correspond to a distribution where P takes on the values 0.5 and 1, only. Thus, the test taker's perception is either no knowledge or total knowledge. The red and green curves are the same as in Figure 2. These two curves serve as baselines and correspond to the case where the examinee correctly perceives that she knows the answer with certainty of a fraction of the questions in the bank, while having no knowledge of the rest.

To illustrate the effect of overestimating one's knowledge two more curves are added to the figure, a blue and a yellow curve. These curves show the power of the NM method where the perceived probabilities deviate from the true ones. The blue and the yellow curves correspond to the cases where the test taker has a hit rate of 95 % and 90 %, respectively, on those questions that she thinks she knows the answer with certainty. This means we are considering a situation where the test taker overrates her knowledge. The drop in power is seen to be quite substantial, especially for values of θ in the range of 0.80 – 0.90. Under these circumstances the NM method

is outperformed by the NCS method. There is an intuitive explanation for this loss in power. For values of θ in this range, it is not unlikely that the number of questions the examinee thinks she knows will exceed the critical value, while the actual number does not. For such a situation, she will not give answers to the other questions and thereby fail the exam. In case the perceived probabilities are the same as the true ones, the examinee would have taken a chance on some of the other questions and thereby increased her chances to pass the exam.

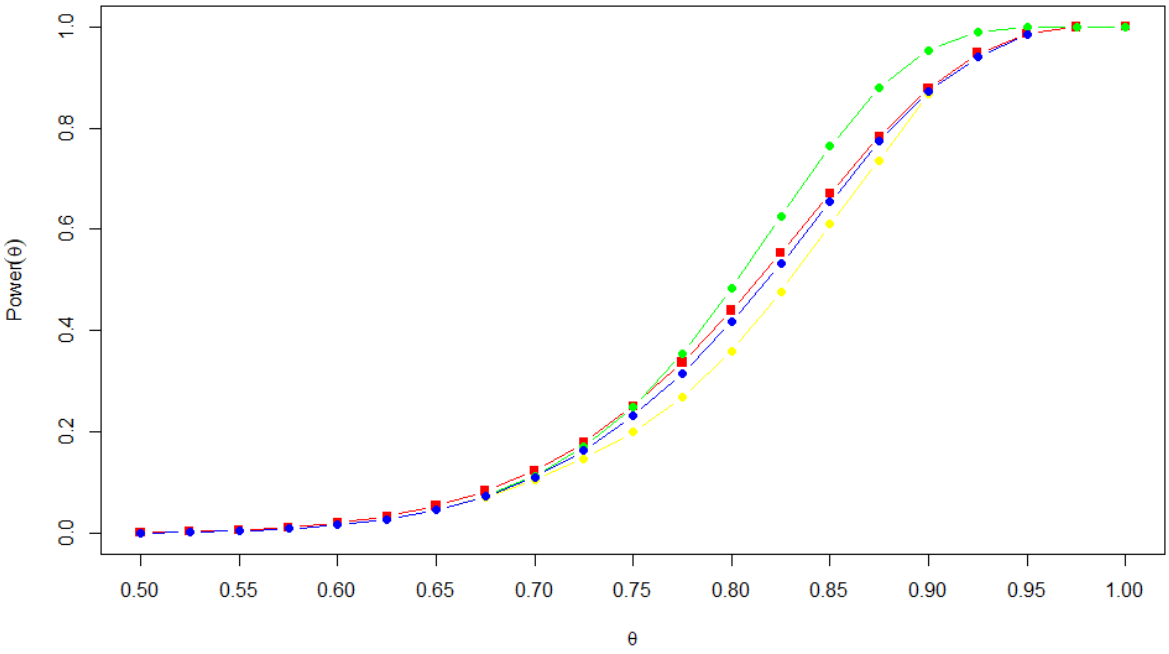

Figure 3. Power comparisons of NCS (red line) with NM (green, blue, and yellow lines). The red and green lines are based on the same situation displayed in Figure 1. The blue line corresponds to a situation where the examinee has a hit rate of 95 % on those questions that she thinks she can with certainty, while the yellow curve represents a hit rate of 90 % on the same questions. The significance level is set at 0.25.

