STEREOTYPE THREAT, GENDER, AND MATH PERFORMANCE: EVIDENCE FROM THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Current Draft: September 15, 2009 [1]
First Draft: January 11, 2008

Thomas E. Wei

Harvard University

Abstract

Stereotype threat posits that when individuals are primed about a fixed characteristic of theirs that is negatively stereotyped in relation to a task (i.e., girls cannot do math), subsequent performance on the task is adversely affected. A limitation of most stereotype threat studies is that they have been based upon small and non-representative samples in lab settings, raising concerns about their external validity. I use micro data from the 1978 to 1999 National Assessment of Educational Progress (NAEP) – a large and representative field assessment of 9, 13, and 17 year-old U.S. children – where through a design quirk, students are randomly assigned to test blocks, some of which include gender prime questions while others do not. I exploit this natural experiment by comparing the gender gap in math test scores of students receiving gender primes to those with placebo questions. I find little evidence of negative stereotype threat and strong evidence of stereotype reactance with girls performing better relative to boys for some gender primes. The impact of gender primes on math test performance appears to be sensitive to the exact phrasing used in the primes.

[1] Comments welcome and can be directed to twei@fas.harvard.edu. I thank Joan Stoeckel from the Educational Testing Service for valuable assistance regarding the NAEP data, as well as Susan Dynarski, Richard Freeman, Brian Jacob, Lawrence Katz, Todd Pittinsky, and seminar participants at Harvard for their support and invaluable comments on earlier drafts. All remaining errors are my own.

Research in psychology and behavioral economics emphasizes the importance of salience on actions and judgments (cf. Kahneman and Tversky, 1984). For instance, recency bias posits that because of mental availability, individuals overestimate the probability of a plane crash after they have seen one reported in the news. Increasing awareness of gender stereotypes in math may similarly prime students' perception and performance. More concretely, societal norms may implicitly or explicitly suggest that math is more suitable for boys. This norm, reinforced through peer interactions, parental attitudes, teacher expectations, and other priming channels, is hypothesized to negatively affect girls' subsequent performance and ultimately reinforce their own stereotypes and behaviors, as well as society's.[2] The result is a self-fulfilling prophecy. This phenomenon is known as "stereotype threat," or "being at risk of confirming, as self-characteristic, a negative stereotype about one's group" (Steele and Aronson, 1995).

Economists have explored many hypotheses for the observed gender gap in labor market outcomes (Goldin, 1994; Altonji and Blank, 1998; Blau and Kahn, 2000), the most common being occupational segregation (MacPherson and Hirsch, 1995). Divergent incentives to invest in human capital as a result of discrimination (Royalty, 1996) and differences in the division of labor (Becker, 1985) are often used to explain why women and men select into different jobs. However, stereotypes about what women can and should do in the labor market can influence whether employers choose to statistically discriminate and whether women choose to train in certain areas. This is particularly relevant in the male-dominated fields of math and science, which generally require more education and pay higher wages. Some psychology experiments

[2] For example, recent work by economists suggests that teachers' subjective assessments of students are influenced by their own biases (cf. Lavy, 2004; Burgess and Greaves, 2009; Hanna and Linden, 2009; Mechtenberg, 2009). Psychologists have also argued that stereotypes can influence women's desire to study math and science, which subsequently affects their career decisions (cf. Spencer, Steele, and Quinn, 1999). However, further empirical research on stereotype threat and the other theories to account for the gender gap in math is warranted, especially given new evidence that the gap arises at a much earlier age than previously thought (Fryer and Levitt, 2008).

have shown that stereotype threat can reduce women's desire to study quantitative fields (Davies et al, 2002; Gupta and Bhawe, 2007). Economists recently began experimentally documenting gender differences in preferences (Croson and Gneezy, 2009), such as for competition (Gneezy, Niederle, and Rustichini, 2003; Niederle and Vesterlund, 2007), but have done less research on stereotype threat and its possible impact on preferences.[3]

Most of the work on stereotype threat has been conducted in lab settings with small and highly selected samples, potentially raising concerns about external validity. On the other hand, the small number of observational studies with large and representative samples typically lack exogenous variation in exposure to primes, which leads to concerns about drawing causal inferences since other confounding factors are likely to be correlated with exposure to primes.

I address the limitations of these approaches by analyzing the micro data from the long-term trend National Assessment of Educational Progress (NAEP), a large and representative test of 9, 13, and 17 year-old U.S. students spanning 22 years. Students are randomly assigned to test blocks, some of which include gender primes while others do not. Comparing the gender gap in math test scores of students receiving gender primes to those receiving placebo questions reveals any possible priming effects on gender gaps in math. Thus, these data provide an unusual natural field experiment, which sheds light on the relevance of stereotype threat outside of the lab.

I find little evidence of traditional stereotype threat, but instead find null effects with some primes and reactance effects with other primes, whereby *girls perform significantly better* relative to boys. In the latter case, the hypothesized mechanism is that explicitly increasing the salience of gender stereotypes induces additional effort from girls to "prove everyone wrong." This accords with evidence that girls more intensely reject gender-math stereotypes than boys in the NAEP as well as with prior studies documenting stereotype reactance (cf. Kray et al, 2004).

[3] See Fryer, Levitt, and List (2008) for a recent analysis of stereotype threat by economists.

This result is robustly observed in most age cohorts and sample years between 1978 and 1999. Although smaller than educational policies such as class-size reductions (Krueger, 2003), the estimated reactance effect size on test scores is a relevant 0.05 test score standard deviations.

These results do not necessarily reject the existence of stereotype threat, but rather suggest that it is complex and mediated by many factors. For example, one possible reason for no stereotype threat in the NAEP is that labs may create relatively more performance anxiety due to increased salience of scrutiny and reduced perception of anonymity. Another explanation is that the NAEP primes are framed differently from most prior lab studies. The phrasing of primes in the NAEP correlates with whether we observe reactance or null effects; moreover, a recent lab experiment using the same primes as the NAEP replicates this pattern of effects (Wei, 2009). These findings suggest that priming can have tangible impacts on performance and that stereotype language may be used to partially *mitigate* gender gaps in math test scores.

