

# An Automatic Quantification of the Randomness of Answering Correctly in Taking Traditional Multiple-choice Tests

Qingsong Gu

Shanghai University of Engineering Science, Shanghai, China

Michael W. Schwartz

St. Cloud State University, Minnesota, USA

**Abstract**—In taking traditional multiple-choice tests, random guessing is unavoidable yet nonnegligible. To uncover the “unfairness” caused by random guessing, this paper designed a Microsoft Excel template with the use of relevant functions to automatically quantify the probability of answering correctly at random, eventually figuring out the least scores a testee should get to pass a traditional multiple-choice test with different probabilities of answering correctly at random and the “luckiness” for passing it. This paper concludes that, although random guessing is nonnegligible, it is unnecessary to remove traditional multiple-choice items from all testing activities, because it can be controlled through changing the passing score and the number of options or reducing its percentage in a test.

**Index Terms**—quantification, random guessing, multiple-choice questions

## I. INTRODUCTION

Multiple choice (MC) is a form of an objective assessment in which respondents are asked to select only correct answers out of the choices from a list. A traditional multiple-choice test (MCT) consists of a collection of questions or incomplete statements, often referred to as *stems*, each having a number of associated *options*. A stem together with its options is often referred to as an *item*. One of the options within each item is the right answer while the others are all wrong answers, often referred to as *distractors* (Bush, 2015). Since the early 1900s, traditional MC item formats have achieved a position of dominance in learning assessment (Swartz, 2006) and multiple-choice tests (MCTs) have been widely used to replace or supplement constructed responses (problems, essays or oral questions), notably due to the pressures of growing students' numbers, budget constraints and the subjectivity of constructed forms of assessment (Becker & Johnston, 1999). Moreover, the wide use of information technology and convenient accessibility of computer networks have greatly helped the development of online MCTs, which allow for more flexibility in delivery as students take their tests at a time and place convenient to them (Krieg & Uyar, 2001; Nicol, 2007). Because of these advantages, traditional MCTs have played, and will continue to play, an indispensable role in educational assessment.

Unfortunately, traditional MCTs came under fire years ago. CQUniversity Australia decided to ban multiple-choice questions (MCQs). According to the Academic Board resolution (24 March 2014), no multiple-choice questions can be used in any exams after Term 2, 2014. Actually, the criticisms of traditional MCTs have never stopped, but this decision indicated that the disadvantages associated with MCTs seemed to be intolerable.

As we know, the most serious disadvantage of traditional MCTs is the limited types of knowledge that can be assessed. MCTs may only measure and foster students' ability to memorise or recall factual information rather than test higher levels of cognitive processes (Becker & Johnston, 1999). MCTs also fail to test students' ability to develop an argument – a skill actively encouraged in higher education (Paxton, 2005). However, one nonnegligible, yet easy to be neglected or forgotten, concern raised by critics of traditional MCTs is that test-takers can guess their way towards a substantial percentage of their grade (Dawson, 2014). Early in 1922, Hahn pointed out that one of the dangers in tests requiring alternative responses was the guessing factor. Students regard MC examinations favourably because: (1) they are perceived as objective, (2) marks can be improved through making correct guesses, and (3) it is possible for students to gain marks on MC tests even if they do not reach the end of the paper in the allotted time (Simkin & Kuechler, 2005).

Although it is often assumed that the effects of guessing are negligible and even some scholars have described the threat of guessing as “overrated” (Haladyna, 2004, p.113), the actual fact is that there is always a chance that examinees will answer items correctly by guessing when an exam includes MC items (Foley, 2016). Additionally, it should be noted that (1) guessing in taking traditional MCTs is random in most cases, having almost nothing to do with the knowledge being tested; (2) guessing in taking traditional MCTs is somewhat encouraged because there is no penalty for answering incorrectly; and (3) random guessing (RG) has become an effective strategy for some “lucky” examinees to “unfairly” pass some decisive examinations.

It is unnecessary to doubt the existence of RG in taking traditional MCTs. What is important is to figure out in what way and to what extent RG plays its role in taking traditional MCTs. In some sense, RG discussed here simply refers to the randomness of answering correctly (RAC), which is to be studied intensively in this paper through an automatic quantification designed with the help of Microsoft Excel.

