Gender Differences in Student Responses to Physics Conceptual Questions Based on Question Context

Laura McCullough
University of Wisconsin-Stout, Physics Department

ABSTRACT

Women's participation in physics remains low, with only 20% of bachelor's degrees and 19% of PhDs in physics being awarded to women. The physics classroom may serve as a barrier to women's participation when the contexts of questions and examples are stereotypically male. To study the possible effects of male contexts on student responses, two versions of a physics conceptual exam were given to students. The first version, a conceptual test in broad use in the field, included stereotypically male contexts; the second, created for the study, included stereotypically female contexts. The physics content of the questions was the same. Examination of the results show that despite identical physics, students can change their responses based on the context presented in the question. Different patterns of change were exhibited: in some instances both men and women did better on female questions, in others only one gender improved. On some questions no change was seen. The conclusion of the study is that the contexts in which physics questions are presented can have an effect on student response, and can show gender differentiation. To encourage participation of all students, instructors need to be aware of the contexts in which they present their material.

Keywords: STEM, Educational Quality, Teaching Quality

INTRODUCTION

The United States has a shortage of talented people entering the fields of science, technology, engineering, and mathematics (STEM). Reports in the last decade have pointed out the dangers of this shortage, and have detailed the need for increasing the numbers of people entering these technical fields (National Academies, 2007 & 2010). One population that is still particularly under-represented in STEM fields is women. This presents us with an opportunity to bring more people into STEM by increasing the participation levels of women.

Data from the National Science Foundation show that women have made some gains in STEM areas, but are still receiving fewer degrees in several fields, particularly physics, engineering, and computer science (NSF, 2011). Figure 1 shows the percentage of bachelor's degrees going to women in STEM fields.

The lower participation of women in science has been studied for many years. Yet despite much research and increasing visibility of this issue, we still do not know exactly why women are less likely to be found in science classes and science careers. In the 2011 report *Women in America*, one of the key findings was this precise problem: "Women earn the majority of conferred degrees overall but earn fewer degrees than men in science and technology." (US Dept of Commerce, 2011). There is a continuing need to study this issue so that we can find out what is

hindering women's participation and what promotes it and implement strategies that will bring more women into the field.