Figure 4 sheds light on the impact of underrating one's knowledge. The red curve, as usual, represents the power for the NCS method and serves as a benchmark. The blue curve represents a situation where the test taker faces two types of questions, either questions she thinks she is 95 % confident with or questions for which she has no knowledge of. Moreover, we assume the hit rate is actually 100 % on the first type of questions, meaning that she underrates her knowledge. A little surprise is that the blue curve in Figure 4 is identical to the green curve in Figure 3. Thus, a slight underestimation of one's knowledge does not affect the power. The reason for this is that the test taker will continue to select the same questions to answer to maximize her probability of passing the exam, just as she would when her perceived

probabilities align with the true probabilities. When the underestimation becomes larger, the power decreases, as shown by the yellow curve. Here, the questions in the bank are either questions she thinks she is 90 % confident with or questions that she has no knowledge of. For the first type of questions, we still assume a hit rate of 100 %. How do we understand why the power decreases at most where θ is in the range 0.85 – 0.90? For such values of θ, it is not an unlikely outcome that the number of questions the test taker thinks she knows the answer of with a confidence of 90 %, is considered too few, causing her not to refrain from taking a chance on the rest of the questions. However, the actual number of questions the test taker knows the answer of could be adequate for her to successfully pass the exam without needing to take chances on the other questions. Thus, underestimation of the knowledge might result in the examinee unnecessarily taking a chance on the questions she does not know and that gambling results in a worse outcome. The power is now close to the power of the NCS method.

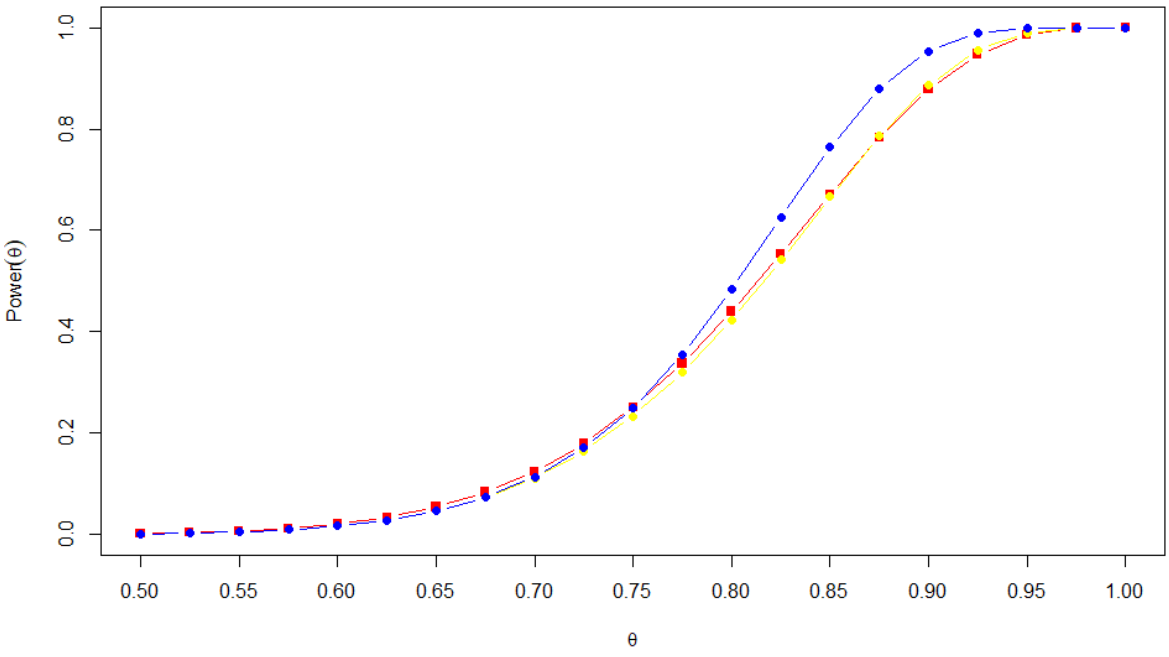

Figure 4. Power comparisons of NCS (red line) with NM (blue, and yellow lines) based on an exam with 20 true or false questions sampled from distributions of Π which only assumes values of 0.5 and 1. The two NM curves both represent situations where the examinee faces two types of questions, either questions she thinks she is almost certain on, while having a hit rate of 100 % on, or questions where she has no knowledge of. Almost certain means 95 % for the blue curve and 90 % for the yellow one. The significance level is set at 0.25.