## II. AUTOMATIC QUANTIFICATION

### A. Definition of Traditional MCTs

A MCT may have different numbers of options: 2, 3, 4, or 5. A MCT is regarded as a true-false (T/F) test when the number of options is 2. In testing practice, 3-option, 4-option and 5-option are often applied for a MCT, but the most widely-used is 4-option, so a traditional MCT usually refers to a 4-option multiple-choice test with only one correct answer in each item. In taking traditional MCTs, testees get points for answering correctly, but no penalty for answering wrongly.

### B. Problems to Be Solved

The automatic quantification in this paper is designed to solve two problems: (1) how to quantify RAC probability (hereafter denoted as “G”) in taking a traditional MCT; (2) how to quantify the “luckiness” for passing the test (hereafter denoted as “L”).

### C. Pre-research

It is not difficult to understand that, in taking traditional MCTs, the theoretical probability of RAC is 0.25 because each item has four options, with only one being the correct answer. The probability of RAC has been studied and interpreted by some mathematical statistics. For example,

$$B(m|n, 1/k) = \binom{n}{m} \left(\frac{1}{k}\right)^m \left(\frac{k-1}{k}\right)^{n-m}. \quad (1)$$

if  $k$  is the number of options in (1), the probability of guessing correctly exactly  $m$  of the  $n$  items is given by the binomial (Budescu & Bar-Hillel, 1993).

Foley (2016) introduced a FERG (full-exam random guessing) strategy to pass traditional MCTs through illustrating the probability of achieving different scores using a FERG strategy, as shown in Fig. 1. In a 100-item exam consisting of 4-option MC items, an examinee likely would obtain “25” (a 0.25 chance of guessing each question correctly multiplied by 100 items) by using FERG on this exam. However, 25 is only the average score one would obtain across many repeated FERG attempts. Approximately half of the time examinees would score higher than 25, and about 15% of the time examinees would score 30 or higher. Theoretically, there is a miniscule, but non-zero, chance that the examinee could score 100% correct.

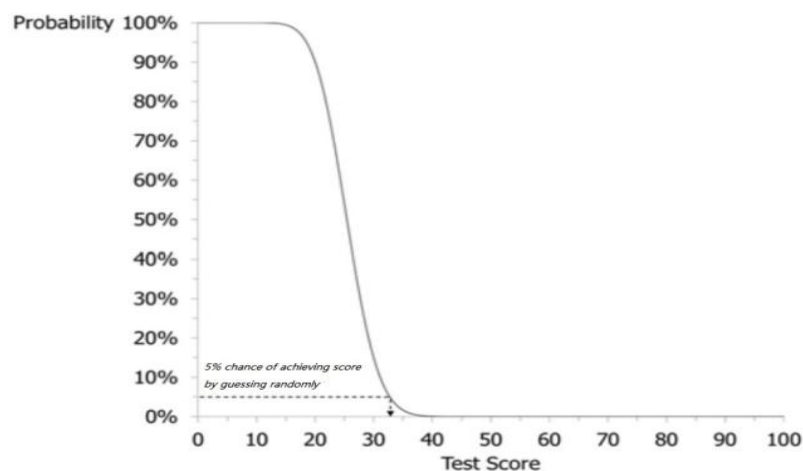


Figure 1. The probability of achieving different scores using a FERG strategy on a 100-item, 4-option MC exam.

Out of curiosity, we designed a game and often played it in the classroom to manually collect the data for the purpose of illustrating G in a traditional MCT. The game used to be played in the following way.

First, students were asked to write down at random any 20 answers for a 4-option and 20-item traditional MCT which had no questions, only to guarantee pure randomness. Students were not allowed to select only A, B, C or D as their answers and the 20 answers were suggested to be listed five by five for their convenience in marking. Normally, this was done within ONE minute. For example, Student A wrote as “ABCDD CBACD BABCD AABDC”.

Then, I picked out at random any one of the students, say Student A, to copy on the blackboard his or her 20 answers which were going to serve as the keys of the virtual no-question test and, according to these keys, other students were required to mark their own answers and write down the number of correct answers. For example, Student B wrote as “BBCAD CAACD DBBCA ACDDC”, earning up to 12 points.

After that, all students were encouraged to voluntarily announce their numbers of correct answers, so the whole class knew on the spot who was the “luckiest” person to achieve the highest score and who was the “least lucky” person to get the lowest one. To be “fair”, another student, say Student C, was picked out at random to copy his or her 20 answers as the keys and, according to the new keys, other students were asked to mark again their own answers and write down the new number of correct answers. This time, the “luckiest” and the least “lucky” persons would most probably change. For example, Student B earned only 5 points according to Student C’s answers “ABDDC CABBA DCCBC ABCCD”.

