American Economic Review: Papers & Proceedings 2008, 98:2, 370–375
http://www.aeaweb.org/articles.php?doi=10.1257/aer.98.2.370

Exploring the Impact of Financial Incentives on Stereotype Threat: Evidence from a Pilot Study

By Roland G. Fryer, Steven D. Levitt, and John A. List*

Motivated in part by large and persistent gender gaps in labor market outcomes (e.g., Claudia Goldin 1994; Joseph G. Altonji and Rebecca M. Blank 1998), a large body of experimental research has been devoted to understanding gender differences in behavior and responses to stimuli.[1] An influential finding in experimental psychology is the presence of stereotype threat: making gender salient induces large gender gaps in performance on math tests (Steven J. Spencer, Claude M. Steele, and Diane M. Quinn 1999). For instance, when Spencer et al. (1999) informed subjects that women tended to underperform men on the math test they were about to take, women's test scores dropped by 50 percent or more compared to a similar math test in which subjects were not informed of previous gender differences. In this latter treatment, men and women perform similarly. Stereotype threat research typically is carried out in the absence of financial rewards for performance.[2]

In this paper, we report the results of a pilot experimental study examining gender differences in math proficiency. Our 2 (presence or absence of a stereotype characterization) × 2 (with and without financial incentives) between subjects design extends the literature in at least two dimensions. First, we test for stereotype threat effects in an environment that also includes financial incentives ($2 per correct answer). By doing so, we pit experimenter demand effects and stereotype threat effects. To the extent that findings in the stereotype threat literature are driven by experimenter demand effects (i.e., women do badly when gender is emphasized because they think the experimenter expects this), paying for performance raises the cost of women accommodating the experimenter, potentially lessening the influence of stereotype threat.[3] The importance of experimenter demand effects is well documented, but estimating the extent to which they are sensitive to price remains an open research question (Levitt and List 2007a). Alternatively, by raising the stakes, the increased financial incentives may serve to increase the stress associated with the test and exacerbate stereotype threat effects.

The second contribution of this research is that we provide one of the first explorations of how the responses of men and women change on a cognitive test when moving from an environment with no pecuniary incentives tied to performance to one where subjects are rewarded with a piece rate scheme (see also Gneezy, Muriel Niederle, and Aldo Rustichini (2003), although this comparison is not emphasized in their work). We view our paper as a complement to the recent

* Fryer: Department of Economics, Harvard University, Littauer Center 208, Cambridge, MA 02138, and NBER (e-mail: rfryer@fas.harvard.edu); Levitt: Department of Economics, University of Chicago, 1126 E. 59th St., Chicago, IL 60637, American Bar Foundation, and NBER (e-mail: slevitt@uchicago.edu); List: Department of Economics, University of Chicago, 1126 E. 59th St., Chicago, IL 60637, and NBER (e-mail: jlist@uchicago.edu). Lint Barrage and Min Lee provided exceptional research assistance. Seda Ertac and Muriel Niederle provided insightful comments. Financial support of the National Science Foundation and the Sherman Shapiro Research Fund are gratefully acknowledged.

[1] On gender research in psychology see A. H. Eagly (1995) and the subsequent debate in the American Psychologist (February 1996); in economics see Rachel Croson and Uri Gneezy (2008).

[2] We were able to find one study that used performance incentives (Lynn McFarland, Dalit Lev-Arey, and Jonathan Ziegert 2003). This work examined stereotype threat as it relates to race, and uses a cognitive test and a personality test. They implemented the incentives via a competitive scheme: people who scored in the top 15 percent in both tests received $20.

