Creating the Perfect Multiple-Choice Question

Testing to see whether randomized answer length influences response selection

Hanna Haddad

Abstract

In looking to construct the most ideal multiple-choice question, this paper looks to address one of the smaller regions within the creation of the multiple-choice question. It is colloquially thought that the longer an answer response, the more obvious and thus the more likely the answer is to be chosen. This survey aimed to address that question, in creating a survey that varied answer choice sizes between small, medium, and large. Unfortunately, the survey found no conclusions, largely a result of small scale and internal validity problems surrounding the construction of the concepts of "small, medium, and large".

Keywords: multiple-choice, answer length, randomized

Note: The following paper was assembled in accordance with the CONSORT checklist

Introduction

Background & Objectives

This study intends to contribute to understandings of how to construct the most perfect multiple-choice question. As a subset of this large category, this study intends to look at the influence that varying answer length has on the answer selection. Overall, it is widely acknowledged that generally longer or shorter correct answer length tends to influence answer response. This phenomenon is enhanced by correct answers being consistently longer or shorter within a battery. The present study is a further test of these forms of bias.

The primary hypotheses being tested are the following:

- *Hypothesis 1*: Under the "longer" answer length experimental condition, modified-answer response rates will be higher than they would be under the "medium" answer length experimental condition. (compare longer with medium)

- *Hypothesis 2*: Under the "shorter" answer length experimental condition, modified-answer response rates will be higher than they would be under the "medium" answer length experimental condition. (compare longer with medium)

Please refer to Appendix A for other potential hypotheses.

Methods

Trial Design

First, subjects will be asked six background questions. The first three of these six questions are demographic questions that ask the respondents sex, age, and level education. The other three background questions are potential modifying variables that ask: to what degree the respondent likes history, whether the individual has taken the Advanced Placement U.S History (APUSH) test, and an attention check.

After that, subjects will be asked four questions, out of a pool of eight questions, each relating to APUSH. Each question will be experimentally varied in one of three conditions (short, medium, long) along the dimension of answer length. The order of the questions and the subsequent answer selections for these questions will be randomized.

It is important to mention that this survey is not a factorial. We experimentally vary the number of answer options available, between four and five answer options (4O vs. 5O) and the type of answer choice available, between the correct or incorrect dimension (C vs. IC). As there are eight questions in the pool, the eight questions are divided among the two different dimensions: two of them receive the 4O x C modification, two of them receive the 4O x IC modification, two of them receive the 5O x C modification, and two of them receive the 5O x IC modification.

Unfortunately, there were several unintended methodological changes that occurred upon execution of the survey:

1. Out of fear that the survey wouldn't get enough 1000 responses at $0.06, we decided to make it 500 responses at $0.12.
2. It was discovered that one of the questions had been modified incorrectly. Instead of modifying to make it 4O x C, we modified the question to be 4O x IC. This made potential stratification along C vs. IC unfruitful.
3. Since there were only eight questions in total, any analysis along 4O vs. 5O and C vs. IC would likely suffer from a limited sample of the independent variable. Thus, both potential hypotheses were thrown out and not used for any data analysis.

Participants

All participants are self-selected members of Amazon Mechanical Turk who have opened, accepted the HIT, and given consent to take the survey processed through Qualtrics. All data collection occurred online between the dates of 05/19/15 and 05/25/15.

Interventions

All individuals had equal probability of receiving any given question. Thus, officially, there was no control and treatment group. However, the interventions that did occur were the answer length

modifications. The interventions that did occur was the variance along answer options between short, medium, and long dimensions.

There was no official metric to establish what was considered short, medium, or long. Overall, the general barometer used was that:

1. The short answer option was shorter than the average of other options
2. The medium answer option was about the same length as the average of other options
3. The long answer option was longer than the average of other options

Sample Size

Respondents (n = 1000) will be drawn from the Amazon Mechanical Turk online labor force specifically those who formally accept the HIT, and give their consent to take the survey. Each respondent will be paid $0.06 upon completion of the survey. Having a large number of respondents would allow for a very fruitful comparison between the answer options of different questions in what was, inherently, a very small and cognitive-focused survey. In the interest of practicality, we reduced the cut to half that and doubled the amount the compensation. The final sample size was n = 471, where each respondent was paid $0.12.