The remainder of the paper proceeds as follows: section I summarizes the related literature and provides an informal framework to reconcile differing results from lab and field settings, section II presents my empirical strategy, section III describes the NAEP data structure and documents basic trends in attitudes about gender stereotypes, section IV tests for stereotype threat, section V performs robustness checks of the main results, section VI discusses possible mechanisms for the findings and then concludes.

I. Stereotype Threat Theory and Evidence: Lab versus Field

The stereotype threat literature is vast (over 300 papers in the past 15 years), so I review only a small subset here. The seminal study is by Steele and Aronson (1995), who consider verbal test score differences between black and white undergraduates. A series of four experiments show that consistent with stereotype threat, race-primed blacks perform worse

relative to whites than non-race-primed blacks after controlling for baseline SAT scores. Many follow-up studies have replicated this result in other tasks and identity domains, such as gender, socio-economic status, and age with non-trivial effect sizes (c.f. Levy, 1996; Steele, 1997; Aronson, Quinn, and Spencer, 1998; Croizet and Claire, 1998; Spencer, Steele, and Quinn, 1999; Blascovich et al, 2001; Good, Aronson, and Inzlicht, 2003; Johns, Schmader, and Martens, 2005). Thus, it appears that stereotype threat exists and can exacerbate the typical test score gap expected from the main culprits, such as school or neighborhood quality and income.

This consensus exists despite the fact that most of the aforementioned studies were conducted in labs with small and highly selected convenience samples (usually 100 students from competitive universities, who are tested on 20 SAT questions). However, Levitt and List (2007) argue that for several plausible reasons, behaviors may diverge in field and lab settings depending on scrutiny levels, participant self-selection, game stakes, and task framing.

Labs feature more researcher scrutiny and less participant anonymity. Participants know that a hypothesis is being tested and that their responses to specific stimuli are being carefully evaluated among a small group of peers.[4] With stereotype threat, increased scrutiny may heighten anxiety and self-consciousness, leading to a greater propensity for threat. Moreover, "experiment-friendly" individuals wishing to please researchers are more likely to volunteer for studies (Rosenthal and Rosnow, 1969; List, 2006). Experimenter demand for finding stereotype threat may thus more likely be met in the lab.

A potential boundary condition for stereotype threat is that participants are "identified with the domain" – in this case, that female participants care about performing well in math

[4] Several experimental economics studies confirm that differing scrutiny levels can lead to differing results in games measuring social preferences (cf. Andreoni and Bernheim, 2009). In many experiments, actual anonymity can be preserved with double-blind procedures; however, *perceived* anonymity may still be low – especially with smaller groups – which may distort behavior.

(Aronson et al, 1999). Consequently, most labs recruit participants who satisfy this criterion. Looser sample selection practices occur in natural experiments, precisely so that theories can be tested for generalizability to broader populations. Depending on how lower domain-identified participants behave under stereotype threat and their prevalence in the sample, aggregate results from field experiments could yield null or even reactance findings.[5]

Labs are low stakes settings given that the experiment does not involve large sums of money, nor do behaviors generate major consequences outside of the experiment. It is not obvious how varying stakes may impact stereotype threat. On the one hand, raising the ante may increase anxiety and pressure, thus magnifying threat. On the other hand, the importance of high stakes tests may dominate potential second-order threat effects. A few experiments have found inconsistent stereotype threat effects with monetary incentives (cf. McFarland, Lev-Arey, and Ziegert, 2003; Fryer, Levitt, and List, 2008). Field settings involve tests with varying stakes for both students and schools. For example, Advanced Placement exams are high stakes for college-bound students but not particularly for schools. Tests based on *No Child Left Behind* can be high stakes for students, teachers, and school districts. Some state accountability tests are used to evaluate schools but are low stakes for students. Similar to most lab experiments, national diagnostic tests such as the NAEP are low stakes for all.

Finally, as demonstrated in experimental economics studies, subtly altering the manner in which tasks are framed – for example, describing games as "community" or "Wall Street" games and characterizing fellow participants as "opponents" or "partners" – can substantially affect

[5] For example, girls who care less about doing well in math might nonetheless reject the stereotype that math is more for boys. Thus, when primed, they may exert more effort to disprove the stereotypical notion and succeed because they do not face the same pressure and anxiety to perform that highly domain-identified girls face. Alternatively, girls who do not identify with the domain may be unaffected by gender primes. Under either assumption, lab experiments, other things equal, provide an upper bound estimate of stereotype threat. Field studies, in contrast, provide a weighted estimate for the effects of highly and lower domain-identified participants. These heterogeneity issues are explored empirically in section IV.

outcomes (cf. Roth, 1995; Ross and Ward, 1996; Burnham, McCabe, and Smith, 2000; Bohnet and Cooter, 2005). In a meta-analysis of the experimental stereotype threat literature, Wei (2009) finds that primes worded in a stereotypical manner (i.e., "boys are better at math than girls" as opposed to "girls are better at math than boys") predict stereotype threat except when these primes are also self-affirming (i.e., soliciting a participant's opinion such as "do you think boys are better at math," as opposed to making a claim such as "it is well-documented that boys are better at math"), in which case *reactance* is more likely. Since most girls strongly reject gender stereotypes regarding math (see section III), allowing them to reaffirm this notion empowers them to disconfirm the stereotype – akin to a "saying-is-believing" effect (Higgins and Rholes, 1978). Primes in the experimental literature are typically not self-affirming, while the NAEP primes are self-affirming. Thus, all other things equal, we should expect to find less stereotype threat and more reactance in the NAEP.

This framework implies that depending on the specific parameters of the field experiment and assumptions about stereotype threat, diverging predictions can be made about how field and lab results will differ. It thus reduces to an empirical question. Nevertheless, it appears that the NAEP natural experiment should be less likely to find stereotype threat and more likely to find reactance. Ultimately, differing results may reflect the confluence of many differing factors. Isolating one precise mechanism is challenging absent a series of nested field and lab studies, which is the impetus for complementing the NAEP field evidence in this paper with the lab experiments conducted in Wei (2009).

