JOURNAL OF CHEMICAL EDUCATION
Article
pubs.acs.org/jchemeduc

Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed To Discourage Guessing

Mark L. Campbell*

Chemistry Department, United States Naval Academy, Annapolis, Maryland 21402, United States

ABSTRACT: Multiple-choice exams, while widely used, are necessarily imprecise due to the contribution of the final student score due to guessing. This past year at the United States Naval Academy the construction and grading scheme for the department-wide general chemistry multiple-choice exams were revised with the goal of decreasing the contribution of the score due to guessing. A variable number of response choices were used based on question type, and students were encouraged not to guess by giving partial credit for answers left blank. Results of the study are presented.

KEYWORDS: *First-Year Undergraduate/General, Testing/Assessment, Problem Solving/Decision Making, Curriculum*

INTRODUCTION

Multiple-choice exams are widely used in high enrollment courses due to their perceived objective nature and their ease in grading and statistical analysis. Various factors are attributed to the reliability of multiple-choice exams including the number of questions and response choices,[1−4] the wording of the questions,[5] response length[6] or order,[7] number right vs formula scoring,[8,9] and polytomous versus dichotomous scoring.[10] A guide to best practices in constructing multiple-choice exams has recently been published in this *Journal*.[11] Exams with a small number of questions are subject to reliability issues related to the contribution of the final score attributed to random guessing.[12] Thus, the objective of this project is to devise an exam scheme to minimize guessing with the goal to improve reliability under the constraints of using a multiple-choice exam with a low number of total questions. We report here the strategy along with analysis of the exam results using the revised method.

METHOD

At the Naval Academy, we give three multiple-choice exams to approximately 1000 students every semester. Two midterm exams are given at the 6-week and 12-week points in the semester and consist of only 25 questions each due to the time constraint (50 min) of the exam period. Over the past 10 years our two "midterm" exams have used questions with four response choices. No partial credit was given for distractor answers nor was a penalty assessed for guessing so that the students were encouraged to answer all 25 questions. Historically, less than 1% of the answers to questions have been left blank indicating the students understood the benefit of guessing even when the student had no knowledge of the question asked. About half of the questions used for exams come from a department exam database[13] so that exams can be designed which fall near a reasonably predictable average, preferably in the 70s. This past year the exam construction and grading were changed in two ways. First, the questions had a variable number of response choices based on the question type. The number of response choices had three variations. Questions with responses requiring reading sentences or requiring thoughtful analysis of each response choice were limited to four response choices. Questions with one-word inspection answers or calculation questions with expected averages of greater than 60% were limited to five response choices. Calculation questions with expected averages of 60% or less had 10 response choices. Second, students were encouraged to leave questions blank if they were unable to reduce their response choices to 3 or less by giving 1/4 credit for a question left blank. The following instruction was printed on the front of the exam and read to students by the instructor prior to each exam:

> Your score will be determined by the sum of 4 points for each question answered correctly and 1 point for each question left blank. If you have some knowledge of a question and are able to limit the answer choices to 3 or less, your chances of selecting the correct answer are improved, and answering such questions is likely to improve your score. **It is unlikely that pure guessing will raise your score; it might lower your score.**

The reasoning behind the three different numbers of response choices is as follows. The reliability of multiple-choice exams depends on both the total number of questions asked and the number of response choices for each question. Reliability is increased as each factor is increased as long as proper exam-writing practices are followed; i.e., multiple questions testing the same concept or use of implausible response choices increase exam time required without improving exam validity and reliability.[11] In fact, asking different versions of the same question or similar questions (clones) can artificially inflate measurements of exam reliability. For questions in which each response must be thoughtfully analyzed individually, the time required to answer the question will depend on the number of response choices. In those instances, four response choices require less time than five or ten response choices. For calculation questions, the required time to select the right answer in a list is minimal compared to the time required to calculate the answer from the given data so that increasing the number of response choices in calculation questions has a negligible effect on the time. Therefore, for calculations, the number of responses has an insignificant effect on the time required for the exam as long as there is a logical

DOI: 10.1021/ed500465q

order to the response choices; i.e., ranked from low to high or high to low as opposed to a random distribution. Certain types of inspection problems do not require careful analysis for each response choice but are answered by picking out the correct answer from a list. In those cases, five response choices are more appropriate as long as all the response choices are plausible. As an example, a question which asks which of the following elements has the most unpaired electrons requires each response choice to be analyzed, while a question asking how many unpaired electrons does oxygen have requires only determining that oxygen has 2 unpaired electrons and looking for that answer. The former question should have only four response choices, while the latter could have five without increasing the time required for answering the question. Because there are normally time limits for exams, the number of response choices must be carefully chosen so that the number of questions asked can be maximized. The choice of either 5 or 10 response choices is a pragmatic choice based on the answer sheet we purchase. The answer sheet we use has five answers per number so that the number of answer choices is maximized in increments of five. We determined five responses were sufficient for easier questions in which the number of random guesses is expected to be small. Ten answer choices were deemed to be sufficient for the more difficult problems; i.e., 15 choices would be overkill.

