

# A Singular Choice for Multiple Choice

Gudmund S. Frandsen  
BRICS, University of Aarhus  
Aabogade 34  
8200 Aarhus N, Denmark  
gudmund@brics.dk

Michael I. Schwartzbach  
BRICS, University of Aarhus  
Aabogade 34  
8200 Aarhus N, Denmark  
mis@brics.dk

## ABSTRACT

How should multiple choice tests be scored and graded, in particular when students are allowed to check several boxes to convey partial knowledge? Many strategies may seem reasonable, but we demonstrate that five self-evident axioms are sufficient to determine completely the correct strategy. We also discuss how to measure robustness of the obtained grades. Our results have practical advantages and also suggest criteria for designing multiple choice questions.

## Keywords

Multiple choice, scoring strategies, grading scales, theory.

## 1. INTRODUCTION

Multiple choice tests have recently been coming into favor as a useful evaluation tool at the university level [3, 10, 8], in contrast to the earlier view that they only support superficial learning [2]. While a new consensus thus seems to be forming about the use of multiple choice tests, there is no corresponding agreement on how answers to such tests should be evaluated. Most authors suggest different scoring strategies that are usually justified by intuition or by ad-hoc adjustments based on experiments. In this work we analyze the concept of scoring strategies and propose five axioms that in our view must be self-evident properties of any correct strategy, and we demonstrate that only a single such strategy satisfies all of these axioms. Importantly, our strategy is sufficiently general to handle *partial* knowledge. We also discuss the step from scoring to grading of tests, which is not trivial, and we provide a model for measuring the robustness of the obtained grades. Through examples we show that this robustness is highly dependent on the exact structure of the test.

## 2. MULTIPLE CHOICE TESTS

We now define the model of multiple choice tests that we consider. A *test* consists of a number of *questions*, say  $n$ , each of which has a number, say  $k_i$ , of possible *options* of which exactly one is correct. The text of the question is usually called the *stem*, and the correct option is called the *key* while the others are known as *distractors*. All questions are assumed to be independent of each other and all distractors are assumed to appear equally likely.

A student *taking* a test will check between 0 and  $k_i$  of the possible options for the  $i$ 'th question. A formal model of this is an  $n$ -tuple of individual answers, each of which is of

the form

$$(k_i, a_i, c_i)$$

where  $k_i$  is the number of possible options,  $a_i$  is the number of checked options, and  $c_i$  is a boolean indicating whether one of the checked options is the correct one.

As an example, consider the following test with two questions:

What is a newt?	
a	<input type="checkbox"/> A small animal
b	<input checked="" type="checkbox"/> A unit of gravity
<hr/>	
What is the correct URL encoding of "100% pure"?	
a	<input type="checkbox"/> 100%+pure
b	<input checked="" type="checkbox"/> 100+%+pure
c	<input checked="" type="checkbox"/> 100%25+pure

For this example, the two given answers are described as (2,1,0) and (3,2,1).

When *scoring* a test, the objective is to assign a numeric score to each student. This score can then be further mapped to a grading scale, which we consider as a secondary stage.

## 3. AXIOMS FOR SCORING STRATEGIES

Numerous different scoring strategies are employed for multiple choice tests, and many of them seem reasonable at a first glance. Since the questions are assumed to be independent, a scoring strategy will be a numeric function  $S(k_i, a_i, c_i)$  that assigns a value to the answer to a single question. A scoring strategy must obey some obvious axioms:

$$\begin{array}{ll} \text{Zero} & S(k, k, 1) = S(k, 0, 0) = 0 \\ \text{Monotonicity} & S(k, a, 0) < S(k, a, 1) \\ \text{Anti-Monotonicity} & a \leq b \Rightarrow S(k, b, c) \leq S(k, a, c) \end{array}$$

The *Zero* axiom states that checking all possible options is the same as leaving them all blank, and both cases contribute nothing. The *Monotonicity* axiom states that correct answers are rated higher than incorrect ones. The *Anti-Monotonicity* axiom states that fewer checked options will never decrease the score.

Thus, for any scoring strategy, the lowest possible score is

$$\sum_i S(k_i, k_i - 1, 0)$$

which corresponds to checking all options except the correct one, and the highest possible score is:

$$\sum_i S(k_i, 1, 1)$$

which corresponds to checking only the single correct option.