This does not devalue the contribution of lab studies but suggests that stereotype theories should be tested in different settings. Several studies have moved in this direction by examining stereotype threat in classroom environments with elementary and high schoolers instead of

college students (Keller, 2002; Keller and Dauenheimer, 2003; Huguet and Regner, 2007; Neuville and Croizet, 2007). The results verify the existence of stereotype threat and accord with studies suggesting that stereotype threat varies by age (Ambady et al, 2001). The concern, however, is whether these studies are true field experiments. Although conducted in classrooms, students are typically aware that an experiment is taking place. Moreover, these studies have used small samples of respondents, usually of French or German nationality.

There are a few studies that use large and representative data to test for stereotype threat. A recent study at GRE test centers shows no evidence that test proctors' gender and race have impacts on performance (Walters, Lee, and Trapani, 2004). Another set of studies uses SAT and Armed Services Vocational Aptitude Battery tests to see if the data accord with what stereotype threat theory predicts (Cullen, Hardison, and Sackett, 2004; Cullen, Waters, and Sackett, 2006). Once again, there is little evidence of traditional stereotype threat; however, these findings should be interpreted with caution since they rely purely on correlational analysis.

The only large-scale randomized stereotype threat field study I am aware of was conducted under the auspices of the College Board, who randomly assigned tests, some of which contained questions about race and gender beforehand, to actual Advanced Placement (AP) calculus and Computerized Placement Test examinees (Stricker and Ward, 2004). The sample included approximately 1,500 students for each test in 1996. The study finds no test differences between identity primed and non-identity primed individuals, although these results have been subject to some debate regarding what constitutes statistical significance (Danaher and Crandall, 2008). The AP study differs from my study on several dimensions, such as the types of gender primes given (AP tests ask students to indicate their race or gender while the NAEP asks students about gender stereotypes), the stakes (AP tests are high stakes for students), the sample (AP test

takers are a fundamentally different population from the nationally representative sample of 9, 13, and 17 year-olds in the NAEP), and the time frame (the NAEP spans 22 years).

In sum, there is strong evidence from lab experiments that stereotype threat exists and has relevant implications for achievement. At the same time, despite clear theoretical reasons for why lab and field settings may generate divergent results, there is a relative paucity of convincing and comprehensive field research, primarily due to methodological challenges and disciplinary preferences. The few existing field studies find limited evidence for stereotype threat, but further work is needed to draw more definitive conclusions.

II. An Empirical Framework to Estimate Stereotype Threat

I use large and nationally representative data that spans over two decades and three age groups. At the same time, I retain the scientific rigor of the other lab studies by exploiting a design quirk in the NAEP that generates a natural experiment with some students randomly receiving gender primes and others not. The micro data, from the long-term trend NAEP math test, is administered every two or four years beginning in the 1970s to students ages 9, 13, and 17 in schools throughout the U.S.

The NAEP contains many test blocks, and although ultimately assessed on comparable material, students are randomly assigned to a fraction of the test blocks to minimize test burden.[6] Some of the test blocks include pre-test background questions, with the type of questions varying considerably across blocks. The background questions are pre-determined but arbitrarily

[6] The NAEP employs a multi-stage sampling design by first selecting PSUs, then schools, and finally students. For each school, eligible students are randomly assigned to take a particular test block. The use of an audio-paced tape format necessitates that all students taking a particular block be in the same room. Because administrators sought to have a reasonable number of students in each testing room, a few of the smaller schools had only one or two test blocks assigned. However, the vast majority of schools were large enough to accommodate multiple test block sessions (Allen, Carlson, and Zelenak, 1999). Thus, the randomization is virtually at the student level – in fact, one of the stated objectives is to give each student an equal chance of being assigned to a test block regardless of the number of blocks administered in his or her school. See section IIIc for randomization checks.

assigned to test blocks. My key identifying assumption is that if a set of background questions contains at least one question about an identity domain of interest such as gender, then that question plausibly acts as an identity prime. Students who receive these blocks form the treatment group, while students who receive test blocks with no background questions or with non-identity priming questions form the control group. Since test blocks are randomly assigned to students, this identifies the effect of primes on math performance.

Test blocks fall into three mutually exclusive and collectively exhaustive categories: no pre-test, placebo, or prime questions. Table 1 summarizes the test blocks for each sample year and age cohort. For example, the age 9 assessment in 1982 includes six test blocks: two blocks with no pre-test questions, three blocks with placebo questions, and one block with prime questions. A "xx" indicates that there are two gender prime questions in that block instead of one. There are eight years – 1978, 1982, 1986, 1990, 1992, 1994, 1996, and 1999. The 1986 to 1999 samples for 9 year-olds do not contain priming questions but do have placebo and no pre-test questions. I include them in the estimation because they provide information on error variance, even if they cannot help identify point estimates of the priming effect.

It is important to properly categorize each test block since prior evidence suggests that gender is not the only identity domain that can have a differential impact on performance. Placebo questions do not relate to known stereotype threat identity domains such as race, age, or socio-economic status. Instead, they generally inquire about opinions regarding math, as well as about math instruction (i.e., how often a math textbook is used, or how often math homework is assigned). Appendix Table 1 provides examples of the types of questions found in the blocks. There is no way to definitively rule out that the placebo questions induce a priming effect, but one test is to compare placebo blocks with no pre-test question blocks. If there are insignificant

performance differences, then it is more plausible that the placebo question blocks serve as valid control groups. This test is implemented in section V.

There are three gender prime questions in the NAEP math assessment. One question is: "How do you feel about this statement: *math is more for boys than girls*. Do you strongly disagree, disagree, undecided, agree, or strongly agree?" There is also a similarly formatted question that reverses the "boy-girl" ordering above, as well as a question asking for gender comparisons with respect to logical ability. These Likert scale questions are useful because they measure explicit attitudes and serve as gender primes. For notational simplicity, I use the following shorthand to describe the primes for the remainder of the paper:

1) "math more for boys" = do you feel that math is more for boys than girls?
2) "math more for girls" = do you feel that math is more for girls than boys?
3) "girls more logical" = do you feel that fewer men have logical ability than women?