Last, we collected all the answer sheets, input the data, and conducted some statistics by means of Microsoft Excel to figure out the average value of G in the played games. Here is a sample of data statistics from five games played by 58 students.

TABLE I.  
THE AVERAGE VALUE OF G FOR A 4-OPTION AND 20-ITEM TRADITIONAL MCT IN FIVE GAMES PLAYED BY 58 STUDENTS

Order	1	2	3	4	5	Average	Percentage
The highest score	11	8	10	13	10	10	50%
The lowest score	1	1	2	1	1	1	5%
The number of students who get 10 to 20 correct answers	3	0	2	4	3	2	3.44%
The number of students who get 5 to 10 correct answers	17	12	11	24	15	16	27.58%
The number of students who get 0 to 5 correct answers	38	46	45	30	40	40	68.97%
The average probability of RAC	4.14	3.92	4.06	4.15	5.58	4.37	21.85%

We played the game many times in the classroom and collected plenty of data. It can be seen from Table I that the data of the same category were quite close between any two games in spite of some minor fluctuations and that the average value of G (represented by 4.37 and 21.85%) was expectedly close to the theoretical one (represented by 5 and 25%).

The data collected from the games were authentic, but it was not an easy thing to let the students know all the results as quickly as they expected because all the numbers of the collected data had to be manually input one by one. For this reason, we designed a Microsoft Excel template with the use of relevant functions to automatically quantify G in taking a traditional MCT, and eventually, to quantify L through figuring out the least RAC scores (hereafter denoted as “M”) for passing a traditional MCT with various G values and the “lucky” RAC scores (hereafter denoted as “N”) for passing the test.

D. Template Design

In this paper, a Microsoft Excel template was designed to automatically quantify the RAC of 1000 testees in taking a 4-option and 20-item traditional MCT. Interestingly, the testees and the test are not authentic but virtual. This template is a one-sheet file with three mutually-related areas that work together through Microsoft Excel functions, such as “CHOOSE”, “RANDBETWEEN”, “SUMPRODUCT”, “COUNTIF”, “MAX”, “MIN”, “SUM” and “AVERAGE”. The following is a detailed introduction to the three areas in the template.

The first area (See Fig. 2) presents the RAC of each testee in taking a 4-option and 20-item traditional MCT. The array “B4:B1003” is about the order of the testees, with most of them concealed for users’ convenience of viewing. The array “C4:V1003” presents all the values of answers which are randomly and automatically given by the computer. The values of RAC scores are listed for each testee in the array “W4:W1003”, and the values of G in the array “Z4:Z1003”. Row 3 is highlighted in red as the keys. The Microsoft Excel functions used in the first area are listed in detail in Table II.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
2	Testees	RG																				RAC Scores	RAC Probabilities			
3	Key	A	B	B	B	C	D	C	D	B	C	A	A	D	A	A	C	C	B	D	D	A	20	100.00%		
4	1	B	A	C	B	D	A	C	A	A	C	C	A	D	A	C	C	D	D	D	A	9	45.00%			
103	100	A	D	D	D	B	A	C	D	A	B	B	B	A	C	D	B	A	D	B	C	5	25.00%			
104	101	C	B	B	A	A	A	B	C	B	B	C	B	A	A	D	A	D	C	D	D	6	30.00%			
203	200	D	D	C	D	C	C	C	B	D	A	D	C	A	B	C	A	C	A	A	D	4	20.00%			
204	201	A	C	A	A	C	D	B	C	A	A	A	C	C	B	C	B	B	A	C	C	6	30.00%			
303	300	A	C	D	C	D	B	D	B	D	B	A	B	D	A	B	A	B	B	D	C	5	25.00%			
304	301	B	A	D	B	C	B	B	A	A	B	A	A	A	B	A	D	B	C	C	D	5	25.00%			
403	400	B	B	C	A	B	D	D	B	D	B	B	C	C	D	C	C	D	C	D	B	5	25.00%			
404	401	D	C	B	B	D	A	B	D	D	A	C	C	A	B	C	B	B	C	D	A	8	40.00%			
503	500	C	B	D	D	B	D	B	C	C	A	C	B	B	A	A	A	D	C	D	D	4	20.00%			
504	501	B	A	C	B	A	B	B	C	D	A	A	C	D	A	B	A	B	C	D	A	6	30.00%			
603	600	C	B	A	A	C	D	D	B	B	D	A	C	D	D	C	B	A	C	B	B	6	30.00%			
604	601	D	B	D	D	C	D	C	B	D	B	D	D	B	B	A	C	A	B	A	5	25.00%				
703	700	C	D	D	C	D	D	B	D	B	C	B	C	A	D	C	C	D	B	3	15.00%					
704	701	C	A	D	A	A	D	D	D	D	B	C	D	A	C	C	B	D	D	A	A	7	35.00%			
803	800	C	D	D	A	D	A	A	C	B	C	B	B	A	A	C	A	C	A	D	B	4	20.00%			
804	801	A	A	D	B	C	D	A	D	A	D	A	A	A	D	B	D	B	B	B	B	8	40.00%			
903	900	D	B	D	D	D	D	B	C	A	A	A	A	D	C	C	D	B	C	B	C	5	25.00%			
904	901	D	D	D	A	D	A	B	A	B	A	D	A	B	C	D	B	C	B	D	A	2	10.00%			
1003	1000	B	D	A	C	C	C	D	C	D	C	C	A	D	B	A	C	C	A	B	D	3	15.00%			

