SCIENCE PBAT

Content Effects on Word Problem Performance

ASSESSMENT: GOOD

SCIENCE LAB REPORT GUIDELINES

TITLE: This is the question that your experiment answers.

DEVELOPING A QUESTION:
You are to develop a question based on something we have done or discussed in class that you are curious about and would like to pursue in greater depth.

INTRODUCTION: Include:

1. Background Information – Explain the necessary background information that you think the reader should know in order to understand your research. (This is the information that found by reading and interviewing people). Do NOT just list information. Synthesize it and relate it to your question, experiment, and/or hypothesis.

2. Your Hypothesis (or hypotheses) – Explain the reasons for choosing your hypothesis (the predicted answer to the question that you are investigating.) Your hypothesis MUST be (at least partially) based on your background research, and you should explain the connection.

MATERIALS:
Just make a list of all the materials that you used.

PROCEDURE:
Explain to the reader EXACTLY how you carried out your experiment, in numbered steps. Don't forget the steps that specifically tell how you controlled your variables and how you collected your data.

RESULTS:
Represent your data using appropriate charts and graphs. Briefly explain what the data mean (what the numbers are telling you).

DISCUSSION: In this section you analyze the data. Include:

1. A. How well did your results (data) support your original hypothesis? Fully explain using evidence from your results to support your answer.

   B. Which data were unexpected and did NOT support your hypothesis? Why did you get these results? You need to explain what might have CAUSED this data to be different from what you had expected or predicted.

2. What patterns do you notice in your data besides the patterns related to your hypothesis? Make a list of at least five unanticipated patterns in the data. What hypotheses do these patterns suggest? What questions do they raise? Fully explain by referring to specific numbers from your data.

3. What problems did you have performing the experiment and how did you solve each of them?

4. How would you improve your procedure if you had more time? Explain the reasons for EACH of the ways you would change the procedure.

5. What questions arose from performing this experiment? Choose ONE of these questions. Write a detailed procedure explaining how you would perform an experiment that would investigate the possible answers to this question.

REFERENCES CITED. Use APA format, and alphabetize the list by author's last name. We highly recommend that you use a bibliography generator such as the one you can find at BibMe.org

Content Effects on Word Problem Performance

Introduction:

Gender bias on tests is an issue that affects the function and reliability of public education to assess students' ability to perform. Tests can best evaluate student ability when created to be taken in impartial and objective circumstances. Subjective judgement is not a basis of general public education- and tests are widely relied on to properly assess students. If we can recognize gender bias, we can tweak our approach to designing tests to not cater to such- but go against the bias and create equal and fair opportunity in all academic settings for students to perform.

Before addressing the effects of bias, the term itself must be defined. Bias is defined by the Merriam Webster dictionary as a prejudice meaning a strong inclination of the mind or a preconceived opinion about someone or something. In this paper however, I will be dealing with content bias. This arises when a test is made to measure something, but in practice contains questions that can put one group at advantage over another. (Jencks & Phillips, 2011). This does not apply to any type of testing, as a subject who has only taken basic algebra would be unfavored on a trigonometry test, in comparison to someone who has taken trigonometry. But this type of bias arises on tests that in principle could be written in a way that does not favor certain groups.

For this experiment, I chose math problems over reading comprehension. Every word problem written could be converted into a equation- but I wanted to see how the content would affect subjects ability to do so. Because word problems can have so much variability in content, they can be constructed with content that is more familiar and approachable to one group than another. For example, a woman may see a math problem

dealing with the subject of football, and may feel unequipped to approach the problem. This is wholly based on stereotypes that exist in our present pop culture and are pressed onto children as they grow up. Of course, there are women who are familiar with masculine things and vise versa.

Content is defined as the topics or matter treated in a written work. Changing the content of the word problems within the realm of what is stereotypically male and female can show how the participants' gender could affect their performance on these word problems.