#### 4. A MULTITUDE OF STRATEGIES

The above three axioms are not enough to completely determine a scoring strategy. Ideally, a scoring strategy should correctly measure the knowledge of the student, but students may of course attempt to guess answers rather than admit ignorance. Thus, the interplay between the scoring strategy and the student's strategy must be considered.

The simplest possible scoring strategy is:

$$S_{absolute}(k, a, c) = \begin{cases} 1 & \text{if } a = c = 1 \\ 0 & \text{otherwise} \end{cases}$$

Here, only single correct options are given credit. This strategy has several problems: partial knowledge is not credited, and since there is no penalty for guessing, students will never leave questions unanswered.

Guessing can be discouraged by assigning a *negative* score to incorrect answers. A simple example strategy could be:

$$S_{harsh}(k, a, c) = \begin{cases} 1 & \text{if } a = c = 1 \\ -1 & \text{if } c = 0 \wedge a = 1 \\ 0 & \text{otherwise} \end{cases}$$

As indicated, this is a rather harsh strategy. A student that knows 70% of the answers and falsely believes that he knows the remaining 30% would score only 40%. This highlights a need to distinguish between guessing and giving wrong answers in good faith.

A more lenient strategy, which is proposed in [8], is:

$$S_{Roberts}(k, a, c) = \begin{cases} 1 & \text{if } a = c = 1 \\ -\frac{1}{k-1} & \text{if } c = 0 \wedge a = 1 \\ 0 & \text{otherwise} \end{cases}$$

Here, the penalty for an incorrect answer is scaled by the number of distractors.

Partial knowledge may be credited in many different ways. A simple scoring strategy could in this case be:

$$S_{partial}(k, a, c) = \begin{cases} 1/a & \text{if } c = 1 \wedge 0 < a < k \\ 0 & \text{otherwise} \end{cases}$$

The idea of giving proportionally reduced credit for multiple options is appealing, but is not the logically correct choice, as we show in the next section.

#### 5. TWO DEFINING AXIOMS

It is easy to check that the above proposed scoring strategies all satisfy the three basic axioms, and that infinitely many other strategies may seem equally reasonable. Rather than

try to argue subjectively for a particular strategy, we will introduce two further axioms that we believe to be self-evident and that will completely define a single scoring strategy.

The first axiom observes that the evaluation of a multiple choice test should be invariant under a simple transformation. Consider the example test shown in Section 2. We could choose to combine the two questions into a single one, and the student would correspondingly combine his answers:

What is a newt and what is the correct URL encoding of "100% pure"?

a	<input type="checkbox"/>	A small animal and 100%+pure
b	<input type="checkbox"/>	A small animal and 100+%+pure
c	<input type="checkbox"/>	A small animal and 100%25+pure
d	<input type="checkbox"/>	A unit of gravity and 100%+pure
e	<input checked="" type="checkbox"/>	A unit of gravity and 100+%+pure
f	<input checked="" type="checkbox"/>	A unit of gravity and 100%25+pure

Clearly, the same amount of knowledge has been displayed by the student, thus the contribution to the numeric score should be unchanged. We can phrase this as an *Invariance* axiom:

$$S(k_1, a_1, 1) + S(k_2, a_2, 1) = S(k_1 k_2, a_1 a_2, 1)$$

This has the unique solution

$$S(k, a, 1) = \log\left(\frac{k}{a}\right)$$

In retrospect, this is clearly the correct strategy, since it measures the number of bits of information that the student has contributed.

We still need to determine the numeric score of an incorrect answer. From an information theoretic point of view the value should be  $-\infty$ , but that is clearly not what we want. Recall that guessing should be punished but wrong answers given in good faith should not. There is only one balance point between these two points of view, namely that the expected outcome from guessing should be zero. This can be expressed by the following *Zero-Sum* axiom:

$$\frac{\binom{k-1}{a-1} S(k, a, 1) + \binom{k-1}{a} S(k, a, 0)}{\binom{k}{a}} = 0$$

For wrong answers where  $a \neq k$  and  $a > 0$ , we again obtain a unique solution

$$S(k, a, 0) = -\frac{a}{k-a} \log\left(\frac{k}{a}\right)$$

So, wrong answers receive a negative score, but a more modest one than in the harsh strategy.