The rationale for giving 1/4 point for a blank answer is as follows. The culture of the Naval Academy requires using a grading scheme in which grades follow 90/80/70/60 cutoffs. The desire was to maintain the same difficulty as previous exams while also maintaining an acceptable exam average (between 70 and 80%). Penalizing guessing by deducting points would inevitably decrease the exam average compared to previous years where guessing was not penalized. Adding 1/4 credit for blank answers has the advantage of statistically resulting in approximately the same average exam score as in previous years because the average guess score for a four-response question statistically results in 1/4 credit for all answers which were random guesses. Adding 1/4 credit also has the advantage of discouraging guessing more compared to penalizing (i.e., negative points) wrong answers.[14] Formula scoring has been shown to result in an increase in exam reliability.[12] Formula scoring has also been shown to improve validity,[15] but exam validity is not the focus of this study. Burton also states that guessing can be said to involve intellectual dishonesty and limiting random guessing provides better information on what the student knows and does not know, to the benefit of teaching.[1] Another important effect of formula scoring is that points are lost for confidently held disinformation which improves test reliability.[16]

■ EFFECT OF GUESSING ON EXAM RELIABILITY

It is generally understood that a well-constructed exam should result in equal scores for students with equal knowledge and that students with more knowledge should score higher than students with less knowledge. Unfortunately, multiple-choice exams have difficulty meeting these goals due to the probabilistic distribution of scores for students who guess on questions for which they do not know the correct answer. Burton has previously published quantitative aspects of how random guessing influences exam scores.[17] He profoundly proposes that most teachers do not fully appreciate the influence of guessing on exam reliability. If teachers understood the statistics of guessing, then it is likely that multiple-choice exams would be constructed differently. Here it is demonstrated how guessing decreases exam reliability. The probability, $P(x)$, of a student guessing x correct answers when guessing n times on an exam with q response choices is given by the Bernoulli distribution formula:

$$P(x) = \frac{n!}{x!(n-x)!}\left(\frac{1}{q}\right)^{x}\left(\frac{q-1}{q}\right)^{n-x}$$

Take the case of a 25 question exam with four response choices. Assume 100 students know the correct answers to $25 - n$ questions and randomly guess on n questions. Table 1 below shows the statistically expected distribution of scores where n is 8, 6, and 4. It is readily apparent from the table that students with the same knowledge show a wide range of scores, while the table also shows that some students with less knowledge get higher scores than those with more knowledge. From the table we can see statistically that 32 students (second row, 21 + 9 + 2 = 32) who know the correct answers to 17 questions will score as high or higher on the exam than 54 students (third row, 18 + 36 = 54) who know the correct answers to 19 questions. And 46 students (third row, 30 + 13 + 3 = 46) who know the correct answers to 19 questions will score as high or higher than 32 students who know the correct answers to 21 questions. The distribution is shifted to higher scores with a wider distribution if there is partial knowledge such that some of the response choices can be eliminated. Partial knowledge can also result in the choice of distractor answers, but partial knowledge on average raises, rather than lowers, test scores with a concomitant increasing scatter of the scores and decreasing exam reliability.[17] It is also worth noting that guessing is likely to benefit the less-prepared student more than the well-prepared student.

Table 1. Distribution of Student Scores[a] Based on Knowing the Answers to 25 − *n* Questions Given Four Response Choices

	Students (*N* = 100) with a Given Score, %							
Number of Guesses (*n*)	68	72	76	80	84	88	92	96
8	10	27	31	21	9	2		
6			18	36	30	13	3	
4					32	42	21	5

[a]The test has a 100-point scale, with 4 points per question.

The benefit of increasing the number of response choices on exam reliability can be seen in the two tables below. Tables 2 and 3 represent the same data as Table 1 except for five and 10 response choices, respectively.