Figure 2. The RAC of each testee in taking a 4-option and 20-item traditional MCT (quantified in the first area).

TABLE II.  
THE MICROSOFT EXCEL FUNCTIONS USED IN THE FIRST AREA

Arrays	Functions	Examples		
		Cell	Formula	Number format
C4:V1003	CHOOSE; RANDBETWEEN	C4	=CHOOSE(RANDBETWEEN(1,4),"A","B","C","D")	Text
W4:W1003	SUMPRODUCT	W4	=SUMPRODUCT(N(C4:V4=\$C\$3:\$V\$3))	General
Z4:Z1003	DIVISION	X4	=W4/20	Percentage

The second area (See Fig. 3) is designed to automatically calculate the G values of all the testees in taking the test. To get a clearer look at the whole data, the 1000 testees are divided into 10 groups. As shown in Fig. 2, the data of the first 100 testees are located in the array "B4:Z103", the second in "B104:Z203", the third in "B204:Z303", and so on. In Fig. 3, the 10 groups follow the order from B1008 to B1017. Values in the array "C1008:C1017" are the maximum RAC scores in each group, and "D1008:D1017" the minimum. In the array "E1008:Y1017" are the numbers of testees with different RAC scores ranging from 0 to 20. In the array "Z1008:AD1017" are the numbers of testees in different ranges of RAC scores. In the array "AE1008:AE1017" are the sums of RAC scores for each group. The array "AF1008:AF1017" deals with the average G values in each group. In Row 1018 are all the average values of different arrays for the 10 groups. The Microsoft Excel functions used in the second area are listed in detail in Table III.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1006	Order	MAX	MIN	The Number of Testees with Different RAC Scores																				The Number of Testees in Different Ranges of RAC Scores				SUM of RAC Scores	Average RAC Probabilities				
1007				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	0-2	3-6	7-8	9-10	11-20				
1008	1	10	1	0	4	6	17	17	18	17	6	10	4	1	0	0	0	0	0	0	0	0	0	0	10	69	16	5	0	495	24.75%		
1009	2	12	1	0	1	4	16	15	14	23	17	6	1	0	2	1	0	0	0	0	0	0	0	5	68	23	1	3	535	26.75%			
1010	3	10	1	0	2	5	8	27	24	16	8	6	1	3	0	0	0	0	0	0	0	0	0	7	75	14	4	0	503	25.15%			
1011	4	10	2	0	0	6	20	23	18	16	9	4	3	1	0	0	0	0	0	0	0	0	0	6	77	13	4	0	482	24.10%			
1012	5	9	0	1	3	5	16	16	19	16	13	5	6	0	0	0	0	0	0	0	0	0	0	9	67	18	6	0	501	25.05%			
1013	6	10	0	1	3	6	17	16	12	16	21	3	3	2	0	0	0	0	0	0	0	0	0	10	61	24	5	0	504	25.20%			
1014	7	10	0	1	1	9	19	17	10	20	12	6	4	1	0	0	0	0	0	0	0	0	0	11	66	18	5	0	492	24.60%			
1015	8	13	0	2	6	7	18	18	19	15	7	3	2	1	1	0	1	0	0	0	0	0	0	15	70	10	3	2	455	22.80%			
1016	9	10	1	0	4	7	11	18	20	22	12	4	1	1	0	0	0	0	0	0	0	0	0	11	71	16	2	0	490	24.50%			
1017	10	9	0	1	1	4	12	20	25	14	10	8	5	0	0	0	0	0	0	0	0	0	0	6	71	18	5	0	513	25.65%			
1018	Average	10	0.6	0.6	2.5	5.9	15	19	18	18	12	5.5	3	1	0.3	0.1	0.1	0	0	0	0	0	0	9	69.5	17	4	0.5	497.10	24.86%			

Figure 3. The average RAC probability of all the testees in taking a 4-option and 20-item traditional MCT (quantified in the second area).