The strategy for estimating stereotype threat is differences-in-differences (*DID*):[7]

$$DID = (\overline{boys}_t - \overline{girls}_t) - (\overline{boys}_c - \overline{girls}_c) \qquad [1]$$

where *t* and *c* indexes individuals in the treatment and control groups, respectively; $\overline{boys}$ and $\overline{girls}$ indicate the test score group means for boys and girls, respectively. Although the test blocks assess comparable material and are of comparable difficulty, they do not contain the exact same set of questions (see Allen, Carlson, and Zelenak (1999) for a more detailed discussion of exam content).[8] This is the motivation for taking double differences rather than simply comparing girls' test score levels between control and treatment groups.[9]

[7] Another possible approach would be to look within students who received multiple test blocks to see if there are performance differences between blocks with and without the prime. Unfortunately, few students received booklets with multiple math blocks, and there was no variation in gender primes between the blocks for those who did.
[8] For example, some blocks may have questions requiring calculator usage or contain a few more questions.
[9] A key identifying assumption is that the *DID* estimate is zero for the treatment and control blocks when the prime questions are removed from the treatment blocks. This condition could be directly tested if there are years that have the same test blocks but without any primes, but unfortunately such years do not exist in the NAEP. Nevertheless, for this to be a problem, the treatment and control blocks would not only have to be different but would also have to

The *DID* approach in equation 1 can be implemented with linear regressions. This has the added advantage of allowing one to control for other variables, although if randomization is successful, this should only affect the precision and not the consistency of the point estimates. Furthermore, since there are many sample years and ages, a regression framework allows for a simple pooling of the many possible comparisons into an aggregate estimate of stereotype threat and for a simple simultaneous inference test. The linear model takes the following form:

$$score_i = \beta_1 * female_i + \beta_2 * treat_i + \beta_3 * (female \times treat)_i + X_i'\gamma + d_{ay} + \varepsilon_i \qquad [2]$$

where *i* indexes the individual, *a* indexes the age cohort, *y* indexes the sample year, and ε is the error term. *score* is the test score, *female* indicates if a student is female, *d* is a set of age by year dummies, and *X* are individual-specific covariates such as race, parental education, grade level, geographic region, and private school attendance. *treat* is a dummy indicating if a student is in the treatment (received a gender prime) or control group (received placebo questions or no background questions). β_1 and β_2 estimate the gender and treatment-control group test score gaps, respectively. The interaction term coefficient, β_3, is the *DID* estimator, although it does not take the double-difference form when covariates are included. Stereotype threat predicts that girls' test scores are relatively more depressed in the treatment group. Thus, a negative estimate for β_3 indicates stereotype threat, while a positive estimate suggests stereotype reactance.

Since the NAEP contains multiple gender primes, it is of interest to examine whether stereotype threat effects differ by type of prime. Equation 2 can be trivially augmented:

$$\begin{aligned} score_i = \; & \beta_1 * female_i + \beta_{2A} * treatA_i + \beta_{2B} * treatB_i + \beta_{3A} * (female \times treatA)_i \\ & + \beta_{3B} * (female \times treatB)_i + X_i'\gamma + d_{ay} + \varepsilon_i \end{aligned} \qquad [3]$$

---

be different in a way that differentially affects boys and girls. This seems less likely but is a potential concern addressed at more length below and in section V.

where the *treat* indicator is decomposed to indicate whether an individual received prime A (*treatA*) or prime B (*treatB*). The untreated control group includes students receiving placebo questions or no background questions. β_{3A} and β_{3B} are the *DID* estimates relative to the non-primed controls for those receiving prime A and prime B, respectively. Basic inference techniques can be applied to determine whether either or both treatment primes are statistically significant and whether the effects of each prime significantly differ from each other.

Since the gender gap may vary by year and age even in the absence of primes, I also estimate equations 2 and 3 fully interacting the age-year dummies (d_{ay}) with the female dummy (*female*), treatment dummy (*treat*), and all standard covariates (*X*). A final issue is how standard errors are estimated. Although the NAEP designers aim to make test blocks comparably difficult even though each block contains different questions, they may not succeed in actuality. Given my estimation strategy, this could be problematic if the difficulty of test questions found in some test blocks but not others varies across genders. To adjust for this uncertainty, I cluster standard errors by age-year-block in all my analyses, since these are the cells to which students are assigned. One might also cluster the standard errors by primary sampling unit (PSU), since the multi-stage NAEP design utilizes PSUs to form nationally representative samples. Though students within PSUs come from schools within the same geographic areas, the large number of PSUs (over a hundred per age-year sample) relative to test blocks suggests that the former is less likely to problematically exhibit serially correlated covariates within clusters. Nevertheless, as a robustness check, I report estimates from both types of clustering in the main analysis.

III. The Long-Term Trend NAEP

A. Data Structure

Commissioned by the U.S. Department of Education, the long-term trend NAEP is a representative, repeated cross-section of U.S. students ages 9, 13, and 17.[10] Since the intent of the assessment was to track student trends in reading and math, test scores are comparable over time; in other words, within each age cohort, the same (or similar) exam is given every two or four years.[11] Unlike the aggregate statistics publicly available on the National Center for Education Statistics website, the micro data are restricted use since they contain student-level demographics and responses to background and test questions from 1978 to 1999 (eight sample years). Each subject-age-year sample contains several thousand students.

Three characteristics of the NAEP are worth mentioning. First, the assessment is given in the student's own school during regular hours, typically by a teacher, substitute teacher, or other local individual. The math assessment is audio-paced to eliminate student reading ability effects. Only cognitive test questions are paced. Although prior stereotype threat studies have used self-paced exams, recent analysis with the NAEP shows virtually no differences by administration method (Perie, Moran, and Lutkus, 2005).

Second, every test booklet begins with a common set of background questions. Although some of these questions may be subtle identity primes (i.e., asking for race or parent education), they are given to all respondents and so should not differentially affect the treatment and control groups. This common set of questions was introduced in 1986, and as seen in section IV, similar results are found in the pre- and post-1986 samples. Also, there is a clear demarcation between the initial set of background questions and the start of a test block, so that students may distinctly compartmentalize them. Finally, to the extent that these subtle questions serve as a prime, they

[10] The NAEP is representative of school children, which with compulsory attendance laws is roughly equivalent to all 9 and 13 year-olds. However, the age 17 samples exclude high school dropouts. The NAEP also oversamples minority populations to ensure sufficient representation from those groups, thus sampling weights are necessary.
[11] It would be interesting to perform the same analysis with reading test scores since the stereotype prediction is reversed for girls and boys. Definitive and randomly distributed gender primes are lacking, however.

should be superseded by the placebo or gender prime questions that appear later on, right before the test questions begin. Note that the respondent's gender is not explicitly asked during the exam (it is either taken from school records or imputed).