Table 2. Distribution of Student Scores[a] Based on Knowing the Answers to 25 − *n* Questions Given Five Response Choices

	Students (*N* = 100) with a Given Score, %							
Number of Guesses (*n*)	68	72	76	80	84	88	92	96
8	17	33	29	15	5	1		
6			26	39	25	8	2	
4					41	41	15	3

[a]The test has a 100-point scale, with 4 points per question.

Optimally, students with the same knowledge should get the same score. As seen from the tables, this goal is not achieved when guessing occurs during the exam. If we can encourage the students to leave answers blank for questions they do not know, then we closer approach the goal; i.e., for the students who know 17 correct answers, if we could get all the students to leave the

DOI: 10.1021/ed500465q

Table 3. Distribution of Student Scores[a] Based on Knowing the Answers to 25 − *n* Questions Given 10 Response Choices

	Students (N = 100) with a Given Score, %						
Number of Guesses (n)	68	72	76	80	84	88	92
8	43	38	15	3	1		
6			53	35	10	2	
4					66	29	5

[a] The test has a 100-point scale, with 4 points per question.

remaining 8 blank, then they would all get the same score. The more answers left blank on those 8 questions, the closer is approached the ideal case.

Another underappreciated consequence of using multiple-choice exams is the effect of guessing on the average of a difficult question. Take the case of a multiple-choice question with four response choices where the students are encouraged to answer all the questions. For a question with an average of 70%, 30% of the responses were answered incorrectly. Statistically, if random guessing is assumed for the incorrect responses, then another 10% of the correct answers were due to guessing so that the "true" average is 60%. Table 4 is a tabulation of the statistical "true" scores for a question based on the measured average. Alarmingly, a multiple-choice question with only four response choices with a 50% average has a true average of only 33.3%. Thus, statistically 1/3 of the questions answered correctly were due to correct guesses. A "difficult" four response question inevitably results in a significant percentage of correct answers being the result of guessing. This table applies best to calculation questions where all the answer choices are plausible and common-mistake distractor responses are not included in the response choices. Implausible response choices wherein partial knowledge allows the student to eliminate some response choices would tend to make the difference between the measured and real scores larger. Including common-mistake distractor responses would make this difference smaller.

Table 4. Statistically Calculated "True Score" from a Measured Score for a Multiple-Choice Question Assuming Random Guessing with Four Response Choices

Measured Score (MS)	True Score[a]
90	86.7
85	80.0
80	72.7
75	66.7
70	60.0
65	53.3
60	46.7
55	40.0
50	33.3
45	26.7
40	20.0

[a] Calculated from the equation: true score = $(4/3)MS-33.3$.

Tables 5 and 6 illustrate how partial knowledge and incorrect information (misconceptions) affect the "true" score for a question with a measured score of 70%. Example 1 in Table 5 is the baseline score (maximum "true" score) for the case of no partial information. In Example 2, partial knowledge allows responses (C) and (D) to be eliminated leaving a choice between the correct answer (A) and response (B). Choosing between (A) and (B) being random means that if 300 students guessed (B), then another 300 also guessed (A). Therefore, only 400 of the correct answers were from students who were certain of their answer resulting in a "true" or "certain" score of 40%. In Example 3, partial knowledge allows response (D) to be eliminated resulting in 150 guesses for responses (A), (B) and (C) so the "true" score is 55%. In Example 4, partial knowledge allows everyone to eliminate response (D). However, 300 students cannot eliminate any of the other three responses so randomly guess among the three choices, while 200 students can eliminate both responses (C) and (D) so randomly guess among choices (A) and (B). Thus, 200 students will guess the correct response resulting in a "true" score of 50%. Finally, in Example 5, 120 students cannot eliminate any responses, 180 can eliminate response (D) and 180 can eliminate both responses (C) and (D). Thus, a total of 180 students will have guessed correctly resulting in a 52% "true" score.

Examples 1 and 2 in Table 6 result in the baseline score (maximum "true" score) for the case of incorrect knowledge and no random guessing. In the case of Example 1, responses (A) and (B) are chosen by all the students based on either correct knowledge (A) or incorrect knowledge (B). In Example 2, all the incorrect responses are chosen in equal amounts because the student population is equally distributed between three equally plausible misconceptions (probably highly unlikely). The main point is that if the students are only choosing among the correct answer and answers which represent the misinformation among all the students, then there will be no guessing and the measured score and "true" score will be equal. In Example 3, 75 students have a misconception which causes them to choose response (B) and 300 students randomly guess. Thus, 75 students guess correctly resulting in a 62.5% "true" score. Finally, in Example 4, 100 students have a misconception causing them to choose response (B), 50 students have a misconception which causes them to choose response (C) and 200 students randomly guess. Consequently, 50 students guess correctly resulting in a "true" score of 65%.