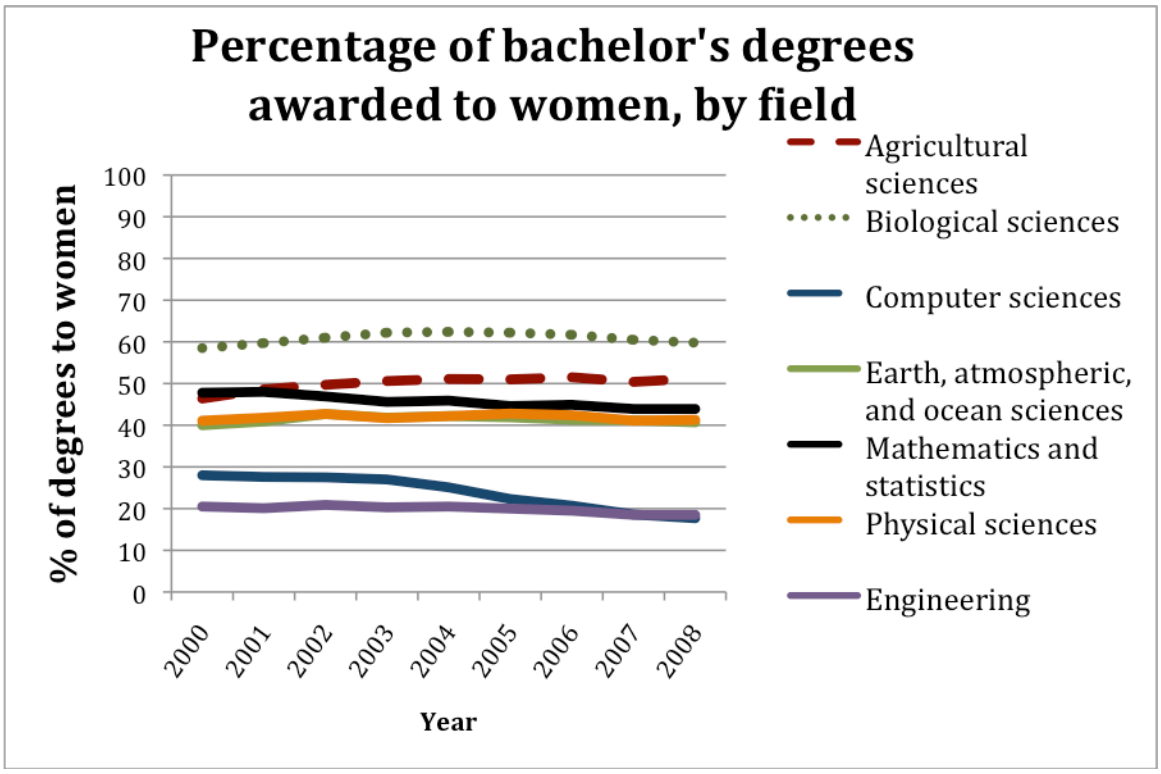

FIGURE 1. BACHELOR'S DEGREES AWARDED TO WOMEN BY FIELD (DATA FROM NSF).

BACKGROUND

This study focuses on one particular arena: the physics classroom. Can the physics classroom promote or hinder women's participation in physics? Most certainly. The area under consideration for this paper is that of tests, and one specific test in particular. The research question addressed is whether or not the context used to frame a test question can affect students' responses to that question. Does a question involving two shopping carts colliding elicit the same response as a question using identical physics, but framing it in terms of two semi trucks colliding?

The context of test questions has been examined extensively in some older studies by Rennie and Parker. In an early paper (Rennie & Parker, 1993) they propose a scheme to categorize problem contexts, looking at language, portrayal of stereotypes, appeal to background experiences, and context. Table 1 gives the criteria for male and female orientations, allegedly neutral orientation, and gender-inclusive orientation.

A small study of eight boys and girls demonstrated that seven of the eight performed better on tasks with non-abstract context (Rennie and Parker, 1996). In a larger study (Rennie and Parker, 1998) they found that context was not strongly linked with performance, but students did show strong preference for different formats. When asked how the context affected their understanding of a multiple-choice item, girls were more likely to say that differences in wording affected their

understanding of the question. Girls were also more likely to find a calculation question with real-life context easier.

TABLE 1. CATEGORIZATION SCHEME PROPOSED BY RENNIE AND PARKER (1993) FOR TEST QUESTIONS.

Criteria	Male Orientation	Female Orientation	Allegedly Neutral Orientation	Gender-Inclusive Orientation
Language	Uses him, he, his	Uses she, her, hers	Uses they, them, their	Uses name of person, use "you"
Portrayal of stereotypes	Males in active role, females in passive role	Females in active role, males in passive role	Genderless people, inactive role (scientist)	Both males and females in active and passive role
Appeal to background experiences	Relevant to stereotyped male experiences	Relevant to stereotyped female experiences	Not relevant to human experiences	Relevant to males and females equally
Context	Decontextualized, abstract	Human, social	Concrete setting	Human, social, environmental

Others have examined the contexts of questions and their effects on performance. A study in South Africa (Enderstein & Spargo, 1998) noted that "the context of the situation used to explore alternative conceptions…substantially affects the frequency with which certain responses are elicited." (pg. 725). Changing a ball rolling off a table to a ball being dropped by a running boy can cause students to view the physics of the situation differently. Another study focused on different types of contextual changes and their effects on a physics test (Stewart, Griffin, & Stewart, 2007). Changes from abstract to concrete, removing or adding figures, and other similar changes showed effects from –3% to +10% in correct responses.

The evidence from the literature supports that idea that changing context can change test response. But can it cause gender bias?

In general, by the time they reach college, young women are less likely to be interested in physics than young men. Several researchers have made arguments that to increase physics interest among women, the context in which physics is taught must be more friendly to women. Murphy & Whitelegg (2006) suggest that questions in a context in which a woman feels incompetent will cause her performance to decline. Repeated exposure to these questions can lead to less interest, less participation, and less achievement. Baram-Tsabari and Yarden (2008) argue for using contexts that are girl-friendly as a way to help increase girls' participation in physics, and perhaps to help make physics more appealing to everyone.

If context affects performance, and performance affects interest and participation, then we need to be sure our tests in the physics classroom are not hindering women's performance and possibly their participation. This study is designed to determine if gender-specific contexts create gender-biased performance on a particular physics test.