Other experiments testing content based effects on math word problems have been conducted. Chipman and Marshall (1991) conducted an experiment which manipulated word problem content. The word problems were organized into four categories: masculine, feminine, neutral familiar, and neutral unfamiliar. The tests were randomized sets of four of each classification, resulting in 32 different tests. Twenty-five male and 25 female subjects were tested in an allotted time of thirty minutes. The results suggested that there were fewer effects of feminine vs masculine content, as opposed to familiar vs unfamiliar content- meaning the masculine and feminine content had less of an effect on performance than familiar vs unfamiliar content.

A former Consortium high school student also conducted a study on how the wording of math questions could affect performance on math tests (Torres, 2014). In her experiment, she created tests using systems of equations word problems. She generated nine problems, three classified as feminine, three as masculine, and three as neutral. These classifications were based on stereotypes. She wrote.. "For example, a female biased questions talks about ticket admission to a fashion show.". Following this, I used

stereotypes of men and women in my experiment.

In her paper, she notes that it would have been more ideal if the math problems were the same level of difficulty. Because of this, I choose to stay within the realm of basic algebra. Her data show the neutral classified questions were attempted the most. In conclusion, it seemed women performed better on male oriented questions, although the difference was not significant. She notes that the male and neutral oriented questions were easier overall, and this may have contributed to her results.

In my experiment, I chose six basic algebra math problems and converted them into masculine, feminine, and neutral versions. The neutral problems were converted into equations, instead of word problems. Three different versions of a test were then generated with a random order of feminine, masculine, and neutral. Each version had two problems of each classification. I also included a page asking about various demographics including gender.

Hypothesis One - Subjects will perform significantly better on the neutral versions of the questions as opposed to the gendered versions because the equations for the neutral questions are provided; meaning subjects do not have to convert word problems into correct equations.

Hypothesis Two - Subjects will do significantly worse on feminine versions of the questions, as opposed to the masculine and neutral versions, because of the data found in Torres (2014).

Hypothesis Three- Female subjects will have significantly higher scores than male subjects, also because of what was found in Torres (2014).

Hypothesis Four - Each subjects' performance on masculine questions will not be

significantly different from their performance on feminine questions because Chipman and Marshall (1991) suggests that familiarity is more important than content bias.

Hypothesis Five - Subjects' performance on questions written for their own gender will not be significantly different from those written for the opposite gender.

Subjects:

Twenty subjects were tested. They were all students at a New York City Public High School. The students' ages ranged from 15-18. The tests were distributed in the Trigonometry and Geometry classes. There were twelve students in Trigonometry, and eight in Geometry. Ten subjects identified as male, nine subjects as female, and one subject as other.

One set of subjects' data from the Algebra Two class was not included in the experiment because they differed too much from the others. A t-test was done to find the p-value, the statistical significance between two sets of data. The p-value is the calculated probability that shows whether or not the null hypothesis should be rejected. In this case, the null hypothesis is that the differences in scores in the two math classes are due to random variation. If the p-value is less than .05, the null hypothesis is rejected. The table below shows the p-values of the three class' scores paired with one another.

Class Tested	**P-Value of Data Set**
Algebra 2 / Trigonometry	0.01
Algebra 2 / Geometry	0.01
Geometry / Trigonometry	0.84

The probability of the Geometry and Trigonometry data sets being similar to the Algebra Two data set is extremely low. This is why the data collected from the Algebra 2 class was not included in my analysis of the experimental data.

Assessment:

The math assessment was created to present three different classifications (masculine, feminine, and neutral) in basic algebra word problems. The problems were taken from a website online through searching “algebra one word problems”. The questions were then pulled the from a math learning site. It contained word problems and steps to solving them. It was then decided what was considered masculine, feminine, and neutral content. For feminine content, beauty products, the color pink, and questions dealing with women were used. For masculine content, footballs and manual labor and questions dealing with men were used. For neutral content, the question was given either in the form of a algebraic equation or had the content deal with items such as printer cartridges and did not specify any genders. For all six questions, one of each classification was generated, changing the content but keeping the base problem and the answer the same. Google Sheets was then used to create three versions of the test. Each version had two masculine, two feminine, and two neutral versions of the questions. The order was then randomized using Google Sheets. Demographic questions were then copied from the census regarding ethnicity and race. An age question was put ranging from 14-19, as well as a question asking the subject to identify their gender, and a question asking the subject to identify their sexual identity. These questions were used as a cover page.