#### 6. THE CORRECT SCORING STRATEGY

Combining the five axioms we have proposed, we arrive at a uniquely determined scoring strategy:

$$S_{axioms}(k, a, c) = \begin{cases} 0 & \text{if } a = 0 \vee a = k \\ \log\left(\frac{k}{a}\right) & \text{if } a > 0 \wedge c = 1 \\ -\frac{a}{k-a} \log\left(\frac{k}{a}\right) & \text{if } a > 0 \wedge c = 0 \end{cases}$$

To the best of our knowledge, this strategy has not been proposed before. Note that any other strategy is bound to

violate one of our five axioms, and reasons for doing so seem difficult to justify.

An interesting special case is a test consisting of only *yes/no* questions. In this situation, our scoring strategy specializes to the following:

$$S_{\text{binary}}(k, a, c) = \begin{cases} 0 & \text{if not answered} \\ \log(2) & \text{if correct} \\ -\log(2) & \text{if incorrect} \end{cases}$$

which is in this case the same as  $S_{\text{harsh}}$ , since a scoring strategy may of course be scaled with an arbitrary factor. The binary strategy simply states that correct and wrong answers must be inverses of each other and a question not answered should yield zero. The fact that wrong knowledge is worse than no knowledge was of course already observed by Socrates who stated “*I know nothing except the fact of my ignorance*”. It seems to be a folklore result that binary multiple choice tests should be graded in this manner.

## 7. FROM SCORING TO GRADING

The scoring strategy assigns a numeric score to a test, but usually this must be transformed into a grade on some fixed scale. The discrete nature of grade scales introduces some additional complications.

For example, who should receive the lowest grade? Clearly, a score of zero must be rewarded with the lowest grade, since no knowledge has been demonstrated. As a consequence, all negative scores must similarly receive this grade. This means that if a student is certain that his true knowledge results in a score of zero or less, then guessing clearly pays off. In reality, this will be a rare case, but the same situation happens at all other grades. If a student is absolutely certain that his true knowledge places him just below a  $B$  grade, then it pays off to perform a limited amount of guessing, as long as the maximum negative score he risks is small enough that he cannot slip below a  $C$  grade. These considerations simply indicate that grading scales ideally should have small intervals.

On the ECTS grading scale, however, the distinction between the grades FX and F could suggest that negative scores could be mapped to the grade F while zero scores would yield the grade FX. On the US grading scale, there is no distinction between these two grade levels.

To obtain a percentage grade, the correct algorithm for our scoring strategy is:

$$\frac{\max(0, \sum_i S(k_i, a_i, c_i))}{\sum_i \log(k_i)}$$

Following this conversion, there are usually standardized means of translating from percentage grades to other grading scales.

Formally, we define a *grading scale* as an increasing vector of numbers in the interval  $]0..1[$ , such as the commonly used Danish scale  $[0.40, 0.50, 0.58, 0.66, 0.74, 0.82, 0.90]$ . In *absolute* grading, these numbers are percentages that delimit the different grades, whereas in *relative* grading they are percentiles. Note that for absolute grading the scheme must depend on the individual test, since the difficulties of

the questions determine the average and distribution of the scores.

## 8. ROBUSTNESS AND CONFIDENCE

A multiple choice test is prepared, the students take the test, and the answers are subsequently scored and graded. An interesting question is to consider how robust the resulting grades are when random guessing by the students is considered.

Intuitively, more questions and more options make the test more reliable, but how many questions and options are necessary to give a *robust* test? Relying on intuition only is problematic. If your test is under-robust you cannot trust the result to the extent you believe, and if it is over-robust you could have saved precious time by generating fewer questions and distractors.

We propose a formal definition of robustness that makes it possible to check in advance whether a multiple choice test  $[k_1, k_2, \dots, k_n]$  (only the numbers of questions and options are relevant) is robust with respect to an absolute grading scale  $[t_1, t_2, \dots, t_g]$ . To motivate our definition let us consider some examples.