THE STUDY AND METHODS

In this study, a commonly used conceptual physics test was examined. The Force Concept Inventory (FCI) is a 30-question, multiple-choice, introductory physics test used across the U.S. by physics instructors at the high school, community college, and university levels (Hestenes, Wells, and Swackhamer, 1992). The test has been shown to have a performance bias favoring males (Kost, Pollock & Finkelstein, 2009; Lorenzo, Crouch & Mazur, 2006, McCullough, 2002). This bias ranges from 10-30% on average, and is seen at all education levels in which it is administered. The background of students is a likely factor in explaining this gap. Yet examining background shows that because males are more likely to have taken physics prior to the test does not explain the improved performance of males. Background does not account for the entire performance gap (McCullough, 2002; Kost, Pollock & Finkelstein, 2009).

In trying to determine possible reasons for this gender bias, the investigator noted the contexts of the questions on the test. The test includes many stereotypically male contexts such as hockey, rockets, trucks, cannonballs, and male figures. This suggested an interesting question: do these stereotypically male contexts contribute to the gender bias in performance on the test?

To determine this, a second version of the test was created which used exclusively stereotypical female contexts: jewelry, shopping, cooking, figure skaters, etc. The physics of the questions was kept identical; only the context of the question was changed. To maintain the validity of the physics content, the questions were reviewed by other physics instructors. The reason the contexts were changed to stereotypically female was to maximize any potential changes to make them more clearly observable. Changing to a more neutral context might produce changes too small to see; if changes are to be notable, the differences should be maximized.

The two versions of the test were given to students in a calculus-based physics course over several semesters. The original version was given first, and the adjusted version was given second, usually one day later. Previous research (Henderson & Heller, 2000) has shown that a retake of this test does not significantly improve student performance. The tests were given both as a pre-test (during the first week of instruction) and as a post-test (during the last week of instruction). Data from pre-test and post-test were kept separate and analyzed separately. Incomplete tests and obvious false answer sets (e.g., all B answers) were removed. This led to slight differences in numbers of students in the samples. Analysis was focused both on overall test score and individual question characteristics.

RESULTS

The biggest question is of course whether or not the revised version of the test changed the overall score of students. Table 2 shows the average score for all students for the pre-test condition and the post-test condition, by version of the test.

At this level of analysis there is not a large enough difference between the versions to suggest any overall conclusion. There is a statistically significant difference ($p<0.01$) between the pre-test scores, but it is of lesser educational significance. This data is combining all people together; are there gender differences in overall score? Table 3 gives the same data by standard gender divisions.

TABLE 2. AVERAGE SCORE ON EACH VERSION OF THE TEST, BY TIME OF ADMINISTRATION.

	Average score out of 30, standard deviation
Original version, pre-test (N=283)	9.2 (30.5%), 4.0
Revised version, pre-test (N=225)	10.6 (35.3%), 4.5
Original version, post-test (N=340)	13.8 (46.1%), 5.4
Revised version, post-test (N=378)	13.2 (43.9%), 6.0

From this data, we can see that the women definitely show lower performance than men, with typically 10-15% lower scores. All of these gender differences are statistically significant differences with $p<0.01$ (2-sample t-test). This bias in favor of men holds with both versions of the test, and is present at both the beginning and end of the class. The revised version of the test does not produce any reduction of the gender gap on overall score.

TABLE 3. AVERAGE SCORE ON EACH VERSION OF THE TEST, BY GENDER AND TIME OF ADMINISTRATION.

Gender	Average score out of 30, standard deviation
Original version, pre-test	
Females (N=99)	7.0 (23.5%), 2.8
Males (N=184)	10.3 (34.3%), 4.0
Revised version, pre-test	
Females (N=93)	8.8 (29.4%), 3.8
Males (N=132)	11.8 (39.4%), 4.6
Original version, post-test	
Females (N=93)	10.7 (35.6%), 4.0
Males (N=247)	15.0 (50.1%), 5.4
Revised version, post-test	
Females (N=121)	11.4 (38.0%), 5.4
Males (N=157)	14.5 (48.4%), 6.1

How reliable are the results? One measure of reliability is Cronbach's alpha. Given the larger standard deviations on the revised tests, one might expect that the reliability may have suffered. Table 4 lists the reliability of all eight conditions of the test, based on the 30 questions of the test.

TABLE 4. CRONBACH'S ALPHA FOR EACH VERSION OF THE TEST, BY GENDER, AND BY TIME OF ADMINISTRATION.