[3] Consistent with this hypothesis, both men and women report a greater level of stress when taking the math test in the stereotype threat treatments than in the baseline. It has long been recognized that the performance of the nonstereotyped group improves when stereotype threat is introduced (this phenomena is denoted "stereotype boost"; see, e.g., Steele and Joshua Aronson 1995; Spencer, Steele, and Quinn 1999), although this increase in performance has generally not been ascribed to increased incentives as we argue in this paper.

flurry of papers in the economics literature that document that women perform worse relative to men on tasks such as completing mazes in a competitive environment (e.g., Gneezy et al. 2003; Niederle and Lise Vesterlund 2007). In these studies, a piece rate form of compensation is used as the baseline against which a competitive incentive environment is compared; in our case, the comparison is no financial incentive versus a piece rate. By combining our two contributions, we also provide an apples-to-apples comparison of the effect of stereotype threat with the effect of modest incentives. We are aware of no other pricing exercise of this type in the stereotype threat literature.

A mixed set of results emerges from our experiment. First, absent financial incentives, we do not reproduce the standard finding that female performance declines in absolute terms when the experimental instructions include a passage emphasizing that men outperform women on tests of this kind. Indeed, of our four treatment cells, the stereotype condition without financial incentives is the variant in which women perform the best. We cannot, however, reject the null hypothesis that women perform identically across all four treatment cells.

Second, and at odds with the experimenter demand effect hypothesis, if anything, the introduction of financial incentives appears to exacerbate gender differences, with or without the presence of stereotype threat language. This second result is closely related to our third finding, which is that male test scores rise when either stereotype characterization or financial incentives are introduced. The number of questions answered correctly by males increases by a statistically significant average of 18 percent in these treatments relative to our baseline. Finally, in exploring potential mediators (self-reports of stress induced by this test, general test anxiety, and proxies for effort), we find consistent impacts of our treatments relative to the baseline. Introducing either the stereotype message or financial incentives increases the stress levels of women more than men. Male effort rises in response to these treatments, whereas it is less apparent that female effort increases.

I. Experimental Design

In the fall of 2007, we recruited University of Chicago students via flyers and e-mail lists to participate in the study.[4] Participants were not informed about the nature of the experiment we would be conducting or the treatment to which they would be assigned. Subjects were promised a $5 show-up fee, plus the chance to earn additional money for participating in an experiment. Initial response induced us to schedule eight sessions over a three-day period.

Upon arriving at the experiment, a subject's experience followed five steps. In Step 1, all subjects signed a consent form that they were willingly participating in a study of "SAT-style math problems." In Step 2, a copy of the instructions was distributed to each subject and the monitor read the instructions aloud. All subjects in each session were placed in one of the four treatment cells.

In the baseline treatment, participants were simply told that, "Today you will take a test that includes 20 questions. You will each have 20 minutes to work on these questions."[5] The experimenter then explained how financial compensation would occur. In our baseline treatment, there was no financial incentive; subjects were simply paid a fixed amount equal to $20 for their participation regardless of their performance on the test. In our "financial incentive" treatment, subjects were given $5 for showing up and $2 per correct answer.

In our stereotype threat treatments, the experimental instructions concluded with the statement: "This is a diagnostic test of your mathematical ability. As you may know, there have been some academic findings about gender differences in math ability. The test you are going to take today is one where men have typically outperformed women." This wording closely parallels that used by Spencer et al.

[4] When designing an experiment in this area, past research suggests that four conditions are necessary conditions to produce stereotype threat: (a) ability evaluation—the test is diagnostic of the targets' ability (Steele and Aronson 1995); (b) domain identification—the subjects care about the domain, and they use the domain as the basis of self-evaluation (Aronson et al. 1999); (c) test difficulty—the test is difficult (Spencer et al. 1999); and (d) stereotype applicability—the stereotype is relevant to the subjects (e.g., Spencer et al. 1999). We were guided by these four conditions in designing our experiment.

[5] Full experimental instructions are available from the authors.

(1999) in their seminal exploration of gender stereotype threat.

Crossing these two dimensions yields four unique experimental cells: (a) a baseline with no stereotype threat and no financial incentives, (b) stereotype threat but no financial incentives, (c) financial incentives without stereotype threat, and (d) both stereotype threat and financial incentives.

Concluding the second step, subjects were informed that their earnings would be paid in private at the completion of the study, at which time they would also be informed of their achievement on the test—the raw number of correct answers. Further, subjects were told that there was no penalty for incorrect answers. Finally, subjects were told that they had 20 minutes to answer the 20 multiple choice math questions. The questions were identical across treatment and were taken from SAT and GRE study guides.