Randomization

In order to assign the 5O vs. 4O and C vs. IC conditions, I divided the questions into two groups: 5O and 4O. Questions 1-4 would be considered the 5O bracket. Questions 5-8 would be considered the 4O bracket. Having established that, I randomized the C vs. IC conditions. To randomize the C condition, I used a random number generator to generate eight numbers between 0 and 100. According to the order, they were assigned to certain questions; i.e. numbers 1 and 2 corresponded with questions 1 and 2.

Once the eight numbers were generated, the top two numbers within the 5O bracket received C treatment, the bottom two would be rolled to receive IC treatment. This was repeated for the other questions in the 4O bracket, such that there were 4 questions assigned C treatment and 4 questions assigned IC treatment.

For questions receiving the IC treatment, I tallied the number of options available, which were: 4, 4, 3, 3 (because correct options are omitted). I then rolled 14 numbers and assigned modification to the highest number within the respective question brackets.

This was the extent of manual randomization. All other randomization occurred on the survey itself. This randomization included:

1. Random selection of 4 questions out of the potential 24 available (3 conditions x 8 questions)
2. Random ordering of answer choices

Results

Participant Flow & Recruitment

The survey was released Tuesday, May 19th and was kept open until Monday, May 25th. It was closed on Monday for analysis.

Initially, there was a brief, but neglible error in the releasing of the survey through Amazon Mechanical Turk. The survey had been released, but there was a minor error with the link. The survey was stopped after collecting 6 responses and the error was fixed in a relaunch. The official survey resulted in the collection of 462 responses. Two responses (not included in the 462) were rejected. Overall, Amazon Mechanical Turk registered a total of 468 responses; each respondent paid $0.12 for their participation.

Qualtrics, however, collected a total of 471 responses. To see if we could find the reason for the difference between Amazon Mechanical Turk and Qualtrics, we checked the user IDs, but they were all different. What, then, is most likely is that some of the participants in the survey had sent the survey to other individuals to take. Either way, the issues faced in the collection of survey respondents were largely minimized, such that these problems, overall, posed very little substantial issue on the survey.

The following chart includes the sample sizes for each of the 24 questions. The 471 respondents received 4 non-duplicated conditions from a potential 24.

Q#	*n*	*Q#*	*n*	*Q#*	*n*	*Q#*	*n*
Q1S	67	Q3S	83	Q5S	70	Q7S	95
Q1M	81	Q3M	75	Q5M	77	Q7M	70
Q1L	83	Q3L	70	Q5L	98	Q7L	76
Q2S	59	Q4S	80	Q6S	71	Q8S	82
Q2M	83	Q4M	76	Q6M	70	Q8M	83
Q2L	83	Q4L	66	Q6L	97	Q8L	84

The average sample size was 80 respondents.

Baseline Data

Overall, the demographics paint a rather stable picture. Thankfully, the survey didn't have a huge female bias, as is often seen in other surveys run through Amazon Mechanical Turk. Additionally, there is a clear bias toward younger respondents. The median age in the survey was approximately 33 years old, which is, without a doubt, not as representative. The median age in surveys administered through other modes (rather than online) tend to hover around 40-45. And, as expected, the education levels among respondents are noticeably lower in comparison the American national averages. Though, that makes more sense since the cohort in this study is younger and has a slight bias against education (due to the nature of the works who use MTurk).