TABLE III.  
THE MICROSOFT EXCEL FUNCTIONS USED IN THE SECOND AREA

Arrays	Functions	Examples		
		Cell	Formula	Number format
C1008:C1017	MAX	C1008	=MAX(W4:W103)	General
D1008:D1017	MIN	D1008	=MIN(W4:W103)	General
E1008:Y1017	COUNTIF	E1008	=COUNTIF(W4:W103,0)	General
Z1008:AD1017	SUM	Z1008	=SUM(E1008:G1008)	General
AE1008:AE1017	SUM	AE1008	=SUM(W4:W103)	General
AF1008:AF1017	DIVISION	AF1008	=AE1008/2000	Percentage
C1018:AF1018	AVERAGE	C1018	=AVERAGE(C1008:C1017)	General

The third area (See Fig. 4) is designed to automatically show the average "luckiness" for each group to pass the test. The array "B1022:B1031" presents the order of the 10 groups of testees. The values in the array "C1022:C1031" in this area are the average G values, equaling those in the array of "AF1008:AF1017" in the second area. In the array "H1022:H1031", 20 is a constant, representing the total scores of the test (hereafter denoted as "T"). In the array "K1022:K1031", 60% is also a constant, representing the supposed percentage of passing scores out of the total scores (hereafter denoted as "P"). In the array "Q1022:Q1031" are the average values of M for passing the test. In the array "W1022:W1031" are the average values of N for passing the test. The values in the array "AA1022:AA1031" represent the average values of L (=N/T). In Row 1032 are all the average values of different arrays for the 10 groups. The Microsoft Excel functions used in the third area are listed in detail in Table IV.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1020	Order	G		T		P		M		N		L														
1021		(RAC Probability)		(Total Scores)		(Percentage of Passing Scores)		(The Least Passing Scores)		(The "Lucky" Scores)		(N/T)														
1022	1	24.75%		20		60%		9		3		13.16%														
1023	2	24.30%		20		60%		9		3		12.84%														
1024	3	22.85%		20		60%		10		2		11.85%														
1025	4	26.00%		20		60%		9		3		14.05%														
1026	5	24.30%		20		60%		9		3		12.84%														
1027	6	25.55%		20		60%		9		3		13.73%														
1028	7	23.30%		20		60%		10		2		12.15%														
1029	8	25.40%		20		60%		9		3		13.62%														
1030	9	23.20%		20		60%		10		2		12.08%														
1031	10	25.35%		20		60%		9		3		13.58%														
1032	Average	24.50%		20		60%		9		3		12.99%														

Figure 4. The "luckiness" for passing a 4-option and 20-item traditional MCT (quantified in the third area).

TABLE IV.  
THE MICROSOFT EXCEL FUNCTIONS USED IN THE THIRD AREA

Arrays	Functions	Examples		
		Cell	Formula	Number format
C1022:C1031	EQUATION	C1022	=AF1008	Percentage
Q1022:Q1031	SUBTRACTION, MULTIPLICATION and DIVISION	Q1022	=H1022*(K1022-C1022)/(1-C1022)	Number
W1022:W1031	SUBTRACTION and MULTIPLICATION	W1022	=H1022*K1022-Q1022	Number
AA1022:AA1031	DIVISION	AA1022	=W1022/H1022	Percentage
C1032:AA1032	AVERAGE	C1032	=AVERAGE(C1022:C1031)	Number; Percentage

E. Data Analysis

The quantification in the Microsoft Excel template introduced above is run automatically, meaning that, whenever the F9 button is pressed, all the variable values in the three areas will be updated immediately. It is so exciting that each second F9 is pressed, the corresponding data of the 1000 testees come out; that is to say, if F9 were pressed 10 times, all the data of 10,000 testees would come out within 10 seconds. Certainly, this template works far more efficiently and conveniently than in the way I used to get the results from the games played in the classroom.

In the first area, RAC is the most important data to be collected for the whole quantification. As is shown in Fig. 5, the values of RAC scores in the first group of testees center around 3-8, ranging from 0 to 12. The similar data can be collected from other groups. Interestingly, no matter how many times I pressed F9, the highest score never touched 14 or more, but in the games I played in the classroom, 14, 15, and even 16 really happened.