Third, starting in 1986, each test contains a math, science, and reading block for 9 and 13 year-olds, but just math and science blocks for 17 year-olds. The reading blocks are not actually scored and are unrelated to the NAEP reading test. Although this may lead to contextual subject effects, I argue that they are most likely minimal. As mentioned, each section of the exam is clearly separated, especially since the math portion is audio-paced, and breaks are given between sections. It thus seems reasonable to think of these sections as separate mini-tests from the students' perspective. More definitively, the consistency of the results across time, ages, and comparison groups, suggests that the booklet orderings are unrelated to the estimates.

Each student is assigned five "plausible values" for the test score given that the exam is too short to precisely yield a single score.[12] The plausible values come from Item-Response Theory, which probabilistically imputes student performance on questions he or she was not given, based on similar students' performance on those questions (Allen, McClellan, and Stoeckel, 2005). In this study, I do not want a proficiency score that is partially determined by how other students responded to similar questions from other blocks, as this would undermine the research design. Therefore, I construct "blockscores," which are the proportion correctly answered by a student in the test block. Although these blockscores may not capture true proficiency, they are based on at least 25 questions, which to the best of my knowledge matches or exceeds the number of questions asked in most prior studies. Also, there is a strong positive correlation between each student's blockscore and imputed proficiency score. These correlations

[12] A block spiral test design is utilized, whereby the exam is partitioned into balanced blocks containing different background and cognitive questions. Each block is then assigned to a student. The implication is that no student receives all the questions on the entire exam, but all questions are still administered in roughly equal proportion.

range from 0.6 to 0.8 and are all highly significant. Mean blockscores by year and age range from 50 to 67 percent, with standard deviations between 16 and 22 percent.

B. Gender Attitudes

Figures 1 and 2 summarize student responses to the gender perception questions. The data are pooled over all sample years (1978-1999) but note that perception questions were not given in some ages and years. Figure 1 shows the "math more for boys" responses by age and gender. A majority (60 to 90 percent) of respondents disagree with the stereotype. Within gender, 13 and 17 year-olds do not markedly differ; but only 60 to 75 percent of 9 year-olds disagree (90 percent in older cohorts), while 10 to 20 percent agree (5 percent in older cohorts).[13]

Differences in attitudes exist between boys and girls regardless of age. Girls reject the "math more for boys" stereotype more strongly than boys do. Amongst 13 and 17 year-olds, more girls strongly disagree while more boys are undecided. Figure 2 shows that 9 year-old girls and boys equally reject the idea that math is more for girls. Interestingly, for 13 and 17 year-olds, more girls strongly disagree than boys (60 percent versus 45 to 55 percent). In fact, more boys actually agree with the "math more for girls" stereotype.

In sum, most students explicitly eschew all stereotypes about gender and math. However, some stereotypes still exist and persist. 9 year-olds are more prone to having stereotypes, whereas attitudes are less malleable between ages 13 and 17. Finally, there are gender differences, most notably that girls appear to more strongly reject all stereotypes.[14]

---

[13] When disaggregated by year, it appears that although there is some variation over time, there does not appear to be a trend. Any variation is driven by respondents switching between intensities of agreement or disagreement. I also examined birth cohorts (age 9, 13, and 17 profiles for those born in 1969, 1973, 1977, and 1979), but there is no apparent time trend.

[14] I also explored attitudes by race, socio-economic status as proxied by parental education, and geographic region. Selected results are presented in Appendix Figure 1 for the "math more for boys" question. In general, the aggregate sample findings also hold for each subgroup. However, blacks and Hispanics are less likely to strongly reject gender stereotypes about math than whites and Asians. This is also true for students with low-educated parents

C. Randomization Checks

Before proceeding with the *DID* estimation, I first verify that the treatment and control groups are balanced over baseline characteristics such as grade, sex, race, parental education, private/public school attendance, region of residence, and household reading materials. The group means and F-statistics for differences in means are given in Table 2, by age and year. I pool the sample years from 1986 to 1999 for the sake of compactness, and since the test booklets and sampling procedures are quite uniform during those years.

In general, the randomization between test blocks appears successful. The two exceptions are 9 year-olds in 1982 and 17 year-olds in 1978. These are driven by the proportion of private school and "other race" students, respectively; when the relevant variable is excluded, the F-tests are insignificant (see the notes in Table 2).[15,16] Of the 152 variables considered (19 for each of the eight age-year groups), 14 were individually significant. Eight of the 14 variables came from the two problematic age-year groups mentioned above. To be cautious, I report the *DID* estimates both with and without controlling for these variables.

IV. The Effect of Stereotype Threat on Math Performance

A. Main Results

---

relative to those with high-educated parents. Regional differences are less pronounced, although Northeast students more strongly reject gender stereotypes while Southeast students are relatively less likely to. The gaps are stronger among girls and younger age cohorts. There are many explanations for these interesting observations, and although identifying the proper ones is a worthy exercise, it is beyond the scope of this paper.

[15] When I run the randomization checks separately by year from 1986 to 1999, all but three sample years yield insignificant joint F-tests. As with the two sample years mentioned above, each has one or two anomalous variables, which yield insignificant joint F-tests when removed. Some of this is explained by the complex sampling design of the NAEP described in section II, which basically randomizes test blocks at the student level, but imperfectly so.

[16] Since my analysis hinges on comparing gender gaps across control and treatment groups, I also conduct these randomization checks separately by gender. The results are similar to the full sample, with mostly insignificant joint F-tests. The only exception is that 17 year-old girls have more issues than 17 year-old boys for a few years, although as was the case before, these issues were driven by one or two anomalous variables, with no obvious perennial culprits.

Table 3 reports *DID* estimates from equations 2 and 3 for the full sample.[17] Each column in each panel corresponds to a separate regression depending on whether the primes are combined (panel A, equation 2) or separated (panel B, equation 3), and whether fully interacted age-year dummies (columns 1 to 4), sampling weights (columns 1, 3, and 5), and/or covariates such as region, race, parent education, grade, and private school attendance (columns 1, 2, and 5) are included. Only the coefficient on the female-treatment interaction term – the *DID* priming estimate (β_3 in equations 2 and 3) – is reported. The estimates, when multiplied by 100, are test score percentage points. Two sets of standard errors are reported below the *DID* point estimates: parentheses for age-year-block clusters and brackets for PSU clusters.