5. Conclusions and Discussion

The paper presents a theoretical comparison of statistical power of the NCS and the NM methods used for true or false questions. The analysis is based on randomly drawing a certain number of questions, constituting an exam, from a large bank of questions. For the NCS method the test taker gets 1 point if the answer to the question is correct, 0 points otherwise. The NM method penalizes an incorrect answer with -1 point, while a correct answer is still rewarded with 1 point. Not answering the question results in 0 points. To compare the two methods, a knowledge parameter θ is defined as the average number of correct answers the test taker would have if she were to answer all questions in the bank.

We compare the two methods in terms of statistical power, i.e., we examine which method, based on the total score on the exam, has the highest probability to correctly pass a test taker who has a sufficiently high value on θ.

As to the result of the comparison, we do not identify a clear winner. The choice of method relies on various factors. For the situation where the variance of Π is small, the two methods are about equal. The explanation is that for such a situation the test taker will give answers to more or less all of the questions on the exam for both methods. Things are getting more exciting when we consider the opposite situation where the variance of Π is large. For this case, the results differ depending on whether the examinee correctly perceives her confidence level or not. If the level is perceived correctly, the power for NM is higher than for NCS. However, if the examinee either overrates or underrates her confidence level, the power of NM typically falls. Similar negative effects of mis-calibration is also found in Budescu and Bo (2015). Which method has the highest power depends on the magnitude of mis-calibration. Also, underrating does not affect the power of NM as much as overrating.

An assumption going through all the power calculations for NM is that the test taker has the ability to choose the strategy that maximizes the probability of passing the exam given her confidence ratings. It is not unlikely that this assumption is too strong in many situations, and it is possible that the test takers' skills in this regard vary, and that this possible variation can have major effects on the probability of passing the exam. To illustrate, suppose that 14 points is the lower limit to pass an exam with 20 questions. Moreover, suppose the test taker is 90 % confident of 18 questions and 55 % confident of 2 questions. To maximize the probability of passing in this case, the test taker should give answers only to the 18 questions being 90 % confident of, even though the expected number of points increases if she gives answers to all

questions. The optimal strategy yields a probability of passing of 0.734 while answering all 20 questions reduces the probability to 0.727. This is a small drop. However, things are getting worse if the test taker gives an answer only to 16 of those questions that she is most sure of. Then, the probability of passing diminishes to 0.514, which must be considered a not negligible drop.

Thus, it would be interesting to investigate the test takers' skills in this regard, both considering the average skill and the variation in skills among test takers. If the average skill is low, many test takers would be better off in terms of power with an NS exam. A large variation in skills would imply that some test takers would be disadvantaged by the NM method, while others would not, certainly this is a negative external effect of this method. Therefore, if this is the case, using the NM method would require a simple rule-of- thumb-guide to the test takers which questions to answer given their confidence levels. Developing such a guide is subject of future research.

Appendix

This Appendix presents the method used to find which questions to give answers to in the NM case and how to calculate the corresponding maximized probability of passing the exam. The method is based on the so-called Poisson binomial distribution being identified as part of a suitable probability model for the score on the exam.

Before we explain how to tackle the maximization issue, we first outline the Poisson binomial distribution. This distribution represents the probability of a discrete random variable X, which is defined as the sum of n independent Bernoulli trials, not necessarily identically distributed, with success probabilities $p_1, p_2, \dots, p_n$. The Binomial distribution is a special case, all success probabilities being equal. Following Wang (1993), the probability mass function of X is given by

$$P(X = x) = \sum_{A \in F_x} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j),$$

where F_x is the set of all subsets of x integers that can be selected from $\{1, 2, \dots, n\}$, while A^c is the complement of A.