To a lesser extent, the overall "real" average of an exam can be estimated from Table 4. However, the "real" average would involve greater uncertainty based on how much deviation there is

Table 5. Statistically Calculated "True" Score from a Measured Score of 70% for a Multiple-Choice Question Assuming Some Partial Knowledge along with Random Guessing for Four Response Choices

Example	(A) (Correct)	(B)	(C)	(D)	Number of Guesses	Guessed Right	"True" Score,[a] %
1	700	100	100	100	400	100	60
2	700	300	0	0	600	300	40
3	700	150	150	0	450	150	55
4	700	200	100	0	500	200	50
5	700	180	90	30	480	180	52

[a] Score for those students who were certain of the answer.

Table 6. Statistically Calculated “True” Score from a Measured Score of 70% for a Multiple-Choice Question Assuming Some Incorrect Information along with Random Guessing for Four Response Choices

Example	(A) (Correct)	(B)	(C)	(D)	Number of Guesses	Guessed Right	“True” Score,[a] %
1	700	300	0	0	0	0	70.0
2	700	100	100	100	0	0	70.0
3	700	150	75	75	300	75	62.5
4	700	150	100	50	200	50	65.0

[a] Score for those students who were certain of the answer.

from the assumed ideal-question type (i.e., questions in which all of the wrong answers result from random guessing). Thus, the more short-answer concept questions with common-mistake distractors and calculation questions with common-mistake lures result in the difference between the measured exam average and the “true” exam average being smaller than in the table.

Mathematically, one can calculate the expected distribution for a 25-question exam taking into account random guessing, partial knowledge and misconceptions. The difficulty is there are an incredible number of possible combinations. Table 7 is the distribution for the case assuming the student knows the answer with certainty for 16 questions, randomly guesses on three questions, uses partial knowledge to eliminate two of the response choices to randomly guess from the two remaining choices, and has misconceptions on the three remaining questions to confidently choose the wrong answers.

Table 7. Distribution of Student Scores[a] Based on Knowing with Certainty 16 Questions, Randomly Guessing on 3 Questions, Using Partial Knowledge on 3 Questions, and Having Misconceptions on 3 Questions

Students (N = 100) with a Given Score, %					
64	68	72	76	80	84
5	21	33	27	11	3

The average score for a student in this situation is 73%. Again, it should be apparent that multiple-choice exams yield a wide distribution of scores for students with the same knowledge due to the probabilistic nature of guessing.

TEN-RESPONSE QUESTIONS: ANALYSIS OF STUDENT DATA

Ten-response questions are not a new idea. The first article describing computer-graded exams in this *Journal* described exams with 10 response choices.[18] Box 1 shows a question from our exam database that has been asked three times as a four-response question and once as a ten-response question. The results for this question for the four different years are shown in Table 8.

Box 1. Question 1

A 1.50 L buffer solution is prepared by mixing 0.55 mol $HC_2H_3O_2$ and 0.35 mol $NaC_2H_3O_2$ in water. Then, 0.05 mol of NaOH is added to the solution. What is the resultant pH of the solution? Assume the volume remains constant throughout. K_a for $HC_2H_3O_2$ is 1.8 × 10^{-5}.

A. 0.10
B. 4.78
C. 4.47
D. 4.56
E. 4.65
F. 4.74
G. 4.84
H. 4.92
I. 5.01
J. 5.10

The relative academic performance of these classes can be estimated by comparing the final exam scores on questions from the same ACS exam; the students from 2005, 2006, 2007, and 2014 averaged 77.7%, 74.6%, 73.9% and 75.7%, respectively. Using the Wilcoxon signed-rank test, 2014 performed worse than 2005 at the 97% confidence level, while 2014 performed better than 2007 at the 95% confidence level and better than 2006 at only the 82% confidence level. The number of students in 2014 who guessed correctly is considerably reduced; i.e., 334 were enticed not to guess and the probability of randomly guessing correctly were reduced from 1 out of 4 to 1 out of 10. The simplest analysis of the data for this question would be if we assume the students from 2014 would have correctly answered the question with only four response choices similarly to the other three years. Using this assumption, we determine that approximately 150 of the correct answers from years 2005 to 2007 were from guessing, close to the value that is calculated based on the question average from Table 4. That is, a four-response question with a 50% average should have about 1/3 of the correctly answered questions due to guessing. However, a more careful analysis of this question indicates it is not the ideal type for which Table 4 would accurately predict the true average. That is, response choice (A) for three of the four years was easily eliminated as a plausible correct answer for anyone with a cursory knowledge of buffers and response choice (D) for 2014 is a better lure than the response choices in the other years.