Age Group	*1*	*%*
20-29	159	34%
30-39	150	32%
40-49	70	15%
50-59	55	12%
60+	35	7%
N/A	2	0%
	471	**100%**

Sex	*n*	*%*
Male	227	48%
Female	244	52%
	471	**100%**

Education Level	*n*	*%*
Less than High School	3	1%
High School / GED	60	13%
Some College	136	29%
2-year College Degree	61	13%
4-year College Degree	166	35%
Masters Degree	40	8%
Doctoral Degree	1	0%
Professional Degree (JD, MD)	4	1%
	471	**100%**

Outcomes and Estimation

Overall, the study had rather inconclusive findings, which quickly becomes apparent in a brief data analysis. I also do not conduct establish confidence intervals simply because the results are so clearly askew. Please refer to the adjacent table and the graph below to see a visual mapping of the data.

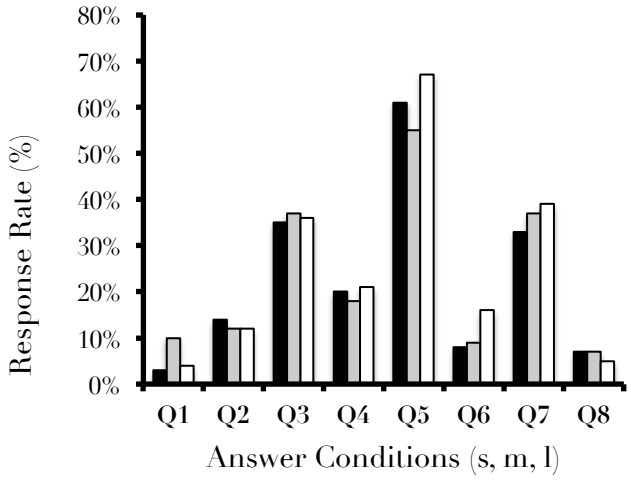
The results for Q2, Q3, Q4, and Q8 all change was more or less marginal and, due to the small sample sizes, are and should be considered, virtually insignificant changes.

The results from Q6 and Q7 show a slight trend where the longer option receives more support.

The results from Q1 shows exactly the opposite of hypothesis to be true.

The results from Q5 are the only that reinforce the existing hypotheses.

		O1	*O2*	*O3*	*O4*	*O5*	*Sample Size*
Q1	S	70%	**3%**	9%	13%	4%	67
	M	78%	**10%**	5%	2%	5%	81
	L	71%	**4%**	10%	7%	8%	83
Q2	S	39%	5%	**14%**	19%	24%	59
	M	37%	5%	**12%**	*18%*	28%	83
	L	36%	*13%*	**12%**	16%	23%	83
Q3	S	29%	6%	17%	**35%**	13%	83
	M	25%	*5%*	21%	**37%**	11%	75
	L	11%	7%	27%	**36%**	19%	70
Q4	S	40%	**20%**	18%	*13%*	10%	80
	M	42%	**18%**	17%	16%	7%	76
	L	29%	**21%**	17%	17%	17%	66
Q5	S	13%	14%	11%	**61%**	N/A	70
	M	18%	16%	12%	**55%**	N/A	77
	L	11%	13%	8%	**67%**	N/A	98
Q6	S	35%	**8%**	39%	17%	N/A	71
	M	23%	**9%**	44%	24%	N/A	70
	L	22%	**16%**	38%	24%	N/A	97
Q7	S	26%	8%	33%	**33%**	N/A	95
	M	23%	11%	29%	**37%**	N/A	70
	L	20%	17%	24%	**39%**	N/A	76
Q8	S	18%	48%	27%	**7%**	N/A	82
	M	19%	47%	27%	**7%**	N/A	83
	L	15%	46%	33%	**5%**	N/A	84



Discussion

Limitations

I believe that a large part of the reason this survey ended up with such flawed and convoluted results, was for several reasons:

1. **The quantification of small, medium and large.** There was no set manner in which these were defined. In some instances, the "small" answer condition was the approximately the same size. There was not really any unifying standard or steadfast mathematical definition.

2. **Changing of meaning.** As I found out the hard way, questions with short answers are already immensely hard to simplify. In some instances, it was very difficult (if not impossible) to reduce the size of the answer options. The adaptation, then, became rather unnatural and served to, oftentimes, convolute or distort the original meaning articulated in the original answer response

3. **Small sample sizes.** In four of the eight questions, I was unable to get any significant results, simply because the sample sizes for each question and for each condition were too small. 80 respondents isn't nearly enough to make any reasonably accurate or respected scientific claims. To do that with a decent level of precision, I would've have needed to collect sample sizes of at least 200-300 on each question, which would've required nearly double the respondents.