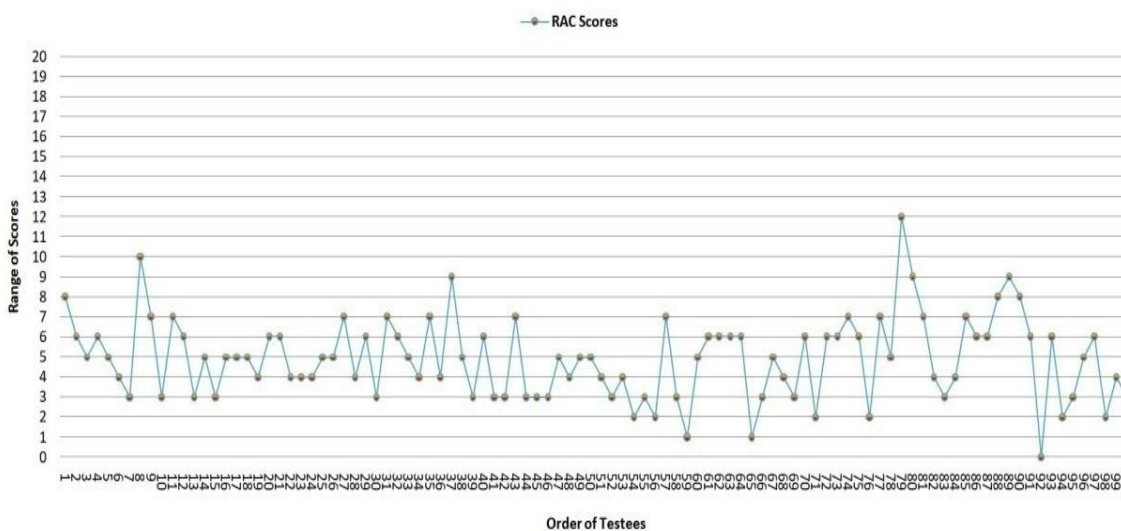


Figure 5. The values of RAC in the first group of testees.

In the second area, the average G values of each group can be seen and understood from various perspectives. In average, the maximum RAC score is 10, and the minimum one close to 1. The average numbers of 100 testees with different RAC scores ranging from 0 to 20 center around those in the range of RAC scores from 3-8, as is illustrated in Fig. 6, and roughly in Fig. 7. From the average sum of RAC scores (497.10 in Fig. 3), it is easy to get the average G value (24.86% in Fig. 3), which are very close to 25%, the theoretical G value.

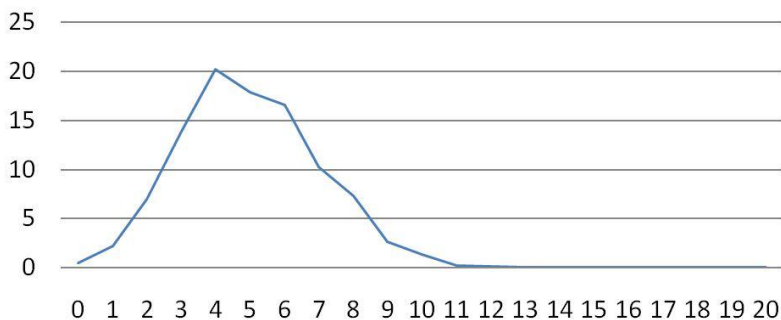


Figure 6. Average numbers of 100 testees with different RAC scores.

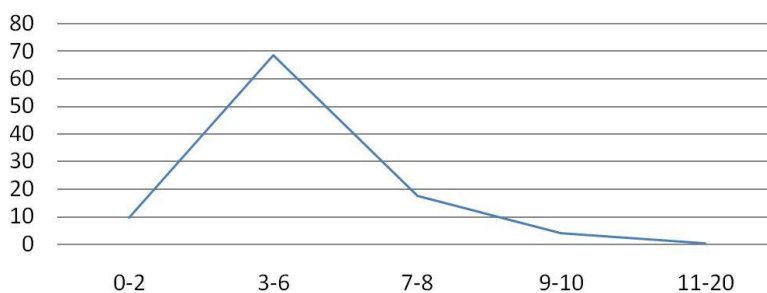


Figure 7. Average numbers of 100 testees in different ranges of RAC scores.