Box 2 shows a question given on the spring six-week exam; the response-choice results are reported in Table 9.

Box 2. Question 2

The vapor pressure of pure water at 25 °C is 23.8 Torr. What is the vapor pressure of water above a solution prepared by dissolving 18.0 g of glucose (a nonelectrolyte, MM = 180.0 g/mol) in 95.0 g of water (MM = 18.02 g/mol)?

A. 0.443 Torr
B. 0.451 Torr
C. 3.80 Torr
D. 4.46 Torr
E. 8.92 Torr
F. 20.0 Torr
G. 23.4 Torr
H. 23.8 Torr
I. 25.2 Torr
J. 27.6 Torr

The number of students who answered correctly in 2014 is only slightly less than 2008 despite 405 students leaving the question blank. Response (A) is a distractor answer in which the student calculated the vapor pressure lowering rather than the actual vapor pressure. Response (B) is the vapor pressure lowering with the error of calculating incorrectly the mole fraction of the solute by neglecting the number of moles of solute in the denominator. Response (F) is a distractor answer in which the mass percent of the solvent is used rather than the mole

DOI: 10.1021/ed500465q

Table 8. Distribution of Responses to the Question in Box 1

	Response Choices for Student Selection											
Year	A	B	C	D	E	F	G	H	I	J	Blank	Students, N
2005	36				496	248		241			2	1023
2006	48				488	207		255			3	1001
2007	168[a]				469	142		230			8	1017
2014	5	8	17	128	317	29	70	36	17	16	334	977

[a] Response choice 0.10 changed to 4.52.

Table 9. Distribution of Responses to the Question in Box 2

	Response Choices for Student Selection											
Year	A	B	C	D	E	F	G	H	I	J	Blank	Students, N
2008	239	195				176	349				3	962
2014	99	57	8	9	9	30	302	24	35	7	405	989

fraction. Response (H) is a distractor answer for those students who think adding a solute has no effect on the solvent vapor pressure. Response (I) is not intentionally meant to be a distractor answer, although for some unknown reason this response was chosen by more students than the other random responses. We can estimate the amount of random guessing in 2014 by looking at the low-selected random responses (C, D, E and J). Those answers total 33; thus, averaging gives about 8 random student responses per letter. Therefore, in 2014 approximately 80 (= 8 × 10) students randomly guessed on this question. Significantly fewer students in 2014 chose distractor answers (A), (B) and (F) than in 2008. If approximately the same number of students in 2008 made similar distractor errors as in 2014, then approximately 600 students in 2008 randomly guessed. Thus, a significant number of the correct answers (~150, 43%) in 2008 were the result of guessing, considerably more than those that were able to randomly guess the right answer in 2014 (~8, 3%). An alternate way of looking at this question would be to assume the same number of students in 2008 guessed who would have guessed in 2014 (80 + 405 = 485) if the 2014 students had not been encouraged to leave answers blank; i.e., the same number of students each year had no idea how to solve the problem or lacked the confidence to make a choice. Dividing these guesses evenly among the four answer choices yields approximately 120 correct answers being attributed to guessing in 2008 (~120, 34%).

Box 3 shows another question given on the spring six-week exam; the response-choice results are reported in Table 10.

Box 3. Question 3

Consider the equilibrium: $2NF_3(g) \rightleftarrows N_2(g) + 3F_2(g)$ Exactly 4.0 mol of $NF_3(g)$ and 6.0 mol of $F_2(g)$ were added to a 2.0 L vessel and the system was allowed to reach equilibrium. If $[N_2]$ = 0.75 M at equilibrium, calculate K_c for this reaction.