Procedure:

1. A cover page was generated advising the teachers of how to distribute the assessment. It asked teachers to advise students to work individually, answer to the best of their ability, and contained a time limit of 20 minutes or less.
2. The tests were distributed in the Geometry, Trigonometry, and Algebra Two classes over the course of one class period.
3. The versions of the tests were placed in a recurring order of Version A, Version B, Version C. The teachers then distributed the stack to the students.
4. The teachers then returned the tests to the researcher.
5. The tests were scored using three marks. True, false, or blank. Any problem left with no answer was marked blank, incorrectly answered problems were labeled false, and correctly answered problems were labeled true.

Results

One of the female classified questions was also incorrectly worded, and was impossible to answer. This caused question five to be discarded when evaluating the results. The comparison amongst female, male, and neutral classified questions became unbalanced because of this.

This chart shows the initial results obtained. The decimals represent the percentage of subjects that got that classification of the question correct.

Table One: Initial Results

Question	1	2	3	4	5	6
Percent Male Version Correct	50%	82%	80%	83%	33.33%	75%
Percent Female Version Correct	67%	100%	33%	80%	N/A	64%
Percent Neutral Version Correct	80%	83%	7%	80%	70%	40%

My first hypothesis states: subjects will perform significantly better on the neutral versions of the questions compared to the gendered versions because the equations for the neutral questions are provided. The data table shows that subjects did not perform significantly better overall on the neutral classified questions. I took the mean of the six numbers for each classification in the data set to compare their average score. When scoring Question five for the female classified questions, I removed Question five from my calculations as it was written incorrectly and impossible to answer, shown in its correct response rate of 0%. Here are the mean correct portions for each type:

Table Two: Mean Percent of Question Types Answered Correctly

Mean Portion Male Correct	67%
Mean Portion Female Correct	69%
Mean Portion Neutral Correct	70%

The mean percent of male questions answered correctly is .67, while the mean percent of correctly answered neutral questions is .7, showing no substantial difference. This goes as well when comparing the mean percent of female questions answers correctly in comparison to the neutral questions, showing a small difference of only .1.

Although the mean percent is higher for the neutral classified questions, it is only by 3% and 1% comparatively.

My second hypothesis states: subjects will do significantly worse on feminine versions of the questions, as opposed to the masculine and neutral versions, because of the data found in Torres (2015). As shown in Table One, the data does not support this hypothesis. You must again take into account the fact that question 5 of the female classification was written incorrectly and therefore cannot be compared to the other percent of correct answers obtained. For Question Two, the feminine question obtained the highest percent of correct answers in that all answers were correct. In questions one and six, the feminine questions were answered correctly the second highest. Question four had similar results for all three types. Only in question three did the feminine classified question do the poorest, coming in at a 33% answered correctly. This contradicts my original claim that subjects will do significantly worse on feminine versions of questions, since this is only apparent in a fifth of the questions' results.

My third hypothesis states: Female subjects will have significantly higher scores than male subjects, also because of what was found in Torres (2015).