A student that knows nothing can by a lucky strike get a full 100% score. Clearly, this is unavoidable. Similarly, if a student has a score in the interval  $[t_{i-1}..t_i[$  and very close to  $t_i$ , then he may increase the grade with probability  $1 - \frac{1}{k}$  by a single guess (with no risk at all if the test has reasonably many questions). This situation is also unavoidable. The two examples show that with sufficiently small probability anything can happen, and a small deviation in the resulting grade may happen with relatively large probability. But we can demand that a robust test has large deviations only with small probability, where large means more than one grade off, and small probability is customarily taken to mean 5% or less.

Formally, we define the *nonconfidence* of a test with respect to a grading scale  $\text{nonc}([k_1, k_2, \dots, k_n], [t_1, t_2, \dots, t_g])$  to be the maximum probability over  $i \in \{1, \dots, g-1\}$  that a student whose true knowledge corresponds to a score in the interval  $[t_{i-1}..t_i[$  gets a score of  $t_{i+1}$  or more by guessing, i.e. that he advances two grades or more. The test  $[k_1, k_2, \dots, k_n]$  is called *robust* with respect to an absolute grading scale  $[t_1, t_2, \dots, t_g]$ , if  $\text{nonc}([\bar{k}_i], [\bar{t}_i]) \leq 5\%$ .

To illustrate these definitions, consider a simplified version of the Danish grading system, namely  $[0.40, 0.50]$ , where a student passes with a score of 50% or more and has a low fail with a score of less than 40%. A test is robust for this grading system if a student who deserves a low fail has probability of at most 5% of passing. What restrictions must a robust test satisfy? We have computed lower bounds on the number of questions needed to obtain a robust test when all questions have  $k$  options:

$k$	$n$
2	155
3	80
4	55
5	40
6	25

Note that the numbers in the table are lower bounds but not necessarily optimal. For example, the entry  $n = 80$  for  $k = 3$  means that any test with fewer than 80 questions that each have 3 options is not robust, but we give no guarantee that a test with 80 questions or more is robust.

Let us make a technical comment on why we cannot give such a guarantee. Most importantly, the non-confidence function is not monotonely decreasing as a function of  $n$ . This may appear strange, but is due to the discrete jumps between possible scores. Consider a test where all questions have 2 options. If  $n = 15$  then a low fail ( $<40\%$ ) can be at most 33.4% and a pass ( $\geq 50\%$ ) is at least 53.3%. Adding one question actually makes it easier for the student to guess a passing score. If  $n = 16$  then a low fail ( $<40\%$ ) can be up to 37.5% and a pass ( $\geq 50\%$ ) can be as little as 50%.

For  $k > 2$ , nontrivial scores may arise from the student selecting more than one option for some questions. This makes it much harder to compute the non-confidence function. We have chosen to compute a lower bound on the non-confidence function rather than the exact function by restricting the student to select a single option for each question. For this reason, the numbers in the above table are only lower bounds on the minimal  $n$  that leads to a robust test. These numbers may appear surprisingly large, but the above grading system can hardly be considered unreasonably fine grained. It is suggested in [10] that  $k = 3$  is always an adequate choice, but our numbers seem to suggest otherwise.

We do not consider the problem of guessing leading to lower grades, since students may completely eliminate this risk by refraining from guessing.

When using an absolute grading scale, we can assert robustness of a test without knowing anything about the students who are going to take the test. When using a relative grading scale things are different. If all students have approximately identical true scores then even a little guessing may lead to scores that rank them in essentially random order, implying grades that are uncorrelated to their knowledge. This may happen no matter how large and well designed the test is, when student proficiency is not known in advance. Relative grading has another drawback. If there are failing grades, it seems unreasonable to have a fixed percentile fail. This view is reflected in the ECTS scale that has an absolute threshold for pass/fail and FX/F combined with a relative distribution of the pass grades A,B,C,D,E. In conclusion, without knowing the actual student scores, one may estimate robustness of a multiple choice test only with respect to absolute (parts of) grading scales.

## 9. RELATED WORK

The literature on multiple choice tests falls in several categories. Many papers analyze the multiple choice paradigm

and provide criteria and tools for designing tests that are valid for higher education [7, 9, 8, 10, 6, 4]. Additionally, much work in psychology and education deal with techniques for designing and phrasing stems, keys, and distractors. This is not really related to our work, so we do not provide references for this area. There is ample documentation that well-designed multiple choice tests correlate with descriptive examinations, some recent example of which are [6, 3].

