Random Guess Scores

Helen Jordan

September 9, 2009

Contents

1 Introduction

The aim of this project is to calculate, for various types of multiple response questions, what percentage score we would expect a student picking responses at random to achieve. It is assumed that the student guesses logically: that is, chooses responses based on the information provided so as to maximise their expected score, without having any subject knowledge (or any common sense which might suggest a certain answer is incorrect!). So we assume that if a student is told there are n correct responses, they will always choose n responses, and if they are given feedback (for example telling them that a certain number of their choices are incorrect), they will never enter an answer which contradicts this feedback.

Some of the calculations included in this report are fairly involved, and they will only be understandable to those with a good knowledge of probability. However, the end result of these calculations should be accessible to anyone with a basic knowledge of mathematics. Tables of simulated values and a RGS calculator for some question types have also been given, as explained in Section 4 on page 7.

2 Single Attempt

2.1 Multiple Response with known number of correct responses

Here students are told that there are n correct responses out of the m total responses, so our random guesser will choose n responses. Let X be a random variable representing the number of correct choices made. Then

$$X \sim \text{HyperGeometric}(m, n, n)$$

Suppose the student is given $\frac{100}{n}\%$ for each correct response. Then

$$\text{RGS} = \frac{100}{n} \times \mathbb{E}(X) = \frac{100}{n} \times \frac{n^2}{m} = \frac{100n}{m}$$

If all choices need to be correct to get the mark, then

$$\text{RGS} = 100 \times \mathbb{P}(X = n) = \frac{100}{\binom{m}{n}}$$

Where

$$\binom{m}{n} = \frac{m!}{(m-n)!\,n!}$$

and

$$n! = n \times (n-1) \times \ldots \times 2 \times 1$$

2.2 Multiple Response with unknown number of correct responses

Suppose that n out of m of the responses are correct, but the student is not told what n is. It is hard to calculate a RGS here, because we don't know how many responses the student is likely to select. Let L be a random variable denoting the number of responses selected by the student. It seems reasonable that the guesser would just select each responses to be in their selection with some fixed probability p, independently of all other selections. In this case

$$L \sim \text{Binomial}(m, p)$$

Again let X represent the number of correct choices. Irrespective of the distribution of L,

$$X|(L = l) \sim \text{HyperGeometric}(m, n, l)$$

Suppose the student is given $\frac{100}{n}\%$ for each correct response, and if they select more than n responses only the worst n are marked. Then

If $L \leq n$,

$$\text{Score} = \frac{100}{n} \times X$$

If $L > n$,

$$\text{Score} = \frac{100}{n} \times \max(0, X + n - L)$$

So

$$\begin{aligned}
\text{RGS} &= \frac{100}{n}\left[\sum_{l=0}^{n} \mathbb{P}(L=l)\mathbb{E}(X|L=l) + \sum_{l=n+1}^{m} \mathbb{P}(L=l)\mathbb{E}(\max(0, X+n-L)\,|L=l)\right] \\
&= \frac{100}{n}\left[\sum_{l=0}^{n} \mathbb{P}(L=l)\frac{n\,l}{m} + \sum_{l=n+1}^{m} \mathbb{P}(L=l)\sum_{k=0}^{l} \mathbb{P}(X=k)\max(0, k+n-l)\right] \\
&= \frac{100}{n}\left[\sum_{l=0}^{n} \mathbb{P}(L=l)\frac{n\,l}{m} + \sum_{l=n+1}^{m} \mathbb{P}(L=l)\sum_{k=l-n+1}^{l} \mathbb{P}(X=k)(k+n-l)\right] \\
&= \frac{100}{n}\left[\sum_{l=0}^{n} \mathbb{P}(L=l)\frac{n\,l}{m} + \sum_{l=n+1}^{m} \mathbb{P}(L=l)\sum_{k=l-n+1}^{l} \mathbb{P}(X=k)(k+n-l)\right] \\
&= \frac{100}{n}\left[\sum_{l=0}^{n} \mathbb{P}(L=l)\frac{n\,l}{m} + \sum_{l=n+1}^{m} \mathbb{P}(L=l)\sum_{k=l-n+1}^{l} \frac{\binom{n}{k}\binom{m-n}{l-k}}{\binom{m}{l}}(k+n-l)\right] \\
&= \frac{100}{n}\left[\sum_{l=0}^{n} \mathbb{P}(L=l)\frac{n\,l}{m} + \sum_{l=n+1}^{m} \mathbb{P}(L=l)\sum_{r=1}^{n} \frac{\binom{n}{r+n-l}\binom{m-n}{n-r}}{\binom{m}{l}}r\right]
\end{aligned}$$

If $L \sim \text{Binomial}(m, p)$, this is

$$\begin{aligned}
&= \frac{100}{n}\left[\sum_{l=0}^{n} \binom{m}{l} p^l (1-p)^{m-l}\frac{n\,l}{m} + \sum_{l=n+1}^{m}\sum_{r=1}^{n} \binom{n}{r+n-l}\binom{m-n}{n-r} r p^l (1-p)^{m-l}\right] \\
&= 100\left[\sum_{l=1}^{n} \binom{m-1}{l-1} p^l (1-p)^{m-l} + \sum_{l=n+1}^{m}\sum_{r=1}^{n} \binom{n}{r+n-l}\binom{m-n}{n-r} \frac{r p^l (1-p)^{m-l}}{n}\right]
\end{aligned}$$