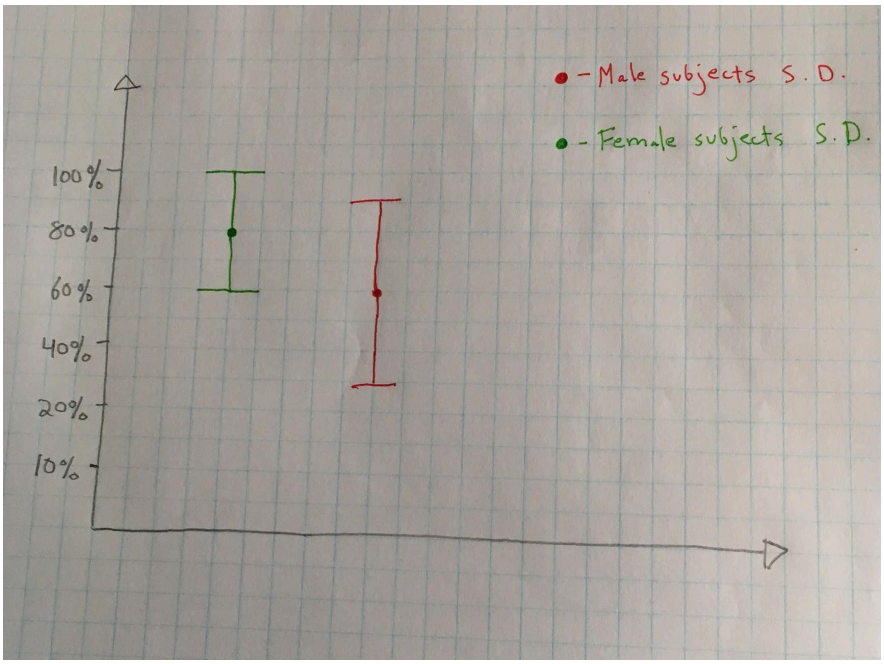
Table Three: Mean Scores by Gender and Standard Deviation

Average Percent Score for Female Subjects	80%	Standard Deviation	20%
Average Percent Score for Male Subjects	60%	Standard Deviation	30%

As you can see, female subjects did generally score higher than male subjects. This supports my hypothesis above. Although, the average percent score is not substantial different between the two pieces of data, female subjects still performed at a higher

correct rate. The female subjects' mean score has a standard deviation of .2, and the male subjects mean score has a standard deviation of .3, showing similar variability within the data sets.

Table Four: Standard Deviation Graphed



As you can see, the two data sets overlap. The male subjects' variability is similar to the female subjects' variability.

Table Five: Score Breakdown

Subject Gender	Percent Correct of Feminine Question (without Question Five)	Percent Correct of Feminine Questions (including Question Five)	Percent Correct of Male Questions	Percent Correct of Neutral Questions
Female	100%	80%	80%	80%
Male	100%	50%	60%	70%

My fourth hypothesis states: each subjects' performance on masculine questions will not be significantly different from their performance on feminine questions because

Chipman and Marshall (1991) suggests that familiarity is more important than gender. Female subjects generally performed the same on feminine and masculine questions, at a correct rate of 80% for each. Male subjects had a bit more variability, but did not perform at a significantly different correct rate between feminine and masculine questions. The difference is only of 10%, although it is worth noting that male subjects did perform better on masculine questions. When looking at the correct rates for feminine questions without Question Five, both female and male subjects performed at a 100% correct rate. This supports my hypothesis in that both genders performed the same on the feminine questions, showing that female subjects could perform the same as male subjects on a feminized question.

My fifth hypothesis states: subjects' performance on questions written for their own gender will not be significantly different from those written for the opposite gender. Table Four shows the proportion of correct answers separated by subjects' gender and question type. The proportions for the feminine questions were divided and calculated separately, given if the subject received the test version with the faulty version of question five, or not. Generally, the female subjects performed the same on each type of question, coming in at a 80% correct rate for each, and a 100% correct rate for the female questions out of five. Although it is shown the female and male subjects did significantly best on the feminine questions out of five, this number was averaged within a much smaller sample size than the others, leaving less room for answering the problem incorrectly. This supports my original hypothesis. The male subjects had more variability in their correct rates, and had an increased correct rate on the neutral and feminine questions vs. masculine questions. The male subjects seemed to perform poorest on the

questions classified as masculine, but this number is not significantly lower than their other correct rates. I would argue that the data from the male subjects also supports my fifth hypothesis, given the smallness of the variability.

Discussion

There are multiple factors that could have affected the outcome of the test results. For one, the classroom environment in which it was taken was not controlled. It was not specified that the room should be silent, and chatting amongst students could have been distracting. Students in the Geometry and Trigonometry class also would have been able to recognize that the subject of the math problems differs from their class subject, and therefore does not affect their grade. This could have affected their overall effort, causing subjects correct rates to be lower.