A. 0.044	F. 2.2×10^2
B. 7.9	G. 2.8×10^2
C. 52	H. 3.2×10^2
D. 1.4×10^2	I. 3.6×10^2
E. 1.9×10^2	J. 4.3×10^2

The students in 2014 had a lower average on this question than the students from 2004. This question was the last problem (#25) on the 2014 exam for half of the students and the 24th for the other half; i.e., we use two forms of the exam to minimize cheating. Thus, it is likely that some of the blank answers were due to students running out of time to work the problem. The only distractor response is (B) which is calculated by not taking into account the exponents in the equilibrium expression. On the basis of the incorrect answers in 2004, the number of students guessing was about 600 so that about 150 of the 542 correct answers (28%) were due to correct guesses. In 2014 about 20 of the 291 correct answers (7%) were likely due to correct guessing.

Two new (not from our database) 10-response questions that were given on the first-semester final exam are shown in Boxes 4 and 5, with results given in Tables 11 and 12.

Box 4. Question 4

When 2.74 g of Ba(s) reacts with excess $O_2(g)$ at 25 °C and 1 atm to form BaO(s), 11,100 J of heat is released. What is ΔH_f° for BaO(s) in kJ/mol?

A. −165 kJ/mol	F. −189 kJ/mol
B. −271 kJ/mol	G. −278 kJ/mol
C. −337 kJ/mol	H. −392 kJ/mol
D, −456 kJ/mol	I. −503 kJ/mol
E. −524 kJ/mol	J. −556 kJ/mol

Box 5. Question 5

A 10.0 g sample of an oxide of titanium, when heated in the presence of excess hydrogen, forms metallic titanium and 3.8 g of water. What is the mass percent of titanium in this oxide?

$$Ti_xO_y(s) + yH_2(g) \rightarrow xTi(s) + yH_2O(g)$$

A. 33%	F. 42%
B. 46%	G. 66%
C. 70%	H. 76%
D. 81%	I. 84%
E. 87%	J. 91%

In the question in Box 4 there was very little random guessing as only three response choices had more than five choices. Response (B) was a distractor answer where the student multiplied the energy release by the moles of barium instead of dividing. Response (G) was off by a factor of 2 due probably to the balanced reaction having a two in front of the BaO when the coefficients are whole numbers. If this question were asked as a four-response question with no incentive for leaving the answer blank, then the number correct based on probability would have been about 540. So the

DOI: 10.1021/ed500465q

Table 10. Distribution of Responses to the Question in Box 3

	Response Choices for Student Selection										
Year	A	B	C	D	E	F	G	H	I	J	Blank	Students, N
2004		132		172			149			542	7	1002
2014	10	64	26	34	17	34	23	22	16	291	452	989

Table 11. Distribution of Responses to the Question in Box 4

	Response Choices for Student Selection											
Year	A	B	C	D	E	F	G	H	I	J	Blank	Students, N
2014	5	97	5	4	5	2	36	3	3	433	421	1014

Table 12. Distribution of Responses to the Question in Box 5

	Response Choices for Student Selection											
Year	A	B	C	D	E	F	G	H	I	J	Blank	Students, N
2014	27	7	12	19	16	7	307	70	11	19	519	1014

421 students who left the question blank all got the same 1/4 credit rather than approximately 105 getting full credit and approximately 316 getting no credit as would have been the case if they had been encouraged to guess with a four-response question. Giving equal credit for equal knowledge is more equitable than randomly distributing full credit to 1/4 of the students with the concomitant result of a more reliable exam.

In the question in Box 5 there was no intention to have distractor answers, although response (A) results if the student answers for the mole percent of TiO_2 instead of the mass percent and response (H) is close to the percent mass for TiO. In this question, there appears to be slightly more random guessing than the previous problem. The students found this problem more difficult than the previous problem with only 307 correct responses and 519 blanks. From the random answer choices, the guesses per answer are about 15. Therefore, only about 15 of the 307 correct answers (5%) were guessed correctly. Again, if the blanks had been randomly guessed for a four-response question, then 130 more students would have answered the question correctly so that almost one out of every three correct answers would be due to random guessing. Thus, 10-response questions with an incentive not to guess are a very good way to reduce the number of students who correctly guess the answer to difficult numerical questions.

Finally, Box 6 shows a five-response question given on the first-semester final exam for comparison with the 10-response questions. The results for this question are shown in Table 13.

Box 6. Question 6

Element E reacts with oxygen to produce EO_2. Identify element E if 16.5 g of it reacts with excess oxygen to form 26.1 g of EO_2.