Now, arrange the exam questions in order from highest to lowest based on the likelihood of providing the correct response. In addition, define p_i, $i = 1, 2, \ldots, 20$, to be the probability of giving a correct answer to the i:th question with respect to this order. Also, in line with this order, the student opts to respond to n of the 20 questions and not answer the others.[5] Define the random variable X_n to be the number of correct answers on the n questions. It follows a Poisson binomial distribution with success probabilities p_1, …, p_n. The score on the exam, denoted by Y_n, can now be expressed as $Y_n = X_n - (n - X_n) = 2X_n - n$, where the first equality follows from the way the scoring system is designed with negative marking used.

For a predetermined passing limit of 13 points, we can solve the maximization problem by calculating $P(Y_n \geq 13) = P(X_n \geq \frac{n+13}{2})$ for various values of n and choosing that value of n for which $P(Y_n \geq 13)$ is maximized. Up to 8 potential values for n should be examined because responding to fewer than 13 questions leads to a zero chance of passing the test. Utilizing the above probability mass function to determine these probabilities can be rather tedious. To sum over all $\frac{n!}{(n-x)!x!}$ elements in the set F_x might even be infeasible in practice unless n is small. Nonetheless, there are other, more efficient methods to determine the probabilities, such as employing a discrete Fourier transform (Hong, 2013), which is a method utilized in the r-package PoissonBinomial, version 1.2.7.

[5] If $p_i > p_j$, it cannot be optimal to answer the j:th question and not to answer the i:th question.

References

Akyol, P., Key, J., & Krishna, K., 2022. Hit or miss? Test taking behavior in multiple choice exams. *Annals of Economics and Statistics*, (147), 3-50.

Bar-Hillel, M., Budescu, D., & Attali, Y., 2005. Scoring and keying multiple-choice tests: A case study in irrationality. *Mind & Society* **4**, 3–12.

Brown, R.W., 2001. Multiple-choice versus descriptive examinations. In: 31st ASEE/IEEE Frontiers in Education. IEEE.

Budescu, D., & Bar-Hillel, M., 1993. To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277-291.

Budescu, D. V., & Bo, Y., 2015. Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, *80*(4), 1105-1122.

Burton, R. F., 2001. Quantifying the Effects of Chance in Multiple Choice and True/False Tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, Vol. 26, No. 1.

Burton, R.F., 2002. Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, *36*(9), 805-811.

Burton, R. F. (2005) Multiple-choice and true/false tests: myths and misapprehensions. *Assessment and Evaluation in Higher Education*, 30(1), pp. 65–72.

Choppin, B. H., 1988. Correcting for guessing. In J.P. Keeves (Ed.), Educational Research, methodology and measurement: an international handbook (pp. 384-386). Oxford: Pergamon Press.

Espinosa, M. P., & Gardeazabal, J., 2010. Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical psychology*, *54*(5), 415-425.

Harden RM, Brown RA, Biran LA, Dallas Ross WP, Wakeford RE., 1976. Multiple choice questions: to guess or not to guess. *Med Educ* **10**: 27–32.

Holzinger, K. J., 1924. On scoring multiple-response tests. *Journal of Educational Measurement, 15*, 445-447.

Hong, Y., 2013. On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*. **59**: 41–51.

Kubinger, K.D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M., 2010. On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*(1), 111-115.

Kurz, T. B., 1999. A review of scoring algorithms for multiple choice tests. *Paper presented at the Annual Meeting of Southwest Educational Research Association*.

Lesage, E., Valcke, M., & Sabbe, E., 2013. Scoring Methods for Multiple Choice Assessment in Higher Education—Is It Still a Matter of Number Right Scoring or Negative Marking? *Studies in Educational Evaluation*, 39, 118-193.

Lichtenstein, S. and Fischhoff, B., 1977. Do those who know more also know more about how much they know? *Organizational behavior and human performance*, *20*(2), pp.159-183.

Lord, F.M., 1953. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika* **18**, 57–76.

Muijtjens, A.M., Mameren, H.V., Hoogenboom, R.J., Evers J.L., van der Vleuten, C.P., 1999. The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Med Educ.* Apr; 33(4):267-75.

Rowley G.L., Traub R.E., 1977. Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement* 1977; **14**: 15–22.

Wang, Y. H., 1993. "On the number of successes in independent trials". *Statistica Sinica*. **3** (2), 295–312.

Zapechelnyuk, A., 2015. An axiomatization of multiple-choice test scoring. *Economics Letters*, Elsevier, vol. 132(C), 24–27.