How to cheat scientifically: an information theoretical approach to solve multiple choice quizzes

***Abstract*—TODO**

I. Introduction

Multiple choice questions (quizzes), with their pros and their cons [1] [2], are largely adopted worldwide for tests and examinations. With this methodology, students are called to answer TRUE or FALSE, or, more generally, choose one right answer over k possible answers to demonstrate their knowledge on a specific topic. It is not a rare case that examiners give to students in advance the complete list of *all* the possible quizzes (questions, the right answer and all the wrong answers) while, on the examination day, students are asked to provide the correct responses to a *subset* of them. This is the case of German and Austrian examinations for Driver License, or the many public competitive exams in Italy.

Do the students really need to learn the subjects to answer correctly to all the tests?

Certainly not, but let we better clarify the problem with a trivial example.
Suppose the universe of all the possible TRUE/FALSE quizzes is composed by just these three questions/answers that you can read way before attempting a test:

1 Is this paper interesting?
Right answer $\rightarrow$ TRUE
2 Have the men three legs?
Right answer $\rightarrow$ FALSE
2 Can a camel go through the eye of a needle ?
Right answer $\rightarrow$ FALSE

On the examination day, you must answer to two questions, randomly chosen from the previous set, and provide the correct answers to both of them. We are sure that nobody can go wrong even with a minimum/no study of the quizzes.

Now let we take another set of questions:

1 In the superstring theory, the series solution of the quartic integrals of the non-functional kind is $e^f \sum_{n,m,p=0}^{\infty} \frac{b^{4n}}{(4n)!} \frac{c^{2m}}{(2m)!} \frac{d^{4p}}{(4p)!} \frac{\Gamma(3n+m+p+1/4)}{a^{3m+m+p+1/4}}$
Right answer (obviously) $\rightarrow$ TRUE
2 Is Sir Thomas North the first translator that brought Plutarch's works to Western Europe in 1559?
Right answer $\rightarrow$ FALSE
2 Is the following chemical reaction correctly balanced $3Cu + 8HNO_3$ g $3Cu(NO_3)_2 + 2NO + 4H_2O$?
Right answer $\rightarrow$ TRUE

Now the probability to give a wrong answer will increase sensibly because if we look at the *meaning*, the topics are more difficult (imagine even only 100 questions of this kind), but the *form* is identical. Indeed, if the change the point of view, the quizzes present the exactly same level of difficulty.

Let we take both the previous sets of quizzes and totally ignore all the meaning, treating them as a simple set of characters on top of which we put a simple algorithm.
In the first quiz set, we rapidly see that if the second character of the question is 'a', then the answer is FALSE, otherwise it is TRUE.
For the second set of quiz, if the third character is 't' then the correct answer is TRUE and FALSE otherwise.

One could object that there is no correlation between the characters of the questions and the right answer of the quiz and this is absolutely correct. However, with a bit of "processing" on the dataset of the quizzes, we can derive, in almost all the relevant cases, some rules like these as we will show later on this paper.

Goal of this work is indeed to i) quantify how much information the students must know only for the purpose of providing correct answers to all the quizzes (and not for demonstrating their knowledge of the topic!), using standard information techniques, and ii) devise "human friendly" algorithms to be applied on a reduced information base for quickly provide the correct answers, without the need of a computer.

In other words, how to cheat on quizzes using an information theoretical approach.

II. Problem Statement

We consider a set of n question, each of which has exactly one corrected answer and $k-1$ wrong answer. We call this a *quiz database*.

Definition 1: We define a "human friendly" algorithm as an algorithm that can be executed easily by a human, without the need of a computer.

Examples of a human friendly algorithm is performing few sums or multiplications with small numbers, while, for instance, applying the Deflate compression is not an easy task to perform without the aid of a computer.

Definition 2: We define a *quiz test* as a random subset of questions taken from a quiz database that must be answered correctly in order to pass the test.

Questions are provided without any ID and without any specific order. Throughout this paper, if no otherwise specified, answers are presented in the same order of the quiz database e.g. if the right answer for a given question in the third in the database, so it will be on each quiz test. This happens often in real cases. However if also the answers are scrambled, a simple sorting function (e.g. order the answer according to the number of characters) can easily provide a way to define a scramble-insensitive order.

Definition 3: Given a quiz database, we define the **quiz problem** as the problem of finding an information base together with a "human friendly" algorithm that allow to answer correctly to any question in the database.

In words, we are searching for a mapping from the strings that compound the quiz database to the correct response.

The optimal solution of this problem is a simple algorithm with the less mnemonic requirement.

This problem is related with the Kolmogorov Complexity [3] if we see the problem as finding the minimum "computability resources" to define all the questions and the right answers. However, we have the constraint that the algorithm must be human friendly, and that our questions and answers are only a small subset of all the possible permutations of the characters in the alphabet.