Question Six had a different answer in neutral form than in feminine and masculine form. The mathematical process was the same, but the answer was defined in terms of x for the neutral version, while the other versions had two answers that relied on finding x. It is noted that question six had more skips than any of the other five, as well as a lower range of correct answers.

The process in which I classified the questions feminine and masculine could have been done differently. I based the content topics on my own associations, but it would have been useful to see what content students associate with femininity and masculinity on average, using a scale ranging from femininity to masculinity with neutral in the middle. This could have helped me create my questions with more of a general consensus of femininity vs masculinity, instead of just my own. The subjects perception of the questions is important.The questions also could have been considered ineffective in

tapping into subjects gender bias, given that the content of the problems was simple word substitution (hairbrush vs. baseballs).

Lastly, I think the placement of the demographics questions could have been different. Instead of having them as a cover page to the test, I would have the subjects answer them after taking the test. This would eliminate any possible stereotype threat.

Going forward, I would want to test the effect of familiar vs unfamiliar content on subjects performance on reading comprehension tests. I would similarly create a test with reading comprehension pulled from the SAT, and have subjects rate the familiarity of the subject and compare this to their performance on the tests.

Conclusion:

If given more time and resources, I would do this experiment again on a larger sample. I enjoyed the creation of the tests, and receiving results and analyzing it, and would want a opportunity to do it better. I would follow up with longer questions that require more content knowledge. Subjects did not perform substantially better overall on the neutral classified questions. Subjects did not perform significantly worse on the feminine versions of questions. Female subjects did generally perform at a higher correct rate than male subjects. Both genders performed the same on feminine questions. Lastly, overall, subjects performance on questions written for their own gender was not significantly different from those written for the opposite gender. This tells me that the various different contents did not significantly affect subjects responses. This was a overall enjoyable experiment and I am glad to have looked into the subject of bias and its impact on worldly things, such as tests.

References

"bias." *Merriam-Webster.com*. Merriam-Webster, 2011. Web. 1 June. 2018.

Chipman, Susan. Marshall, Sandra. “Content Effects on Word Problem Performance: A Possible Source of Test Bias?” *American Educational Research Journal* 28, (1991): 897-915. Web. June. 2018.

Jencks, C., & Phillips, M. (Eds.). (2011). *The Black-White test score gap*. Brookings Institution Press.

Race and Ethnicity. Census on Race and Ethnicity. United States Census Bureau, January, 2017.

Torres, S. (2014). Does Gender Bias Within a Test Affect the Subjects Performance on the Test? (Science PBAT). NY, New York.

APPENDIX A - Quiz Variants and Instructions for Teachers

****This page is for the math teachers only and should not be distributed to the students as part of the test.****

1. Please (to the best of your ability) evenly distribute the three versions of the test. I have labeled them test a, b, and c at the top left corner of each.
2. This test should be given with a maximum time limit of twenty minutes. (Obviously if you have less time available in class, that's okay.)
3. It should be done individually, with no group work or contact between students.
4. The answers are important in the sense that I want every person to answer honestly and to the best of their ability. It is most ideal for each question to have a actual answer, and not unfinished answers so I can clearly mark answers correct or incorrect. It is not required for students to show their work, but they must answer without the use of a calculator.
5. Please return the stack to Teacher T. with the name of the math class on top.
6. Thank you for your time!! I greatly appreciate your assistance!!

Test A

Part 1- Please answer each multiple choice question below.

What is your age?

a) 13-14
b) 15-16
c) 17-18
d) 19-20

What is your ethnicity?

a) Hispanic or Latino
b) Non-hispanic

What is your race? (circle all that apply)

a) White
b) Black / African American
c) Asian
d) Native American or Alaska Native
e) Native Hawaiian or Other Pacific Islander
f) Other

What is your sexual orientation? (circle all that apply)

a) Gay
b) Bisexual
c) Lesbian
d) Pansexual
e) Straight
f) Other _____ (please specify)

What is your gender?

a) Male
b) Female
c) Other _____ (please specify)

Turn the page to begin part 2.