4. **Weak Data Analysis.** For the most part, the results and data analysis are rather simple. If I had the time and the willpower, I could've analyzed the data through the following stratifying variables: attn_check, taken_exam & char_length. Appendix C shows other work done to potentially look at character length among answers in order to establish a more scientific backing to the definitions of small, medium, and large.

5. **Loss of Moderating Variables.** Because of the problems faced in having such a limited sample of dependent variables, I had to lose two important moderating variables (4O vs. 5O & C vs. IC) that I had reserved for complex data analysis. That could've, potentially, contributed to some meaningful conclusions regarding my hypothesis.

6. **Small Sample of Dependent Variables.** There were only eight questions, which is in no way a substantial or significant to make large claims about the hypotheses I had created. I needed, at least, double the questions on rotation.

Generalizability

Thankfully, this is one of few areas where my survey was not entirely awful. Overall, the survey population was relatively representative of the MTurk population and, perhaps, the larger American population. It was markedly younger and less educated, but that could be considered a problem inherent in the nature of the platform used.

Aside from that, as there are no conclusive findings, there is nothing worth generalizing to the larger environment. Even if this survey was successful, it belongs within one of many experiments to help solve the riddle of what makes the perfect multiple-choice question.

Interpretation

Now that I look back at this all, I find it laughable that I had thought I would get substantial results from running this survey. Overall, the survey was too small and too messily conducted (see internal validity issues mentioned in limitations section) to derive any noteworthy conclusions. That being said, I do not believe that this hypothesis has been disproven; in fact, I believe more research must be done!

Appendix A: Optional Hypotheses

Hypothesis 3: There will be no effect between question where there were either 4 or 5 answer options. (compare five_option with four_option)

Hypothesis 4: There will be no effect between questions where either a correct or incorrect answer was modified. (compare correct_answer with incorrect_answer)

Hypothesis 5: There will be a bias against questions where the average was closer to the shorter experimental condition. (compare closer_equal with closer_shorter)

Hypothesis 6: There will be no observable effect in the analysis between individuals who have and have not taken the AP U.S. History test. (stratify any data with taken_exam)

Hypothesis 7: There will be no observable effect in the analysis between individuals who did and did not pass the attention check. (stratify any data with attn_check)

Appendix B: Variable Construction

Direct Variables
These variables do not require any reconstruction; the variables gathered are simply a collection of direct responses from the survey.

***add no-response coding; yes is always coded higher than no; correlations = corr x y**

Name: sex
Type: moderating
Description: this variable measures ones given sex.
Coding: (1) Male (2) Female

Name: age
Type: moderating
Description: this variable measures ones age
Coding: 2015 - [year_born]

Name: taken_exam
Type: moderating
Description: this variable will measure whether or not an individual has taken the AP U.S. History exam before. It could be potentially used as a moderating variable.
Coding: (1) Yes (2) No (9) Don't know/remember (.) Missing data

Name: enjoy_hist
Type: moderating
Description: this variable measures the degree to which an individual enjoys the subject of history.
Coding: (1) Enjoy (2) Weakly enjoy (3) Indifferent (4) Weakly dislike (5) Dislike

Name: attn_check
Type: moderating
Description: this variable measures whether one is reading the question descriptions.
Coding: (1) Correct (2) Incorrect

Name: Q[1-8][S, E, L]
Type: independent
Description: these are responses to a singular question that will be recombined for other analyses.
Coding: (1) Correct (2) Incorrect

Constructed Variables
These variables are combinations of certain variables to yield certain analyses.