The results are generally robust across these various specifications. The female main effect (not reported) is a significant -0.02, suggesting that girls score about two percentage points lower than boys in this sample. The pooled *DID* estimate of the effect of any gender prime on test scores (panel A) is positive (0.002 to 0.003) but generally insignificant, providing little evidence of traditional stereotype threat. When I split the estimates by type of gender prime (panel B), a different pattern emerges. The pooled *DID* estimate for those receiving the "math more for boys" prime is positive and highly significant (0.006 to 0.009), suggesting stereotype reactance. For those receiving the "math more for girls" prime, the estimate is negative and mostly insignificant (-0.002 to -0.007), suggesting no stereotype effect. Test statistics from an F-test comparing the "math more for boys" and "math more for girls" *DID* estimates using both age-year-block and PSU clusters are reported below panel B. Across all specifications, the differences between the two primes are significant, which is consistent with the theoretical

[17] In this and the ensuing analysis, no obvious differences were found based on whether the treatment group had one prime or two primes. I thus report results that do not distinguish between treatment groups with one or two primes. In particular, the "girls more logical" prime only shows up in a few sample years and is always bundled with the "math more for girls" prime. I combine these groups with the ones only receiving the "math more for girls" prime, which is sensible given the framework in section I – namely, that both primes are counter-stereotypically framed.

framework in section I.[18] Not all primes are created equal, and pooling the primes thus produces insignificant overall effects.[19]

Table 4 presents analogous regression results by age. For each age cohort, *DID* estimates are presented for "any prime" (panel A) and separately by type of prime (panel B), both with and without standard covariates pooling across years. All specifications include sampling weights and year dummies that are fully interacted with all covariates except for the female-treatment interaction term. Standard errors are clustered by age-year-block. The effects are insignificant for 9 year-olds; however, this is based on just two sample years (1978 and 1982) since the "math more for boys" question was the only gender prime in those years. 13 and 17 year-olds have complete data and look similar to the aggregate samples in Table 3: the effect of any prime is positive but not strongly significant, but the "math more for boys" prime is positive and significant while the "math more for girls" prime is negative and insignificant. F-tests of differences in these primes are highly significant across all specifications. Although both 13 and 17 year-olds exhibit similar patterns, the "math more for boys" reactance effect appears strongest for 13 year-olds (*DID* estimates of about 0.009 versus 0.007 for 17 year-olds). As a reference point, the female main effect is about -0.02 for 13 year-olds and -0.04 for 17 year-olds (both are significant), indicating that the gender gap in math favors boys and widens with age. In sum, there is little evidence of traditional stereotype threat in the NAEP but considerable evidence of stereotype reactance for some primes. This finding implies that stereotype language actually may be able to reduce the test score gender gap that exists under normal conditions.

[18] As an additional robustness check, I ran the regressions with a full set of test block fixed effects. Since the test blocks are well balanced by gender, it is not surprising that the results are quite similar to the results from Table 3: the *DID* estimate of any prime is an insignificant 0.0014, while the effect is a significant 0.0071 for "math more for boys" and an insignificant -0.0019 for "math more for girls" primes.

[19] Table 3 reports full sample results, with control groups receiving either no pre-test questions or placebo pre-test questions. Appendix Table 2 shows analogous regressions but instead defines the control groups as only those receiving placebo pre-test questions. Although the reactance effects appear to be somewhat stronger, the results are quite similar to Table 3 overall, thus I include the full set of control groups for the remaining analyses.

B. Effect Sizes

The upper bound *DID* estimate for the "math more for boys" prime is about 0.9 test score percentage points, which with a pooled standard deviation of about 18 percent, yields an effect size of 0.05 test score standard deviations. Although not nearly as large as the 0.20 standard deviation estimates from class-size experiments (Krueger, 2003), this effect is non-negligible.

To gain a better sense of the magnitudes, I estimate the average test score gender gap for the full sample of control groups, including the same covariates in Table 3 and applying student weights (I also conducted this analysis without covariates and/or weights and obtain similar results). The gap is 1.9 test score percentage points in favor of boys (for example, boys averaged 60 percent answers correct, while girls averaged 58.1 percent answers correct). With a pooled *DID* estimate of 0.9 percentage points, this implies that stereotype reactance reduces the gap to 1.0 test score percentage points, or by 47 percent of the original gap.

The gender gap widens for older age groups, so it is useful to examine the reactance magnitudes by age. I focus on 13 and 17 year-olds since each of these groups has complete data for the "math more for boys" prime (all eight sample years). For 13 year-olds, the estimated gap is 1.2 test score percentage points in favor of boys. From Table 4, the "math more for boys" pooled *DID* estimate for 13 year-olds is 0.9 percentage points, which implies that stereotype reactance reduces the gap to 0.3 test score percentage points, or by 75 percent of the original gap. For 17 year-olds, the estimated gap is 3.9 percentage points in favor of boys. The "math more for boys" pooled *DID* estimate for 17 year-olds is 0.7 percentage points, which implies that the gap is reduced to 3.2 percentage points, or by 18 percent of the original gap.

Thus, the effect appears to be strongest for 13 year-olds. Nevertheless, to the extent existing gender gaps are of practical magnitudes for both age cohorts, the stereotype reactance

effect sizes are non-trivial. They highlight the importance of framing effects in that subtly manipulating stereotype language can actually reduce the test score gender gap.

C. Distributional Differences

There are several approaches to examine whether one segment of the ability distribution is driving the stereotype reactance effects, but the preferred method is to estimate quantile regressions from the test score distribution.[20] Instead of OLS, I estimate conditional quantile functions of the same form as equations 2 and 3 for 21 quantiles ranging from 0.05 to 0.95 in increments of 0.05, in addition to the 0.01 and 0.99 quantiles. I approximate standard errors for the point estimates using the bootstrap method with 1,000 replications clustered by age-year-block.[21] Figures 3 and 4 plot as a dark solid line the quantile regression coefficients on the female-treatment interaction term (*DID* estimate of stereotype threat) for the 21 quantiles labeled on the horizontal axis. The dotted lines show 95% confidence intervals, and for reference, the dashed line gives the mean OLS treatment effects presented in section IVa.