The problem can be viewed also as the problem of finding an hash function to apply on each question that return a number in the interval $[0{-}k]$ corresponding to the right answer. Hence, it is more proper to refer to it as a *mapping function*.

III. An algebraic solution

In the most challenging scenario, the correct answer have equal probabilities, resulting in the higher degree of entropy of the information. Than we suppose to make m random *tests* on each question. Each test is basically a simple hash function whose result is a number in the range $[0-k]$ with (almost) equal probability, hence $t(\cdot) \rightarrow [0-k]$.

For example, if $k = 2$ (TRUE/FALSE quizzes) a good test can be to check if the first character of the question is in the range [A-L]. In this case the testing function returns 1, and 0 otherwise. Another test can be done in the same manner but with reference to the second character of the question etc.

Applying the m tests on the n questions, we obtain a matrix $m \times n$ matrix that we call A. Then, by calling b the vector of all the positions of the right answers, we can write the simple algebraic equation:

$$\mathbf{Ax} = \mathbf{b} \tag{1}$$

Inverting the matrix we can easily found $\mathbf{x}$ [1]. This procedure can be rapidly done by a computer using the quiz database. Then we just need to memorize $\mathbf{x}$ that is an information base that permits to answer correctly to any quiz. Indeed, it suffices to apply all the m tests to a question to obtain a vector and then make the scalar product with $\mathbf{x}$ (in the Galois Field k) to obtain the right answer. All these operations are extremely fast and easy (basically small sum and multiplications), even for humans.

Solving 100 TRUE/FALSE quizzes on quantum physics (given in any language!) requires no more than memorize three telephone numbers i.e. 100 bits or equivalently 30 decimal numbers.

A. Example: the driver license

Let we take $n = 10$ random TRUE/FALSE quizzes from a study questions for the Illinois State driver license test.

1 Unless authorized to do so, drivers may not break into the line of a funeral procession. (TRUE)
2 It is permissible to pass on a two-lane, two-way roadway within 100 feet of an intersection or railroad crossing. (FALSE)
3 Drivers may open car doors on the side on which traffic is moving only when it can be done safely and without interfering with traffic. (TRUE)
4 Slow vehicles should use the left-hand lane except when passing or making a left turn? (FALSE)
5 When a motorist is turning right and a bicyclist is approaching on the right, let the bicyclist go through the intersection first before making a right turn. (TRUE)
6 It is permissible for anyone to wear a headset while driving. (FALSE)
7 In urban areas, drivers moving out of an alley, building, private road or driveway need not come to a complete stop before entering the roadway if the roadway is clear of traffic. (FALSE)
8 When on a two-lane roadway, drivers must stop their vehicles when approaching a stopped school bus with its red warning lights flashing and its stop signal arm extended. (TRUE)
9 When following a truck at night, it is important to dim your headlights. (TRUE)
10 A parent or legal guardian may request that the drivers license of a minor be cancelled at any time prior to age 18. (TRUE)

Then we define 10 tests: we take the first 10 characters (without commas and white-spaces) of each question and for each of them we assign the value 1 if the character is in the range [A-L], and 0 otherwise. For the answers, we assign 1 to TRUE and 0 to FALSE.

For the previous set of quizzes, applying eq. 1 we have:

[1] In the rare case the matrix presents a singularity, we just need to add/change a test

$$
\underbrace{\begin{pmatrix}
0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1
\end{pmatrix}}_{\mathbf{A}}
\underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \end{pmatrix}}_{\mathbf{x}}
=
\underbrace{\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}}_{\mathbf{b}}
$$

That admits a solution for $\mathbf{x} = (1, 1, 0, 1, 0, 1, 0, 0, 0, 0)$. Knowing these ten bits (not necessary memorized as bit array), we can solve any of the previously presented quiz. For instance, if we must answer to the question 'It is permissible for anyone to wear a headset while driving.' (question 6), we apply the same tests on the first 10 characters of the question to obtain the vector $\mathbf{v} = (1, 0, 1, 0, 0, 1, 0, 0, 1, 0)$. Then from the scalar product $\mathbf{v}{\cdot}\mathbf{x}$ we obtain 2 whose modulo k is 0. Hence the correct answer is FALSE, no headset while driving.

References

[1] V. L. Clegg and W. E. Cashin, "Improving multiple-choice tests. idea paper no. 16." 1986.

[2] L. C. Jacobs and C. I. Chase, *Developing and Using Tests Effectively. A Guide for Faculty.* ERIC, 1992.

[3] M. Li and P. M. Vitányi, *An introduction to Kolmogorov complexity and its applications*. Springer, 2009.