Name: five_option
Type: moderating
Construction: Q[1-4][S, E, L]
Description: this is a five-option variable analysis.
Coding: (1) Correct (2) Incorrect

Name: four_option
Type: moderating
Construction: Q[5-8][S, E, L]
Description: this is a four-option variable analysis.
Coding: (1) Correct (2) Incorrect

Name: correct_answer
Type: moderating
Construction: Q[3, 5, 7][S, E, L]
Description: this variable measures the modification of the correct answer to allow analysis to reveal if modification of the correct answer held an impact.
Coding: (1) Correct (2) Incorrect

Name: incorrect_answer
Type: moderating
Construction: Q[1, 2, 4, 6, 8][S, E, L]
Description: this variable measures the modification of the incorrect answer to allow analysis to reveal if modification of the incorrect answer held an impact.
Coding: (1) Correct (2) Incorrect

Name: closer_equal
Type: moderating
Construction: Q[1, 3, 4, 6, 8][S, E, L]
Description: this variable is composed of questions where the average is closer to the equal experimental condition than the shorter or longer experimental conditions. The purpose of doing this is to reveal, through analysis, whether there is any influence of the average non-modified answer options on the experimental condition.
Coding: (1) Correct (2) Incorrect

Name: closer_shorter
Type: moderating
Construction: Q[2, 5, 7][S, E, L]
Description: this variable is composed of questions where the average is closer to the shorter experimental condition than the equal experimental condition. The purpose of doing this is to reveal, through analysis, whether there is any influence of the average non-modified answer options on the experimental condition.
Coding: (1) Correct (2) Incorrect

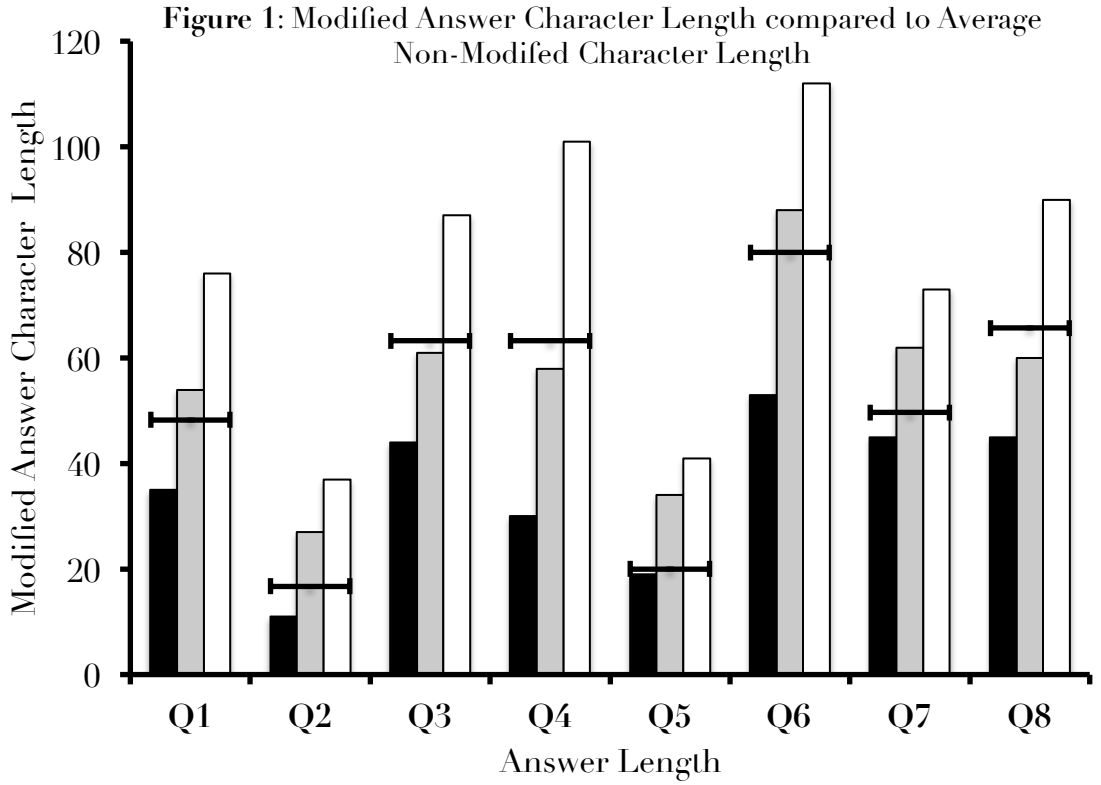
Name: shorter
Type: moderating
Construction: Q[1-8][S]
Description: this variable is a combination of all the "shorter" experimental conditions.
Coding: (1) Correct (2) Incorrect