A. C B. Ti C. S. D. Sn E. Mn

This question had the most answers left blank of any five-response question given during the year. The question is a calculation question disguised as an inspection question; i.e., the question is answered by calculating the molar mass of the element from the given data and then selecting the element from the periodic table based on the molar mass. There was no intention of using any distractor answers other than choosing elements that form commonly known dioxides; thus, responses (A) through (D) represent students randomly guessing. Therefore, about half of the students did not know how to calculate the correct answer (251 + 5 × 50 = 501). About 50 of the 554 correct answers (9%) were due to correct guesses. Half of the students who did not know the answer chose to guess while the other half left the answer blank. This percentage of guess-to-blank ratio is much higher than seen for the ten-response questions.

Table 13. Distribution of Responses to the Question in Box 6

	Response Choices for Student Selection						
Year	A	B	C	D	E	Blank	Students, N
2014	78	42	56	33	554	251	1014

Statistical Analysis of Blanks

This past year our department gave five exams in which a variable number of response choices were used and blank answers were given 1/4 credit. The sixth exam given, the final exam for the second-semester course, was a combination of a two-semester ACS exam with additional department-written questions so that the rules for ACS exams were used. A total of 180 different questions were asked on the five exams. Of the 180 questions, 102 were questions with four response choices, 61 were questions with five response choices and 17 were questions with 10 response choices. For the four-response questions, less than 5% involved calculations. For the five-response questions, 67% involved calculations. For analysis purposes, the four- and five-response questions were divided into two subgroups based on the z-score of the question average for that question type. For the four-response questions with $z > -1$, 3% of the questions were left blank, while for the questions with $z < -1$, 5% were left blank. For the five-response questions with $z > -1$, 5% of the questions were left blank, while for the questions with $z < -1$, 9% were left blank. For the 17 “difficult” computational questions with 10 response choices, the average percentage of questions left blank was 33%. In the five exams there were 181,075 total response opportunities (~1000 students answering 180 questions). Of those, 128,481 (71.0%) were answered correctly, 39,963 (22.1%) were answered incorrectly, and 12,631 (7.0%) were left blank. The averages of the five exams (in the order given) were 82.3%, 71.1%, 73.2%, 70.2%, and 65.4% and the cumulative average was 72.7%. The percentage of students who left zero answers blank on these exams (in the order given) was 41%, 30%, 19%, 27%, and 32%, respectively. The percentage of students who left 20% or more of the questions blank was 10.3%, 10.3%, 7.4%, 14.2%, and 6.7%.

Cronbach's Alpha

Cronbach's alpha[19] is used to measure exam reliability. Cronbach's alpha was calculated for the five exams with the new format and grading rules. In the order given, the Cronbach's alphas are 0.77, 0.70, 0.88, 0.72, and 0.68. For comparison purposes, the Cronbach's alphas for the exams given the previous year were 0.65, 0.61, 0.85, 0.70, and 0.71. Thus, four of the five alpha values are higher (more reliability) for the new format although three of the five differences are very small. The exam with the smaller alpha compared to the previous year curiously had fewer answers left blank (30% fewer than the previous exam) than the other four exams despite the exam having the lowest average of the five. Perhaps the students who realize they are struggling with an exam are more likely to take a chance and guess answers compared to when they have more confidence as they progress through the exam.

Student Surveys

At the end of the year, students were asked to fill out a survey with questions about the new format and method of scoring exams. When asked for the primary reason for leaving questions blank (Box 7), the greatest percentage of students chose that they were uncertain of the correct answer (56.4%). The second highest primary reason for leaving a question blank was due to running out of time on the exam (31.0%). By a 4:3 ratio, the students preferred the new method of scoring the exam compared to previous years.

Box 7. Survey Item about Leaving Questions Blank

This past semester the chemistry department continued with a new way common exams were graded. In previous years, exams were scored by giving 4 points for every correct answer. Strategically, the test takers were encouraged to answer all 25 questions in order to maximize their scores. This year, the common exams were graded by giving 4 points for each correct answer and 1 point for every question left blank. The new method was utilized in an attempt to prevent random guessing with the goal of providing a more accurate measure of the test taker's knowledge. Which grading method did you prefer?