If the student scores only if they have selected all of the correct choices and none of the incorrect ones, then things are much simpler:

$$\begin{aligned}
\text{RGS} &= 100 \times \mathbb{P}(L=n) \times \mathbb{P}(X=n|L=n) \\
&= \frac{100\mathbb{P}(L=n)}{\binom{m}{n}}
\end{aligned}$$

If $L \sim \text{Binomial}(m, p)$, this is

$$= 100\, p^n \,(1-p)^{m-n}$$

2.3 Drag and Drop

2.3.1 Each choices may be used an unlimited number of times

In this case, each of the n boxes is filled with one of the m choices. Each time the random guesser fills a box, they pick one of the m possible choices at random (independently of their choices for the other boxes), so they have a probability $\frac{1}{m}$ of choosing the correct answer.

So if the student is given $\frac{100}{n}\%$ for each correct response, then

$$\text{RGS} = \frac{100}{n} \times \frac{n}{m} = \frac{100}{m}$$

If the student scores only if they have filled all the boxes correctly, then

$$\text{RGS} = 100 \times \left(\frac{1}{m}\right)^n$$

2.3.2 Each choice may only be used once

In this case, there are m choices, each of which may be put into at most one of the n boxes, so there are n correct answers, and $m-n$ incorrect answers which should go unused. Let the boxes be labelled $1, 2, \ldots n$. Let L be a random variable representing the number of "correct" answers selected, not necessarily in the right boxes. Given $L = l$, without loss of generality, suppose that these correct answers are contained (in some order) in boxes $1, 2, \ldots l$. Let i_j be the choice placed in box j. Then the probability that precisely k boxes are correct is

$$\frac{\binom{l}{k} D_{l-k}}{l!}$$

where D_n is the number of derangements of a set of i objects: that is, the number of permutations such that no object is mapped to itself. A standard result is that $D_n = n! \sum_{i=1}^{n} \frac{(-1)^i}{i!}$. Now

$$L \sim \text{HyperGeometric}(m, n, n)$$

So

$$\mathbb{P}\left(\text{precisely } k \text{ boxes are correct}\right) = \sum_{l=k}^{n} \frac{\binom{n}{l}\binom{m-n}{n-l}}{\binom{m}{n}} \frac{1}{k!} \sum_{i=0}^{l-k} \frac{(-1)^i}{i!}$$

So

$$\text{RGS} = \frac{100}{n} \sum_{k=1}^{n} \sum_{l=k}^{n} \frac{\binom{n}{l}\binom{m-n}{n-l}}{\binom{m}{n}} \frac{1}{(k-1)!} \sum_{i=0}^{l-k} \frac{(-1)^i}{i!}$$

3 Multiple Attempts

Now we assume that students are allowed t attempts at the question, and that the maximum score available to them at the jth attempt is $100 \times \left(1 - \frac{j-1}{t}\right)$. We assume that credit is only given for partially correct answers at the final attempt, so that the student gets a score of $100 \times \left(1 - \frac{j-1}{t}\right)$ if they get everything right at the jth attempt, or $\frac{100}{n} \times \frac{k}{t}$ if they get k choices right at the final attempt.

Suppose that the student has not got the question entirely correct. Then there are two possibilities for what information the student will be given:

1. They will simply be told that their answer is incorrect. They are not told how many of their choices are correct.
2. They are told how many of their choices are correct.

3.1 Multiple Response with known number of correct responses

Suppose (1) holds, so that students are not told how many of their choices are correct. Let A_j be the event that all choices are correct on the jth attempt. Then

Claim:

$$\mathbb{P}(A_j) = \frac{1}{\binom{m}{n}} \ \forall j = 1, \ldots, t$$

Proof

$$\begin{aligned}
\mathbb{P}(A_j) &= \mathbb{P}\left(\cap_{i=1}^{j} A_i^{\mathrm{c}}\right) \mathbb{P}\left(A_j | \cap_{i=1}^{j} A_i^{\mathrm{c}}\right) \\
&= \mathbb{P}\left(\left[\cup_{i=1}^{j} A_i\right]^{\mathrm{c}}\right) \frac{1}{\binom{m}{n} - j + 1}
\end{aligned}$$

Since the events A_j are disjoint, $\mathbb{P}\left(\cup_{i=1}^{j} A_i\right) = \sum_{i=1}^{j} \mathbb{P}\left(A_i\right)$, so

$$\begin{aligned}
&= \left(1 - \frac{k-1}{\binom{m}{n}}\right) \frac{1}{\binom{m}{n} - j + 1} \\
&= \frac{1}{\binom{m}{n}}
\end{aligned}$$

as required.