Gender	Cronbach's alpha			
	Pre-test Original	Pre-test Revised	Post-test Original	Post-test Revised
Females	.414	.647	.655	.811
Males	.688	.755	.811	.859

As one would hope, the post-test reliabilities are better than the pre-test numbers. The reliability of the test for men tends to be much better than for women. If women show poorer performance, they may be guessing more, which would lead one to expect lower reliability scores. The

reliability of the revised version is an improvement over the original version, often a large one, which is a bit of a surprise. Given the overall scores were similar, and the distributions not too dissimilar, one might have expected to see similar reliabilities. Without further research, it is not possible to state why the revised version has a higher reliability. One possibility is that the consistent daily-life contexts of the questions puts them in the same "space" for students, whereas the original includes a mix of both abstract and contextualized problems.

The overall scores suggest that while the revised version is not harming anything, it is not helping anything either. Yet average scores do not tell the whole story, as it is possible that there may be significant effects on individual questions that are washed out by viewing the test performance as a whole. Examining individual questions can provide an interesting case study on how contexts may affect performance. Two curious cases are described below.

Question #19: Blocks vs. joggers

In a question aimed at eliciting students' knowledge of motion and kinematics, students look at a figure of two blocks moving to the right, shown at successive time intervals (Figure 2). The question is asking about the relative speed of the two blocks.

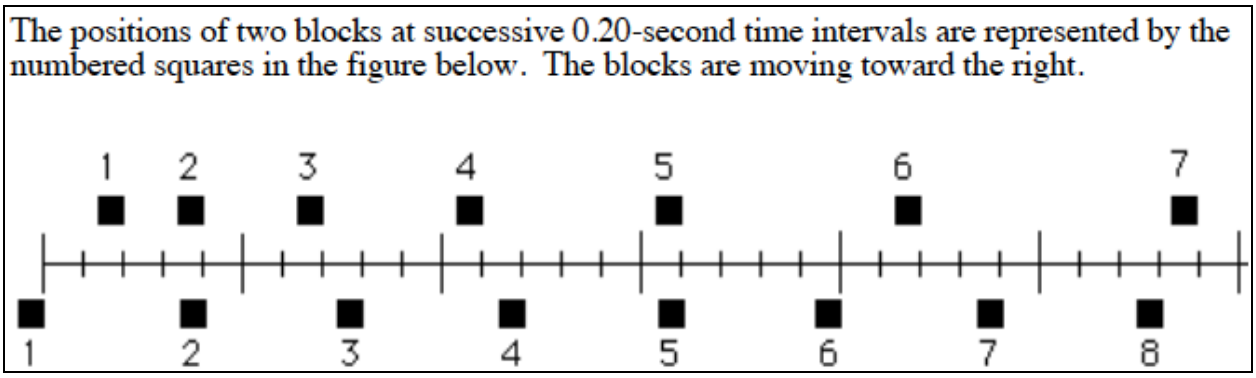

FIGURE 2. QUESTION 19 ON ORIGINAL TEST.

In the revised version, the blocks are replaced with two female joggers (Figure 3).

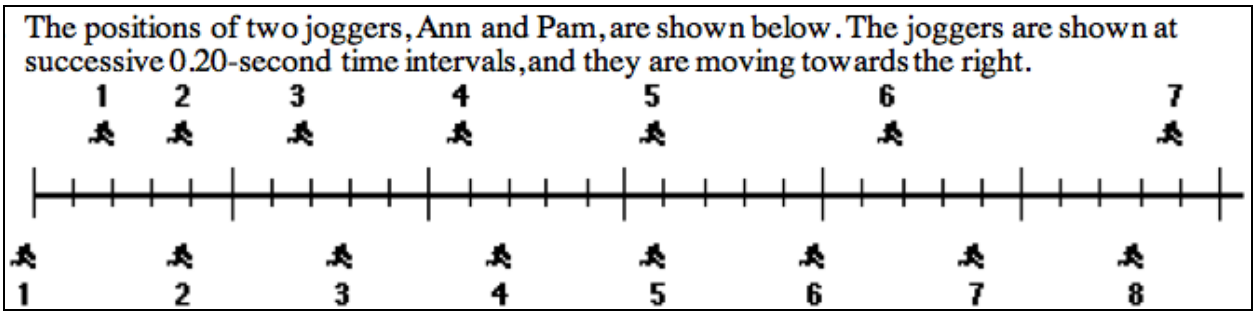

FIGURE 3. QUESTION 19 ON REVISED TEST.

The change is minor; the physics is identical. The front of the foot of the runner matches with the front edge of the block. Yet to students these questions provoke different responses.