In the third area, the first and most important thing is to calculate M.

$$(T - M)G + M = TP \quad (2)$$

$$M = \frac{T(P - G)}{1 - G} \quad (3)$$

The essential relations among G, T, P and M can be demonstrated in (2), through which M can be calculated as in (3). From M, it is easy to calculate N as in (4) and L in (5), which both clearly indicate the “luckiness” in passing the test. In this template, the automatic quantification is conducted for a 4-option and 20-item traditional MCT, the passing score of which is 12, 60 percent of the total scores, so T is constantly 20, P is constantly 60%, and R is approximately 25%. It can be seen from this area that the average M for the test is 9, which is 3 less than 12, and the average L is around 13%, which is more than half of R. Actually, the “luckiness” in passing the test cannot be fully demonstrated by M in this template, which looks like a small number because T is only 20, but L, presented as a percentage, can directly show the “luckiness”. The average L values come out as expected and can be objectively calculated, but the result is astonishing since 13% is not a small percentage.

$$N = TP - M = TP - \frac{T(P - G)}{1 - G} = T \frac{G(1 - P)}{1 - G} \quad (4)$$

$$L = N/T = \frac{G(1 - P)}{1 - G} \quad (5)$$

#### F. Discussion

No matter how many times F9 is pressed in the template, the average L values center around 13% when P is 60%. The possibility of guessing at MC items means some unfairness when it comes to testing, in particular to selection assessment. If there are two examinees of equal (moderate) level of ability, then one of them might fail the test while the other passes (Kubinger, Holocher-Ertl, Reif, Hohensinn & Frebort, 2010). It is indisputable that RG in taking a traditional MCT does diminish the validity and reliability of the test, but how to minimize the negative effects of RG has been a headache ever since the practice of traditional MCTs.

To deal with guessing effects in a fair manner, Birnbaum (1968) developed from Item Response Theory (IRT) a 3-PL model, which provided an item-discrimination parameter and an item-guessing parameter. Also, a lot of research has been done on scoring rules to solve this guessing problem. Number Right (NR) is the simplest scoring rule which just counts the number of correct responses. Since 1953, the Scholastic Aptitude Test (SAT) employed a formula-scoring rule which corrected for random guessing by penalizing incorrect responses, while being neutral regarding omitted items (Budescu & Bar-Hillel, 1993). Although many approaches have been proposed to reduce RG in taking traditional MCTs, none would be superior.

In the Microsoft Excel template designed in this paper, L considerably reflects how much RG does in taking a traditional MCT. Since L equals N/T, L seemingly has to change with both N and T, but it is clear in (5) that L is only connected with P and G.

Although M and N change with T, L has little to do with T. As we know, the reliability of a multiple-choice test is partly governed by the number of items it contains; however, longer tests are more time consuming to take, and for some subject areas, it can be very hard to create new test items that are sufficiently distinct from previously used items (Bush, 2015). Consequently, it is not the right choice to reduce RG by simply raising the number of items. The feasible way is to change P and G.

The automatic quantification of G and L in the template is done on a 4-option and 20-item traditional MCT. When P is 60%, L is approximately 13%. However, when P is 50%, L will be about 17%, a bit more than 13%, which exactly verifies what CQUniversity Pro Vice-chancellor Professor Rob Reed (Hinchliffe, 2014) said, “With a pass mark of 50 per cent which is the standard pass mark across Australia, you can know about a third of the answer and guess about

two thirds at that 50 per cent level.” More L variants in Fig. 8 are calculated as in the array “C3:F7”. For example, the formula of C3 is “=(1/C2)\*(1-\$B\$3)/(1-1/C2)”, in which “1/C2” is a G value.

	A	B	C	D	E	F
1	L		Number of options			
2			2	3	4	5
3	P	50%	50%	25%	17%	13%
4		60%	40%	20%	13%	10%
5		70%	30%	15%	10%	8%
6		80%	20%	10%	7%	5%
7		90%	10%	5%	3%	3%

Figure 8. L changes with P and G.

### III. CONCLUSION

While it is true that the probability of obtaining a high score through guessing is small (Sax, 1989), it cannot be said that RG is negligible. No doubt, with the help of RG, some testees really passed one or more decisive assessments in an unfairly lucky way. Even so, it is unnecessary to remove traditional MC items from all testing activities, but possible to reduce the percentage of them at least. Now that RG has little to do with the number of items, P and G may be changed to keep RG under control. A successful example is the Knowledge Test for a Minnesota driver’s license. The test contains 40 questions in the form of traditional MC and T/F and the testee must obtain a score of 80 percent to pass. First, the number of items in the test is moderate, so the test doesn’t take a long time and the questions can be renewed on a large scale to reduce repetition. Moreover, MC and T/F are allocated appropriately to cover the knowledge and understanding of Minnesota traffic laws and road signs. Most important of all, 80% is a high P value, so L in the test is very low, or rather, RG in the test is limited to a satisfying range.