Name: equal
Type: moderating
Construction: Q[1-8][E]
Description: this variable is a combination of all the "equal" experimental conditions.
Coding: (1) Correct (2) Incorrect

Name: longer
Type: moderating
Construction: Q[1-8][L]
Description: this variable is a combination of all the "longer" experimental conditions.
Coding: (1) Correct (2) Incorrect

Name: char_length
Type: moderating
Construction: count character count in each answer
Description: this variable measures the character length of each answer.
Coding: N/A

Appendix C: Graphs and Tables



	Modified Options				*Non-Modified Options*				
Questions	*Shorter*	*Equal*	*Longer*	*Average*	*NMO-1*	*NMO-2*	*NMO-3*	*NMO-4*	*NMO-5*
Q1	35	54	76	48.25	45	***M***	57	51	40
Q2	11	27	37	16.75	10	13	***M***	26	18
Q3	44	61	87	63.25	52	47	83	***M***	71
Q4	30	58	101	63.25	66	***M***	40	65	82
Q5	19	34	41	20.00	18	23	19	***M***	N/A
Q6	53	88	112	80.00	71	***M***	87	82	N/A
Q7	45	62	73	49.67	43	51	55	***M***	N/A
Q8	45	60	90	65.67	76	60	61	***M***	N/A

A table illustrating what is shown in the graph above.

Appendix D: Question Details

QUESTION 1	*The development of the early nineteenth-century concept of "separate spheres" for the sexes encouraged all of the following EXCEPT*
Question 1: Shorter	accepting women as intellectual equals of men
	idealizing the home as a safehaven
	designating the home as the appropriate place for a woman
	emphasizing childrearing as a prime duty of a woman
	establishing a moral climate in the home
Question 1: Medium	accepting women as intellectual equals of men
	idealizing the home as a haven in a competitive world
	designating the home as the appropriate place for a woman
	emphasizing childrearing as a prime duty of a woman
	establishing a moral climate in the home
Question 1: Longer	accepting women as intellectual equals of men
	idealizing the home as a haven free from the burdens of a competitive world
	designating the home as the appropriate place for a woman
	emphasizing childrearing as a prime duty of a woman
	establishing a moral climate in the home
QUESTION 2	*Under the Articles of Confederation the United States central government had no power to*
Question 2: Shorter	levy taxes
	make treaties
	declare war
	request troops from states
	amend the Articles
Question 2: Medium	levy taxes
	make treaties
	declare war against enemies
	request troops from states
	amend the Articles
Question 2: Longer	levy taxes
	make treaties
	declare war against perceived threats
	request troops from states
	amend the Articles
QUESTION 3	*Which of the following has been viewed by some historians as an indication of strong anti-Catholic sentiment in the presidential election of 1928?*
Question 3: Shorter	The increased political activity of the Ku Klux Klan
	The failure of the farm bloc to go to the polls
	Alfred E. Smith's choice of Arkansas senator Joseph T. Robinson as his running mate
	Alfred E. Smith's failure to carry the South
	Herbert Hoover's use of "rugged individualism" as his campaign slogan 3
Question 3: Medium	The increased political activity of the Ku Klux Klan
	The failure of the farm bloc to go to the polls
	Alfred E. Smith's choice of Arkansas senator Joseph T. Robinson as his running mate
	Alfred E. Smith's failure to carry a solidly Democratic South
	Herbert Hoover's use of "rugged individualism" as his campaign slogan 3
Question 3: Longer	The increased political activity of the Ku Klux Klan
	The failure of the farm bloc to go to the polls
	Alfred E. Smith's choice of Arkansas senator Joseph T. Robinson as his running mate
	Alfred E. Smith's failure to receive mass support from a predominantly Democratic South
	Herbert Hoover's use of "rugged individualism" as his campaign slogan 3
QUESTION 4	*The American Federation of Labor under the leadership of Samuel Gompers organized*