Part 2- Please answer these questions to the best of your ability in the allotted boxes. There are six questions, so please make sure to turn the page.

1. 1, A factory contains a series of water tanks, all of the same size. If pump 1 can fill 12 of these tanks in a 12 hour shift, and Pump 2 can fill 11 tanks in the same time, then how many tanks can the two pumps fill, working together, in 1 hour?	2. Ahmed buys a football online for $32.00 per football including tax, and the shipping is $4.00. If he decides to buy more footballs, and it doesn't change the shipping cost, how many can he buy with $100.00?

3. Solve for x: 2x-14 = 42	4. Julie has $50, which was $8 more than twice what Jane has. How much money does Jane have?
5. Solve for x: 2x+4=b	6. A class of 50 women is divided into two groups. One group has 8 less than the other. How many are in each group?

Test B

Part 1- Please answer each multiple choice question below.

What is your age?

a) 13-14
b) 15-16
c) 17-18
d) 19-20

What is your ethnicity?

a) Hispanic or Latino
b) Non-Hispanic

What is your race? (circle all that apply)

a) White
b) Black / African American
c) Asian
d) Native American or Alaska Native
e) Native Hawaiian or Other Pacific Islander
f) Other

What is your sexual orientation? (circle all that apply)

a) Gay
b) Bisexual
c) Lesbian
d) Pansexual
e) Straight
f) Other _____ (please specify)

What is your gender?

a) Male
b) Female
c) Other _____ (please specify)

Now, turn the page to begin part 2.

Part 2- Please answer each question to the best of your ability in the allotted boxes. There are six question, so please make sure to turn the page.

1. If rate A is 12 units per 12 hours and rate B is 11 units per 12 hours, how many units can the two rates produce together in one hour?	2. Shayla buys eye shadow online for $32.00 per unit including tax, and the shipping is $4.00. If she decides to buy more eyeshadow, and it doesnt change the shipping cost, how many can she buy with $100.00?

3. Tom spent 42$ on baseballs. This was 14$ less than twice what he spent for his baseball bat. How much was the bat?	4. Solve for x: 2x+8=50
5. There are b pink soft balls. This is four more than twice the number of purple softballs. How many blue softballs are there?	6. A class of 50 men is divided into two groups. One group has 8 less than the other; how many are in each group?

Test C

Part 1- Please answer each multiple choice question below:

What is your age?

a) 13-14
b) 15-16
c) 17-18
d) 19-20

What is your ethnicity?

a) Hispanic or Latino
b) Non-Hispanic

What is your race? (circle all that apply)

a) White
b) Black / African American
c) Asian
d) Native American or Alaska Native
e) Native Hawaiian or Other Pacific Islander
f) Other

What is your sexual orientation? (circle all that apply)

a) Gay
b) Bisexual
c) Lesbian
d) Pansexual
e) Straight
f) Other _____ (please specify)

What is your gender?

a) Male
b) Female
c) Other _____ (please specify)

Now, turn the page to begin part 2.

Part 2- Please answer each math problem to the best of your ability in the allotted boxes. There are six questions, so please make sure to turn the page.

1. A salon contains an arrangement of hairspray cans, all of the same size. If hairdresser 1 can use 12 of these spray cans in a 12 hour shift, and hairdresser 2 uses 11 in the same amount of time, then how many cans can the two hairdressers use, working together, in 1 hour?	2. A student bought printer cartridges online for $32.00 per including tax, and this shipping is $4.00. If they decided to buy another printer cartridge, and it doesnt change the shipping cost, how many can they buy with $100.00?

3. Jane spent 42$ on bras. This was 14$ less than twice what she spent for her dress. How much was the dress?	4. Tyrone has $50, which was 8$ more than twice what John has. How much money does John have?
5. There are b blue baseballs. This is four more than twice the number of black baseballs. How many black baseballs are there?	6. Solve for x: x+(x-8)=50