[20] Another method is to take the five NAEP plausible test score values and divide the sample into deciles using the full test score distribution for each age cohort and year; then estimate the standard *DID* model for each decile, which yields ten estimates that when plotted against decile, illustrates distributional heterogeneity. One concern is that actual test score is an endogenous outcome potentially influenced by the treatment, so it would be problematic to select samples based on test scores. However, I do not use the blockscores (fraction correctly answered), rather I use the plausible values computed from Item-Response Theory. The methods are complicated but essentially impute scores for students based on how other students with similar background characteristics performed. Thus, the plausible values are not based solely on how an individual student answered test questions and so are less likely to be contaminated by treatment status. Nevertheless, I also perform this exercise in the following alternative way: using the control samples, I run a separate OLS regression of test scores (fraction correctly answered) on a set of pre-determined characteristics including gender, race, region, parental education, grade level, and private school attendance for each age and year. With the estimated coefficients, I predict student test scores based on background characteristics. Using the deciles from these predicted test scores, I can produce an analogous set of distributional comparison plots. This approach partitions students by how well we would expect them to do based on fixed characteristics, then compares that with how they actually perform to see if treatment effects differ on baseline expectations. Because both approaches rely on background characteristics to predict test scores, we expect similar results. This is qualitatively so, although estimates using the latter method tend to be more imprecise. The basic conclusions accord with the formal quantile estimates (that there may be some distributional heterogeneity, although no single segment dominates the stereotype reactance findings), with the exception that they miss the stereotype threat at the top of the distribution for students receiving the "math more for girls" prime and the strengthening of stereotype reactance at the top of the distribution for students receiving the "math more for boys" prime.

[21] This approach to measuring treatment heterogeneity is similar to that in Bitler, Gelbach, and Hoynes (2006).

Students in the NAEP are not explicitly selected to be highly identified with the math domain, as they typically are in lab experiments. One way to indirectly identify these students in the NAEP is to examine those at the top of the test score distribution. Such a proxy is standard practice in the psychological literature when explicit student attitudes about performing well in math are unavailable (cf. Steele, 1997; Cullen, Hardison, and Sackett, 2004).[22] If highly domain-identified students are most susceptible to stereotype threat, as the theory predicts (Aronson et al, 1999), then we should see negative estimates at upper quantiles.

Figure 3 shows quantile effects for the pooled sample receiving the "math more for boys" prime. The estimates are imprecise at the bottom of the distribution, but in general, significantly positive estimates (stereotype reactance) are present throughout most of the distribution with magnitudes similar to those estimated via OLS. The reactance effect appears to strengthen slightly (estimates become more positive) at the top of the distribution.

Figure 4 shows results for the pooled sample receiving the "math more for girls" prime. As with the OLS results, there is virtually no stereotype effect for most of the distribution. The OLS estimate obscures the significantly negative *DID* estimates in the 85th, 90th, and 95th quantiles however, consistent with the notion that highly identified students are most susceptible to stereotype threat. Studies have shown that although mean gender test scores differences in math are relatively small, the difference in variances is large, with significantly fewer girls in the upper tails (Hedges and Nowell, 1995). This is relevant to the extent that stereotype threat adversely impacts highly identified girls who are capable of high performance. This may

[22] Another way to measure highly identified students is to examine attitudes. The NAEP asks "how important or not is math?", "are you willing to work hard to do well in math?", and "do you really want to do well in math?" Casual analysis suggests that students who think math is more important, who are willing to work hard, and who really want to do well in math, tend to have higher test scores. I cannot directly use this information to estimate my *DID* models though because these questions were only presented to a fraction of students. Since responses are closely related to test scores, which all students have, I directly use test scores as the indicator of domain identification.

partially account for female under-representation in upper echelons of quantitative occupations if stereotype threat weakens potential top scorers' confidence so that they avoid such professions.

In sum, there is some evidence of traditional stereotype threat at the top of the test score distribution for those receiving the "math more for girls" prime. However, stereotype reactance is, if anything, stronger at the top of the distribution for those receiving the "math more for boys" prime. It is unclear what is driving these differences but the results rule out the hypothesis that reactance is driven by students at an extreme or narrow part of the distribution. They also suggest that insignificant stereotype threat in the NAEP overall is not solely due to sample selection differences with regard to highly domain-identified students.[23]

V. Robustness Checks

A. Leveling Up or Down?

I have shown that when treated with the "math more for boys" prime, the gender gap shrinks relative to with no primes or placebo primes. This narrowing of the gap can be due to boys performing worse, girls performing better, or some combination of the two. Moreover, this performance fluctuation can be driven by priming effects that differentially affect boys and girls, test block differences in difficulty that differentially affect boys and girls, or some combination of the two. In a perfect experiment, all groups would have received the same test questions, in which case we could directly compare girls (boys) in the treatment group with girls (boys) in the control group.[24] Absent this in the NAEP, concern about differences in test difficulty across

[23] I also conducted a subgroup analysis examining stereotype threat in groups based on region of the U.S., race, average school test score, school racial composition, and parental education level. The conclusion from this descriptive analysis is that stereotype reactance is robust across many groups, although there is heterogeneity in magnitudes. It would be interesting to see whether these differences are related to differences in stereotype attitudes across subgroups, as noted in section IIIb. However, explaining this heterogeneity is beyond the scope of this paper.

[24] One way to approximate this gold standard is to examine performance on questions appearing in both treatment and control groups. Unfortunately, none exist. A second option is to first code each question on content and difficulty level and then compare questions with similar codings. Unfortunately, many of the questions in the NAEP

blocks that are common to both genders motivates the use of differences-in-differences as the estimation strategy instead of just comparing group means. Even so, this identification strategy is potentially problematic if differences in test block difficulty differentially affect boys and girls. The limitations of the NAEP make it difficult to definitively resolve these issues; still, I provide heuristic evidence from the data and draw on recent experiments as evidence for robustness.