In Step 3 the subjects completed the exam. Each subject was seated in a classroom with dividers placed between subjects to mitigate spillovers of answers across subjects. After taking the test, but before learning their results, in Step 4 subjects completed a survey that asked about a variety of background characteristics, and anxiety toward test-taking in general and on this test in particular. Step 5 concluded the experiment, with subjects being paid their earnings in private.

A total of 79 men and 61 women participated in the experiment. Subjects were primarily business school students and undergraduates. The number of subjects of a particular gender in a given experimental cell varied from 12 to 21. Subjects in the financial treatments earned slightly less than those in the nonfinancial treatments ($22 versus $25).

II. Results

Table 1 summarizes the results. Background characteristics proved not to be particularly well balanced across treatment groups. Thus, the results we report in the table condition on a range of predetermined characteristics: race, gender, number of college math courses, self-reported SAT math score, MBA student, and age. These covariates together explain approximately 35 percent of the observed variation in test scores.[6] The patterns observed in the raw data are consistent with those presented in the table, but noisier.

The columns in Table 1 correspond to different outcomes, each of which is reported separately by gender. The first two columns correspond to the number of questions answered correctly. Columns 3 and 4 present the self-reported level of stress felt on this test on a ten-point scale; columns 5 and 6 are self-reported responses to test-related anxiety more generally. The remaining columns reflect whether the subject reported guessing at the end of the test, the total number of questions answered (whether correct or incorrect), and whether the subject reports that they are at the seventy-fifth percentile or above in math among those in the University of Chicago community. Odd columns present coefficients for males; even columns correspond to female subjects.

The top row of Table 1 reports means by gender for each of these outcomes in our baseline treatment (neither stereotype threat nor financial incentives). The next three rows report estimated treatment effects for each of the treatment cells separately. In all cases, the omitted category is the baseline treatment, making the coefficients relative to baseline. Standard errors are in parentheses. The final row of the table pools our three treatment cells to report an overall impact of any treatment relative to the baseline. Although theory might generate very different predictions across our three treatments, for most of our outcomes the three treatments induced similar responses, making this pooled estimate perhaps of some interest.

The first two columns of the table present our findings with respect to the number of questions answered correctly. In our baseline treatment, the average score among men is roughly one additional correct answer greater than the average score among women (8.16 versus 7.25). Interestingly, regardless of which treatment is imposed, male test scores rise by over a point,

[6] In our data, being male, Asian, young, pursuing an MBA, having taken multiple college math courses, and having high SAT math scores are associated with more correct answers. For all of the outcomes we consider, we control for these observable characteristics using OLS regressions constraining the impact of these covariates to be identical across all treatment cells, using the residuals from these regressions as the basis for the results reported in Table 1.

TABLE 1—EXPERIMENTAL RESULTS

	Questions correct		Stress on current test		General test anxiety		Total questions answered		Guess at the end of test?	
	(1) Male	(2) Female	(3) Male	(4) Female	(5) Male	(6) Female	(7) Male	(8) Female	(9) Male	(10) Female
Mean in baseline	8.16	7.25	4.47	3.44	2.57	2.09	16.00	16.94	0.53	0.69
Estimated treatment effect of										
Sterotype threat	1.17	0.87	0.48	1.96*	−0.48*	0.30	2.38*	0.32	0.25*	0.15
	(0.80)	(0.966)	(0.73)	(0.83)	(0.21)	(0.27)	(0.98)	(1.35)	(0.12)	(0.14)
Financial incentives	1.48	0.32	0.16	1.71	−0.42	0.38	3.04**	1.51	0.30*	0.23
	(0.80)	(1.03)	(0.73)	(0.89)	(0.21)	(0.28)	(0.98)	(1.43)	(0.12)	(0.15)
Stereotype treat and financial incentives	1.85*	0.05	0.59	1.04	−0.47*	0.41	2.45*	0.26	0.29*	0.15
	(0.83)	(0.92)	(0.76)	(0.80)	(0.22)	(0.25)	(1.02)	(1.29)	(0.13)	(0.14)
Estimated treatment effect of										
Any treatment	1.49*	0.40	0.40	1.53*	−0.46*	0.37	2.63**	0.62	0.28**	0.17
Relative to baseline	(0.66)	(0.77)	(0.60)	(0.67)	(0.18)	(0.21)	(0.81)	(1.08)	(0.10)	(0.11)