In Table 5 we see that the relatively minor change of an abstract block to a figure of a jogger improved students' scores by 11-18%. All but the men's post-test are statistically significant at p=0.05. The physics is the same; the visual is changed very little. Somehow making it more concrete and less abstract allowed all students to perform better on this question.

TABLE 5. AVERAGE % CORRECT ON QUESTION 19 BY TEST VERSION, TIME OF ADMINISTRATION, AND GENDER.

Gender	Question 19 Pretest: Average % correct	
	Original version	Revised version
Women	32	48
Men	42	58
	Question 19 Post-test: Average % correct	
	Original version	Revised version
Women	34	52
Men	50	61

Question #4: Vehicles vs. Shopping Carts

A different pattern is seen with question 4. Here the change consisted of switching the objects involved in a collision. The original question 4 reads as in Figure 4.

4. A large truck collides head-on with a small compact car. During the collision:
   (A) the truck exerts a greater amount of force on the car than the car exerts on the truck.
   (B) the car exerts a greater amount of force on the truck than the truck exerts on the car.
   (C) neither exerts a force on the other, the car gets smashed simply because it gets in the way of the truck.
   (D) the truck exerts a force on the car but the car does not exert a force on the truck.
   (E) the truck exerts the same amount of force on the car as the car exerts on the truck.

FIGURE 4. QUESTION 4 ON ORIGINAL TEST.

The revision changes the vehicles to shopping carts, as seen in Figure 5.

4. Imagine a head-on collision between a very full shopping cart and an empty cart. Both carts are moving very quickly. During the collision,

   (A)the full cart exerts a greater amount of force on the empty cart than the empty cart exerts on the full cart.
   (B)the empty cart exerts a greater amount of force on the full cart than the full cart exerts on the empty cart.
   (C)neither exerts a force on the other, the empty cart gets smashed simply because it gets in the way of the full cart.
   (D) the full cart exerts a force on the empty cart but the empty cart doesn't exert a force on the full cart.
   (E) the full cart exerts the same amount of force on the empty cart as the empty cart exerts on the full cart.

FIGURE 5. QUESTION 4 ON REVISED TEST.

Here again we see students responding differently to the different contexts (Table 6), but in this case the changed question produced worse results.

TABLE 6. AVERAGE % CORRECT ON QUESTION 4 BY TEST VERSION, TIME OF ADMINISTRATION, AND GENDER.

	Question 4 Pretest: Average % correct	
	Original version	Revised version
Women	16	12
Men	15	18
	Question 4 Post-test: Average % correct	
	Original version	Revised version
Women	34	23
Men	39	26

Based on the pre-test, we see that students are choosing the correct answer less often than pure guessing would show. This suggests that there are strong misconceptions present that the distractors/wrong answers on the question are drawing out. There are minor gender differences on the pre-test; both men and women have strong beliefs that do not match current physics understanding.

On the post-test, we see a significant difference between the tests. The statistically significant result is the men's post-test at $p=0.025$. Despite the women's post-test result not being statistically significant, it is of interest to note that a similar drop occurred for women, with the revision dropping the percentage of correct scores by over 10%, and bringing the numbers down close to what guessing would predict.

When one examines the 30 questions of the test, one sees a small pattern of positive changes within the questions with the new test. For the pre-instruction condition, women improved their score by over 5% on 18 questions. Men improved by more than 5% on 13 questions. The revised version decreased average scores by more than 5% on 4 questions for women, and only 1 for men. So even though the revision did not change the overall score by a great deal, nearly half the questions showed some small improvement with the revision.

For the post-instruction tests, women improved their score by over 5% on 10 questions, and men on 4 questions. The revision hurt performance by more than 5% on 4 questions for women, and 9 for men. So after students have learned some physics, the revised version does a worse job at improving performance on a question-by-question level, and is more likely to hurt men's performance than help on any single question.

Another interesting way to look at the data is to see how many questions helped both men and women improve their scores. On the pre-test, 11 of the 30 questions showed improvement for both men and women by over 5%. On the post-test, this dropped to only 2. At the beginning of the class, there were no questions on which both men and women did worse on the revised test by more than 5%. At the end of the test, 3 questions had both men and women scoring worse by over 5%. These analyses show that there is a definite difference between the pre-test condition

and the post-test condition. Further research will be needed to tease out more of these differences and any underlying reasons.