So

$$\begin{aligned}
\mathrm{RGS} &= 100 \times \left(\frac{2 + 3 + \ldots + t}{t \times \binom{m}{n}} + \sum_{k=1}^{n} \frac{k}{nt} \times \mathbb{P}\left(k \text{ correct at } t\text{th attempt}\right)\right) \\
&= 100 \times \left(\frac{2 + 3 + \ldots + t}{t \times \binom{m}{n}} + \sum_{k=1}^{n} \frac{k}{nt} \times \frac{\binom{n}{k}\binom{m-n}{n-k}}{\binom{m}{n}}\right) \\
&= \frac{100}{t \times \binom{m}{n}} \times \left(\frac{1}{2}(t-1)(t+2) + \sum_{k=1}^{n} \binom{n-1}{k-1}\binom{m-n}{n-k}\right) \\
&= \frac{100}{t} \times \left(\frac{(t-1)(t+2)}{2 \times \binom{m}{n}} + \frac{1}{m-n+1}\right)
\end{aligned}$$

If credit is only given for entirely correct answers, then:

$$\begin{aligned}
\mathrm{RGS} &= 100 \times \frac{1 + 2 + \ldots + t}{t \times \binom{m}{n}} \\
&= \frac{50\,(t+1)}{\binom{m}{n}}
\end{aligned}$$

Now suppose (2) holds, so the student is told how many of their choices are correct. At this stage, the situation becomes too complicated for the RGS to be calculated precisely, so I wrote a program to simulate the random guesser.

In this simulation, all $\binom{m}{n}$ possible combinations of n options are generated and one is chosen at random to represent the n correct choices. Independently of this, a combination is picked as the random guesser's first attempt. The number of correct choices in this first attempt, c_1 say, is recorded. If the guesser's first attempt is not entirely correct, those combinations which do not share precisely c_1 choices with the first attempt are discarded, and one of the remaining possible combinations is chosen as the second attempt. This process continues until the student gets the answer entirely correct or runs out of attempts, at which stage the answer is given a score. The simulation is run a sufficient number of times to give a resulting average RGS which is stable to the required precision, in this case to the nearest whole percentage point.

3.2 Multiple Response with unknown number of correct responses

Here I wrote programs to simulate both situations (1) and (2). In both cases I made the assumption that the random guesser would start off by choosing at random between all possible subsets of questions, and then eliminate inconsistent subsets as before. This assumption is equivalent to supposing that in the first attempt the guesser selects each possible choice with probability $\frac{1}{2}$, independently of all other choices.

The only difference between the two simulations is in the way in which possible subsets of choices are eliminated. If (1) holds, and the student is simply told that their selection is incorrect, then only the subset chosen will be eliminated. If (2) holds, and the student is told c of their choices are correct, then all remaining choices must share precisely c choices with the previous selection.

3.3 Drag and Drop

Again, I wrote programs to simulate all possible situations:

- If each choices may be used an unlimited number of times, then to calculate all possible options we must consider all permutations of the choices, including each choice an unlimited number of times in the permutation.
- If each choice may only be used once, then to calculate all possible options we must consider all permutations of the choices, only including each choice once in the permutation.

Then, if (1) holds, and the student is simply told that their permutation is incorrect, then only the permutation chosen will be eliminated. If (2) holds, and the student is told c of their choices are correct, then all remaining permutations must have precisely c choices in the same position as they are in the previous permutation.

4 How to use the RGS tables and calculator

4.1 RGS tables

I have compiled four spreadsheets containing tabulated random guess scores for some of the different question types:

Multiple Response questions

- **multiple_response_1** Contains data for all multiple choice or multiple response questions where (1) holds: if the student does not get the question correct at first attempt, they are simply told that their answer is incorrect
- **multiple_response_2** Contains data for all multiple choice or multiple response questions where (2) holds: if the student does not get the question correct at first attempt, they are told how many of their choices are correct

In both cases, each tab contains data for a different number of total options. The rows represent the number of these options which are correct, and the columns show the number of attempts allowed. The upper table on each tab contains scores for when students are told how many options are correct, and the lower table contains scores for when they are not.

Drag and Drop questions

- **drag_and_drop_once_only** Contains data for all drag and drop questions where each choice may be put into at most one box
- **drag_and_drop_unlimited** Contains data for all drag and drop questions where each choice may be put into an unlimited number of boxes

In both cases, each tab contains data for a different number of boxes to be filled in. The rows represent the number of choices to put in the boxes, and the columns show the number of attempts allowed. These figures assume (2) holds: if the student does not get the question correct at first attempt, they are told how many of their choices are correct. If this assumption is not valid, the RGS calculator can be used (see Section 4.2).

4.2 RGS calculator

I have written a program which asks a number of questions to determine the precise question type, then outputs a simulated value for the RGS for that question. It will find random guess scores for a wider range of questions than those described in the rest of this report. For example, if after getting a question wrong, a student is told precisely which options were correct (instead of just the number of correct answers), the program can simulate the RGS. It also will output the RGS in the situation where all options need to be correct for the student to score at all (which can dramatically reduce the RGS in some cases).

To run this program:

- Open a command prompt and navigate to the folder `RGS_calc`
- Type `RGS_calc.exe` into the command prompt
- Follow the instructions!