Notes: The values in this table are regression estimates for dummy variables corresponding to different treatment cells in our 2×2 design, which crosses financial incentives and stereotype threat treatments. The omitted category is our baseline treatment (nonfinancial incentives, no stereotype threat); all estimates are relative to that baseline. The dependent variable is listed at the top of each column. Included in the regressions, but not shown in the table, are controls for race, gender, number of college math classes, self-reported SAT math score, whether the student is getting an MBA, and age. Each entry in the table is from a different regression. Results are shown separately for men and women. Standard errors are in parentheses. The top row of the table reports sample means by gender in the baseline treatment.

* Denotes significance at 0.05 level.

** Denotes significance at 0.01 level.

although the increase is statistically significant only at the $p < 0.05$ level in the treatment that interacts stereotype threat and financial incentives (coefficient of 1.85 with a standard error of 0.83). Pooling across the treatments, the overall effect on males is also positive and statistically significant at conventional levels.

More importantly for our purposes, at odds with the literature on stereotype threat, we find little effect of stereotype characterization for women. Indeed, we find quite the opposite: women are at their best in the stereotype threat only treatment. In the financial incentives only treatment, women's test scores rise by 0.32 (standard error of 1.03) from the baseline, or less than one-fourth as much as men's scores increase, although this difference is not statistically significant. Relative to men, women lose ground in the treatment with both financial incentives and stereotype threat.[7]

The next four columns explore test-related stress. Columns 3 and 4 report subject responses to the question, "How stressful was this test for you?" with responses given on a ten-point scale, with one corresponding to "not at all" and ten meaning "a great deal." Men report only a slight increase in stress associated with adding incentives, and these changes are not statistically significant. On the other hand, while women have a lower level of stress in the baseline treatments, they experience a statistically significant increase in two of the three treatments. Pooling across treatments, the average increase in stress for women is 1.53 points (standard error equal to 0.67).

[7] Similar results are obtained when we exclude those subjects who self-identified as having below-average math ability; stereotype threat is thought to be most powerful among those who place high value on the activity. If cueing gender stereotypes is the mechanism through which stereotype threat statements drives a wedge between gender performance, one might expect that treatment to trigger very different responses by gender to self-reported math proficiency. The fraction of men ranking themselves highly in math increases by 50 percent after participating in a treatment involving incentives, although imprecise estimates leave the difference statistically insignificant. Women's self-reported math ranking is not sensitive to the treatments.

Columns 5 and 6 present results corresponding to general test-related anxiety, as opposed to the stress felt on this particular test. Our anxiety measure is the average response on a four-point scale across four questions (e.g., "I feel very panicky when I take an important test," "During examinations I get so nervous that I forget facts that I really know").[8] The results displayed in columns (5) and (6) clearly demonstrate that treatment exposure influences how subjects respond to these questions. Men exposed to the incentive treatments report *lower* levels of test-taking anxiety (coefficient of −0.45 with a standard error of 0.17), whereas women in the incentive treatments report *higher* levels of test anxiety in the incentive treatments (coefficient of 0.37 with a standard error of 0.21). Thus, both of our measures of stress/anxiety paint a consistent pattern in which introducing incentives increases self-reported stress for women, but not for men. The differential stress response provides one mediator for explaining why women's relative performance declines in our treatments. This result strengthens insights gained from the data in Spencer et al. (1999), who report that their stereotype manipulation had a marginally significant effect on anxiety.