Two references are particularly relevant to our work. The paper [4] discusses myths and misapprehensions about multiple choice tests. We note that the use of negative scores and the distinction between guessing and “good faith” ignorance figures prominently in the discussions. The paper [5] surveys scoring strategies for multiple choice tests that reward partial knowledge. Three different strategies are presented here. The strategy:

$$S_{scale}(k, a, c) = \begin{cases} \frac{k-a}{k} & \text{if } c = 1 \\ 0 & \text{otherwise} \end{cases}$$

violates the *Invariance* and *Zero-Sum* axioms; the strategy (originally from [1]):

$$S_{Akeroyd}(k, a, c) = \begin{cases} 1 & \text{if } a = c = 1 \\ 0.5 & \text{if } c = 1 \wedge a = 2 \\ 0.25 & \text{if } c = 0 \wedge a = 0 \\ 0 & \text{otherwise} \end{cases}$$

violates the *Zero*, *Monotonicity*, *Invariance* and *Zero-Sum* axioms; and the strategy:

$$S_{Bush}(k, a, c) = \begin{cases} \frac{k-a}{k-1} & \text{if } c = 1 \\ \frac{-a}{k-1} & \text{otherwise} \end{cases}$$

violates the *Invariance* and *Zero-Sum* axioms. While strictly incorrect, the intuitions behind these grading strategies seem to circle around ideas that are similar to our axioms: penalties for incorrect answers and scaling with the number of distractors.

The previously mentioned strategy:

$$S_{Roberts}(k, a, c) = \begin{cases} 1 & \text{if } a = c = 1 \\ -\frac{1}{k-1} & \text{if } c = 0 \wedge a = 1 \\ 0 & \text{otherwise} \end{cases}$$

does in fact coincide with our strategy (after suitable scaling with a factor of  $\log(k)$ ) for the special case where partial knowledge is not allowed and all questions have the same number of options. Thus, our more general strategy may be seen as a pathway for lifting some limitations inherent in the underlying design of multiple choice tests.

Regarding the concept of robustness, we are not aware of any related work. The papers [6, 8] examine the correlation between grades obtained by multiple choice test and by descriptive examinations, but the innate robustness of a single test is not quantified.

## 10. CONCLUSION

We have presented five axioms for multiple choice scoring strategies that we believe to be self-evident. They uniquely determine a novel scoring strategy that also deals with partial knowledge and questions with varying numbers of options. Our contribution seems to provide a logical answer to

a question that previous authors have either struggled with or explicitly avoided.

The Web site <http://www.brics.dk/Multiple> contains an online service for generating, scoring, and evaluating multiple choice tests using the results presented in this paper.

## 11. REFERENCES

- [1] F. Akeroyd. Progress in multiple-choice scoring methods. In *Journal of Further and Higher Education*, 6. Carfax Publishing, 1982.
- [2] J. Biggs. *Teaching for Quality Learning at University*. Open University Press, 1999.
- [3] R. W. Brown. Multiple-choice versus descriptive examinations. In *31st ASEE/IEEE Frontiers in Education*. IEEE, 2001.
- [4] R. F. Burton. Multiple-choice and true/false tests: myths and misapprehensions. In *Assessment and Evaluation in Higher Education*, 30(1). Carfax Publishing, 2005.
- [5] M. Bush. A multiple choice test that rewards partial knowledge. In *Journal of Further and Higher Education*, 25(2). Carfax Publishing, 2001.
- [6] D. W. Farthing and D. McPhee. Multiple choice for honours-level students? a statistical evaluation. In *Third Annual CAA Conference*. CAA Centre, 1999.
- [7] R. Lister. Objectives and objective assessment in cs1. In *SIGCSE 2001*. ACM, 2001.
- [8] T. S. Roberts. The use of multiple choice tests for formative and summative assessment. In *ACE 2006*. Australian Computer Society, 2006.
- [9] D. Traynor and J. P. Gibson. Synthesis and analysis of automatic assessment methods in cs1 – generating intelligent mcqs. In *SIGCSE 2005*. ACM, 2005.
- [10] K. Woodford and P. Bancroft. Multiple choice questions not considered harmful. In *ACE 2005*. Australian Computer Society, 2005.