Question 4: Shorter	skilled workers in craft unions in order to achieve economic gains
	all workers in "one big union"
	unskilled workers along industrial lines
	workers and intellectuals into a labor party for political action
	workers into a fraternal organization to provide unemployment and old-age benefits
Question 4: Medium	skilled workers in craft unions in order to achieve economic gains
	all industrial and agricultural workers in "one big union"
	unskilled workers along industrial lines
	workers and intellectuals into a labor party for political action
	workers into a fraternal organization to provide unemployment and old-age benefits
Question 4: Longer	skilled workers in craft unions in order to achieve economic gains
	all indviduals in the industrial and agricultural sectors into "one big union" with centralized power
	unskilled workers along industrial lines
	workers and intellectuals into a labor party for political action
	workers into a fraternal organization to provide unemployment and old-age benefits
QUESTION 5	*All of the following concerns were addressed during the "Hundred Days" of the New Deal EXCEPT*
Question 5: Shorter	banking regulation
	unemployment relief
	agricultural adjustment
	court restructuring
Question 5: Equal	banking regulation
	unemployment relief
	agricultural adjustment
	restructuring of the Supreme Court
Question 5: Longer	banking regulation
	unemployment relief
	agricultural adjustment
	restructuring of Supreme Court procedures
QUESTION 6	*Which of the following was true of the French-American Alliance formed in 1778?*
Question 6: Shorter	It contributed little to the American victory in the Revolutionary War.
	It restricted French naval activity to the high seas.
	It influenced the British to offer generous peace terms in the Treaty of Paris in 1783.
	It specifically prohibited the deployment of French troops on North American soil.
Question 6: Equal	It contributed little to the American victory in the Revolutionary War.
	It restricted French naval activity to the high seas, far from the North American coast.
	It influenced the British to offer generous peace terms in the Treaty of Paris in 1783.
	It specifically prohibited the deployment of French troops on North American soil.
Question 6: Longer	It contributed little to the American victory in the Revolutionary War.
	It restricted French naval activity to the high seas, far from the British colonies on the North American coast.
	It influenced the British to offer generous peace terms in the Treaty of Paris in 1783.
	It specifically prohibited the deployment of French troops on North American soil.
QUESTION 7	*Liberty of conscience was defended by Roger Williams on the grounds that*
Question 7: Shorter	all religions were equal in the eyes of God
	Puritan ideas about sin and salvation were outmoded
	theological truths would emerge from the clash of ideas
	states are inappropriate as spiritual agencies
Question 7: Equal	all religions were equal in the eyes of God
	Puritan ideas about sin and salvation were outmoded
	theological truths would emerge from the clash of ideas
	the state was an inappropriate agency in matters of the spirit
Question 7: Longer	all religions were equal in the eyes of God
	Puritan ideas about sin and salvation were outmoded

	theological truths would emerge from the clash of ideas
	the state was an improper and ineffectual agency in matters of the spirit
QUESTION 8	*Which of the following was true of a married woman in the colonial era?*
Question 8: Shorter	She would be sentenced to debtors' prison for debts incurred by her husband.
	She generally lost control of her property when she married.
	She had no legal claim on the estate of her deceased husband.
	She had equal legal rights over her children.
Question 8: Equal	She would be sentenced to debtors' prison for debts incurred by her husband.
	She generally lost control of her property when she married.
	She had no legal claim on the estate of her deceased husband.
	She had equal legal rights over her children as her husband.
Question 8: Longer	She would be sentenced to debtors' prison for debts incurred by her husband.
	She generally lost control of her property when she married.
	She had no legal claim on the estate of her deceased husband.
	She had legal rights over her children that were exactly the same as those of her husband.