Effort is a second possible channel through which our manipulations might operate. Although we do not directly observe effort, we gathered two crude proxies: whether the subject reports guessing at the end of the time period, and the total number of questions answered (regardless of whether the correct response is given). As shown in columns (7) and (8), men leave an average of four (20 percent) of the questions blank in the baseline and women fail to answer roughly three (15 percent) of the questions, despite the fact that there is no penalty for guessing. Similarly, only about half of the men report guessing at the end of the baseline, with roughly two-thirds of women reporting that they guessed (columns (9) and (10)).

Consistent with the earlier test score results, exposure to any of our treatments has a clear impact on the effort exerted by males, but has little apparent impact on women. Men answer a statistically significant two to three extra questions in the treatments, and the percent reporting that they guess at the end of the time period increases by 25 to 30 percentage points. The point estimates for women are positive on both of these measures, but in each case smaller in magnitude (only one-fourth as large on number of questions answered) and statistically insignificant. Even if the subjects were purely guessing on the additional questions to which they gave responses, this channel can account for nearly 40 percent of the gender difference in test results observed.

[8] The correlation between the responses to these four questions was extremely high, leading us to combine them into a single index. The correlation between this general index of test anxiety and the amount of stress felt on this particular test is approximately 0.4.

III. Conclusion

Researchers of stereotype threat have interpreted their findings as a force that influences women's participation in math-related curricula and professions (Spencer et al. 1999, 6–7). Interestingly, economists have had little to say in this literature. This paper presents a first step in that direction. Our results thus far raise more questions than provide definitive answers. To what do we attribute the absence of stereotype threat behavior in our data when the prior literature has generated such powerful results? Is the greater responsiveness of men to financial incentives a pattern that generalizes beyond our study? If that is the case, should we expect to see women self-selecting away from careers with productivity related rewards toward fixed-wage jobs? We believe that the issues we touch upon in this paper are ripe for further exploration.

REFERENCES

Altonji, Joseph G., and Rebecca Blank. 1998. "Race and Gender in the Labor Market." In *Handbook of Labor Economics Volume 3C*, ed. Orley Ashenfelter, and David Card, 3143–3259. Amsterdam: Elsevier.

Aronson, J., Michael J. Lustina, Catherine Good, Kelli Keough, Claude M. Steele, and Joseph Brown. 1999. "When White Men Can't Do Math: Necessary and Sufficient Factors in Stereotype Threat." *Journal of Experimental Social Psychology*, 35(1): 29–46.

Croson, Rachel, and Uri Gneezy. 2008. "Gender Differences in Preferences." Unpublished.

Eagly, A. H. 1995. "The Science and Politics of Comparing Women and Men." *American Psychologist*, 50(3): 145–58.

Eagly, A. H. 1996. "Differences Between Women

and Men: Their Magnitude, Practical Importance, and Political Meaning." *American Psychologist*, 51(2): 158–59.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *Quarterly Journal of Economics*, 118(3): 1049–74.

Goldin, Claudia. 1994. "Understanding the Gender Gap: An Economic History of American Women." In *Equal Employment Opportunity: Labor Market Discrimination and Public Policy*, ed. Paul Burstein, 17–26. Edison, NJ: Aldine Transaction.

Levitt, Steven, and John List. 2007a. "Viewpoint: On the Generalizability of Lab Behavior to the Field." *Canadian Journal of Economics*, 40(2): 347–70.

Levitt, Steven, and John List. 2007b. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives*, 21(2): 153–74.

McFarland, Lynn, Dalit Lev-Arey, and Jonathan Ziegert. 2003. "An Examination of Stereotype Threat in a Motivational Context." *Human Performance*, 16(3): 181–205.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, 122(3): 1067–1101.

Spencer, Steven, Claude Steele, and Diane Quinn. 1999. "Stereotype Threat and Women's Math Performance." *Journal of Experimental Social Psychology*, 35(1): 4–28.

Steele, Claude. 1997. "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance." *American Psychologist*, 52(6): 613–29.

Steele, Claude, and Joshua Aronson. 1995. "Stereotype Threat and the Intellectual Test Performance of African Americans." *Journal of Personality and Social Psychology*, 69(5): 797–811.