### ACKNOWLEDGMENTS

First, this paper is fully indebted to the grant from the 2017 Overseas Research Scholar Program for Young and Middle-aged Teachers of Higher Education in Shanghai, which was sponsored by Shanghai Municipal Education Commission. Second, this paper was supported by College English Listening (s201718001), one of the 2017 Key-curriculum Construction Projects for Undergraduates of Higher Education in Shanghai. Third, the authors would like to give special thanks to the White Horse Gang who encouraged and supported us during the time when we were writing this paper in St. Cloud.

### REFERENCES

- [1] Assessment Policy and Procedure (Higher Education Coursework) (current version - from Term 1, 2018). (2018). <https://www.cqu.edu.au/policy?collection=policy&form=policy&query=Assessment+Policy+and+Procedure+%28Higher+Education+Coursework%29+%28current+version+-+from+Term+1%2C+2018%29> (accessed 28/4/2018).
- [2] Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record* 75, 348–357.
- [3] Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability. In: Lord, F.M. and Novick, M.R. (eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 395–479.
- [4] Budesu, 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.
- [5] Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education* 40.2, 218–231.
- [6] Dawson, P. (2014). Why it’s not just assessors who benefit from multiple choice(s). <https://theconversation.com/why-its-not-just-assessors-who-benefit-from-multiple-choice-s-31973> (accessed 28/4/2018).
- [7] Foley, P. B. (2016). Getting Lucky: How Guessing Threatens the Validity of Performance Classifications. *Practical Assessment, Research & Evaluation* 21.3, 1–23.
- [8] Hahn, H. H. (1922). A Criticism of Tests Requiring Alternative Responses. *The Journal of Educational Research* 6.3, 236–240.
- [9] Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- [10] Hinchliffe, J. (2014). CQ University scraps multiple choice exams in an Australian first. <http://www.abc.net.au/news/2014-09-23/cqu-scraps-multiple-choice-exams-in-an-australian-first/5763226> (accessed 14/4/2018).
- [11] Krieg, R. G., & Uyar, B. (2001). Student performance in business and economics statistics: Does exam structure matter? *Journal of Economics and Finance* 25, 229–241.
- [12] Kubinger, D. K., Holocher-Ertl, S., Reif, M., Hohensinn, C. & Frebort, M. (2010). On Minimizing Guessing Effects on MultipleChoice 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.

- [13] Minnesota Driver's Manual (no date). [https://dps.mn.gov/divisions/dvs/forms-documents/Documents/Minnesota\\_Drivers\\_Manual.pdf](https://dps.mn.gov/divisions/dvs/forms-documents/Documents/Minnesota_Drivers_Manual.pdf) (accessed 28/4/2018).
- [14] Multiple choice (no date). [https://en.wikipedia.org/wiki/Multiple\\_choice](https://en.wikipedia.org/wiki/Multiple_choice) (accessed 28/4/2018).
- [15] Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education* 31, 53–64.
- [16] Paxton, M. (2000). A linguistic perspective on multiple choice questioning. *Assessment and Evaluation in Higher Education* 25, 109–119.
- [17] Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadworth Publishing Company.
- [18] Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: what is the connection? *Decision Sciences Journal of Innovative Education* 3, 73–97.
- [19] Swartz, M. S. (2006). Acceptance and Accuracy of Multiple Choice, Confidential-Level, and Essay Question Formats for Graduate Students. *Journal of Education for Business* March/April, 215–220.



**Qingsong Gu**, associate professor from School of Foreign Languages at Shanghai University of Engineering Science, earned a MA degree in English from Shanghai Maritime University and has been teaching College English in China for more than 15 years. His research focuses on Teaching English as a Second or Foreign Language. Currently, he is a research scholar at St. Cloud State University.



**Michael W. Schwartz**, associate professor and director of Intensive English Center at St. Cloud State University, earned a Ph.D degree in Educational Linguistics from University of New Mexico. His teaching includes world Englishes, second language writing, discourse analysis and teacher development. He was a sponsor of Qingsong Gu who completed his research program at St. Cloud State University from August 26, 2017 to August 25, 2018.