(1) Old scoring method (4 points per correct answer)

(2) New scoring method (4 points per correct answer and 1 point for questions left blank)

■ SUMMARY

Our student population is reluctant to leave answers blank. Only 7.0% of exam questions were left blank, similar to another study for a nonchemistry course.[20] Thus, students were three times as likely to answer a question incorrectly as to leave the answer blank. Giving 1/4 credit for a blank answer is indisputably appropriate for questions where students run out of time because an answer to such a question has to be at random. It is certainly more equitable to give all students in this situation the same credit rather than randomly giving full credit to the students who were lucky enough to have guessed correctly. The appropriateness of giving 1/4 credit to the students who are not sure of the correct answer is less clear because students with partial knowledge might be penalized for leaving a blank rather than making an educated guess. This effect has been widely debated, but exam reliability has been shown to improve when using formula scoring.[9] Ten-answer questions with an incentive to not guess definitively reduce the amount of random guessing. Difficult numerical problems in which there are only 4 or 5 answer choices and formula scoring is not used result in many of the correct answers being due to guessing.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: campbell@usna.edu.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Burton, R. F. Quantifying the Effects of Chance in Multiple Choice and True/False Tests: Question Selection and Guessing of Answers. *Assess. Eval. High. Educ.* **2001**, *26*, 41−50.

(2) Burton, R. F. Sampling Knowledge and Understanding: How Long Should a Test Be? *Assess. Eval. High. Educ.* **2006**, *31*, 569−582.

(3) Bodner, G. M. Statistical Analysis of Multiple-Choice Exams. *J. Chem. Educ.* **1980**, *57*, 188−190.

(4) Budescu, D. V.; Nevo, B. Optimal Number of Options: An Investigation of the Assumption of Proportionality. *J. Educ. Meas.* **1985**, *22*, 183−196.

(5) Cassels, J. R. T.; Johnstone, A. H. The Effect of Language on Student Performance on Multiple Choice Tests in Chemistry. *J. Chem. Educ.* **1984**, *61*, 613−615.

(6) Stanley, J. C.; Hopkins, K. D. *Educational and Psychological Measurement and Evaluation*; Prentice Hall: Englewood Cliffs, NJ, 1972.

(7) Tellinghuisen, J.; Sulikowski, M. M. Does the Answer Order Matter on Multiple-Choice Exams? *J. Chem. Educ.* **2008**, *85*, 572−575.

(8) Lord, F. M. Formula Scoring and Number-Right Scoring. *J. Educ. Meas.* **1975**, *12*, 7−11.

(9) Muijtjens, A. M. M.; van Mameren, H.; Hoogenboom, R. F. I.; Evers, F. L. H.; van der Vleuten, C. P. M. The Effect of a 'Don't Know' Option on Test Scores: Number-Right and Formula Scoring Compared. *Med. Educ.* **1999**, *33*, 267−275.

(10) Grunert, M. L.; Raker, J. R.; Murphy, K. L.; Holme, T. A. Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Partial Credit. *J. Chem. Educ.* **2013**, *90*, 1310−1315.

(11) Towns, M. H. Guide to Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *J. Chem. Educ.* **2014**, *91* (9), 1426−1431.

(12) Mattson, D. The Effects of Guessing on the Standard Error of Measurement and the Reliability of Test Scores. *Educ. Psychol. Meas.* **1965**, *25*, 727−731.

(13) Hartman, J. R.; Lin, S. Analysis of Student Performance on Multiple-Choice Questions in General Chemistry. *J. Chem. Educ.* **2011**, *88*, 1223−1230.

(14) Traub, R. B.; Hambleton, R. K.; Singh, B. Effects of Promised Reward and Threatened Penalty on Performance of a Multiple-Choice Vocabulary Test. *Educ. Psychol. Meas.* **1969**, *29*, 847−861.

(15) Lord, F. M. Formula Scoring and Validity. *Educ. Psychol. Meas.* **1963**, *23*, 663−672.

(16) Burton, R. F. Multiple Choice and True/False Tests: Reliability Measures and Some Implications of Negative Marking. *Assess. Eval. High. Educ.* **2004**, *29*, 585−595.

(17) Burton, R. F.; Miller, D. J. Statistical Modelling of Multiple-Choice and True/False Tests: Way of Considering, and of Reducing, the Uncertainties Attributable to Guessing. *Assess. Eval. High. Educ.* **1999**, *24*, 399−411.

(18) Hinkley, C. C.; Lagowski, J. J. A Versatile Computer-Graded Examination. *J. Chem. Educ.* **1966**, *43*, 575−578.

(19) Cronbach, L. J. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* **1951**, *16*, 297−334.

(20) Ebel, R. L. Blind Guessing on Objective Achievement Tests. *J. Educ. Meas.* **1968**, *5*, 321−324.

DOI: 10.1021/ed500465q