CONCLUSIONS

The evidence from this study supports the finding in the literature that changing the context of a physics problem can affect how students respond to the question. Changing a commonly-used physics test from one in which the questions have mostly stereotypically male contexts to one in which it has all stereotypically female contexts did not have a significant impact on the overall score on the test for either men or women. Yet within individual questions there was a large variety of patterns displayed.

The revised version of the test had much stronger reliability scores than the original version did. And both versions of the test showed a performance bias in favor of men. The revised test was unable to close this gap, but it did not widen it.

By examining the patterns on individual questions, it may be possible to develop theories on what types of changes to context cause gender-differentiated changes in responses. With further research, it may be possible to develop a version of the test that does not exhibit any gender performance bias related to context. This is the ultimate goal of this research program.

ACKNOWLEDGMENTS

The author acknowledges the support of several colleagues who helped significantly with the development of the revised test version: Patricia Heller, Thomas Foster, and David Meltzer.

REFERENCES

Baram-Tsabari, A. and Yarden, A. 2008. "Girls' biology, boys' physics: evidence from free-choice science learning settings." *Research in Science and Technological Education*, 26, no. 1 (April 2008): 75-92.

Enderstein, L. and Spargo, P. 1998. "The effect of context, culture and learning on the selection of alternative options in similar situations by South African pupils." *International Journal of Science Education*, 20 no. 6: 711-736.

Henderson, C., and Heller, P. 2000. "Common Concerns about the Force Concept Inventory." Presentation at American Association of Physics Teachers meeting, January 2000, Kissimmee, FL. http://groups.physics.umn.edu/physed/Talks/Charles1-00.pdf

Hestenes, D., Wells, W., and Swackhamer, G. 1992. "Force Concept Inventory." *The Physics Teacher*, 30 no. 3: 141-151.

Kost, L., Pollock, S. and Finkelstein, N. 2009. "Characterizing the gender gap in introductory physics." *Physical Review Special Topics-Physics Education Research*. 5 no. 1: 010101. http://prst-per.aps.org/toc/PRSTPER/v5/i1

Lorenzo, M., Crouch, C.H., and Mazur, E. 2006. "Reducing the gender gap in physics." *American Journal of Physics*, 74 no. 2: 118-122.

McCullough, L. 2002. "Gender, Math, and the FCI." *Proceedings of the 2002 Physics Education Research Conference.* S. Franklin, J. Marx, & K. Cummings, Eds. Rochester NY.

Murphy, P., and Whitelegg, E. 2006. "Girls and physics: Continued barriers to 'belonging'." *The Curriculum Journal*, 17 no. 3 (September 2006): 281-305.

National Academies. 2007. *Rising Above the Gathering Storm.* Washington, DC: National Academies Press. http://www.nap.edu/catalog.php?record_id=11463#orgs

_____. 2010. *Rising Above the Gathering Storm, Revisited.* Washington, DC: National Academies Press. http://www.nap.edu/catalog.php?record_id=12999

National Science Foundation. 2011. *Women, Minorities, and Persons with Disabilities in Science and Engineering*. Arlington, VA: National Science Foundation (NSF 11-309). http://www.nsf.gov/statistics/wmpd/

Rennie, L. and Parker, L. 1993. "Assessment in physics: Further exploration of the implications of item context." *Australian Science Teachers Journal*, 39 no. 4: 28-32.

Rennie, L. and Parker, L. 1996. "Placing physics problems in real-life context: Students' reactions and performance." *Australian Science Teachers Journal*, 42 no. 1: 55-59.

Rennie, L. and Parker, L. 1998. "Equitable measurement of achievement in physics: High school students' responses to assessment tasks in different formats and contexts." *Journal of Women and Minorities in Science and Engineering*, 4 no. 2: 113-127.

Stewart, J., Griffin, H. and Stewart, G. 2007. "Context sensitivity in the Force Concept Inventory." *Physical Review Special Topics-Physics Education Research*. 3 no. 1: 010102. http://prst-per.aps.org/toc/PRSTPER/v3/i1

U.S. Department of Commerce. 2011. Women in America. Washington, DC: Department of Commerce. http://www.whitehouse.gov/administration/eop/cwg/data-on-women

AUTHOR INFORMATION

Dr. McCullough is the chair of the Department of Physics at the University of Wisconsin-Stout. She has a Ph.D. in Science Education and has studied gender and science issues as well as physics education research for 15 years. She may be contacted at McCulloughL@uwstout.edu.