Figure 5 shows, using the full NAEP sample, student-weighted mean test scores (fraction of questions answered correctly) by control and type of treatment prime, separately for boys and girls. The gender gap in the control groups persists in the "math more for girls" prime (null stereotype threat); however, the gap shrinks with the "math more for boys" prime (stereotype reactance). This replicates the main differences-in-differences results from section IVa.

Boys and girls perform worse in absolute terms for all treatment primes relative to controls in the pooled sample, on average. Thus, one interpretation of the results is that this reflects a priming effect that adversely affects boys and girls alike. Although this hypothesis is difficult to rule out mechanically, it seems harder to justify on theoretical grounds. It is not obvious why the priming effect would be so much stronger for both boys and girls under the "math more for girls" treatment prime relative to the "math more for boys" prime. Moreover, these graphs reflect the full NAEP sample averages. When disaggregated by age, we see a similar pattern for 13 year-olds; however, for 17 year-olds, the treatment groups actually have higher mean performance relative to controls. In other words, both girls and boys improve with the treatment prime, but girls' scores increase by more (stereotype reactance). Unless we believe that 17 year-olds are polar opposites of 13 year-olds, it seems unlikely that these results are purely explained by priming effects that equally affect both genders.

---

are confidential due to planned reuse on future tests. Moreover, many questions that are available are at best difficult to interpret (i.e., one question is: "X times 1 = X true when any no. substituted for X"; another is: "estimate buy between 5 & 10 13 [Pencils for $1.00"), and so any coding, even if mechanically feasible, would be arbitrary.

Another explanation of the results is that test blocks with treatment primes happen to be harder than control test blocks in a way that differentially affects boys and girls. This seems unlikely, particularly for the "math more for boys" prime since the mean scores for the treatment and control blocks within each gender lie within one percentage point. This suggests that the test blocks may have slightly different degrees of difficulty but not substantially enough to explain the differential gender effects, especially since the test blocks are designed to be similar. Though still within a modest five percentage point range, the gaps between the controls and "math more for girls" prime are significantly larger. Nevertheless, if these results only reflect differential test block difficulty effects, then extremely difficult blocks should exacerbate the gender gap relative to easier blocks, rather than narrow it or leave it unchanged, as is the case here.

Furthermore, the results come from the full NAEP sample, which includes many pooled independent samples across three age cohorts and 22 years. It is possible that test designers failed to make all test blocks comparably difficult. However, given that test blocks vary across age and year, it is unlikely that the most difficult test blocks always happened to be assigned treatment primes. In fact, the 17 year-old sample of treatment blocks appear to actually be, on average, easier than the controls, whereas the opposite is true for 13 year-olds.

Figure 6 disaggregates the *DID* estimates from Table 3 by estimating equation 2 for each prime, age, and year, and then plotting the resulting *DID* estimates for the 13 and 17 year-olds, who have complete data. "Math more for boys" estimates are generally positive (stereotype reactance), while "math more for girls" estimates are close to zero or even slightly negative (null stereotype threat), consistent with the pooled estimates. The estimates have insignificant time trends, which is reassuring given that attitudes about gender stereotypes have not changed substantially over time (see section IIIb). The reactance findings for 17-year olds ("math more

for boys" prime) are attenuated for a few years. This is driven by differences in the number of test questions between treatment and control blocks, particularly in the earlier sample years when there are multiple control blocks. For example, one control block has 68 questions compared with 42 questions in the treatment block. The other control blocks in that age-year sample are within a few questions of the treatment. When such outliers are removed, the pooled estimate of reactance (from Table 3) increases from 0.009 to 0.011. Moreover, the attenuated reactance estimates in Figure 6 increase to magnitudes similar to the other age-year sample estimates.

That the main results are quite consistent even when disaggregated by age and year suggests that differential differences in test block difficulty are not responsible for the observed effects (the results are also robust to error clustering by age-year-block as discussed in section II and shown in Table 3). Instead, both treatment blocks tend to be slightly more challenging than the control blocks on average, but not in such an enormous way as to differentially affect the genders. Differences in the gender gaps between treatment and control groups thus robustly identify the effects of gender primes on math performance.

Nevertheless, the question of whether these primes induce boys to do worse or girls to do better remains. Given existing theory and evidence, I argue in section VI that it is girls doing better, but concede that this is not a question I can definitively answer with the NAEP data alone. However, the results from a recent experimental study of 187 undergraduates provides potent evidence on this issue (Wei, 2009). A math test containing 20 GRE multiple choice questions was administered to students randomly assigned to a control (no prime), a "math more for boys" prime, or a "math more for girls" prime group. The primes were worded exactly as they are in the NAEP. More importantly, because all students received the exact same test questions, mean scores can be directly compared in levels to identify priming effects.

The results are strikingly similar to the NAEP findings. Women receiving the "math more for boys" prime correctly answered an average of 2 more questions relative to control group women, from a baseline control mean of 8 (significant at the 5 percent level). Men receiving the "math more for boys" prime correctly answered 0.5 fewer questions relative to control men, from a baseline control mean of 11 (statistically insignificant). The priming effect thus reduced the original 3-question gender gap by about 50 percent, which is in the same ballpark as the NAEP findings in section IVb. Furthermore, the "math more for girls" prime reduced scores by a statistically insignificant 0.6 questions for women and 0.8 questions for men.

The experiment finds stereotype reactance for the "math more for boys" prime and no stereotype threat effects for the "math more for girls" prime. This effect is driven by women performing better and not by men performing worse. Although the sample and setting differ, the lab replication of the NAEP findings solidifies the claim that stereotype reactance is robust for certain types of primes and that this effect is due to girls leveling up, not boys leveling down.

B. Other Checks

One concern is if the placebo questions induce zero effects, which is necessary for these blocks to be valid controls. It is comforting that the effects with the no-question controls and with the placebo-question controls do not exhibit differential patterns (see Appendix Table 2). To be more rigorous, I estimate the *DID* models from equation 2, but instead use the placebo question blocks as the treatment and the no-question blocks as the control. I also compare no-question blocks with each other and placebo question blocks with each other. If nothing unusual is occurring, then the estimated *DID* effects should be small and insignificant.

The pooled *DID* estimates (analogous to those reported in Table 3) range from an insignificant 0.001 to 0.003 (with a standard error of approximately 0.006). Estimates

disaggregated by age and year also yield insignificant *DID* estimates.[25] This lends plausibility to the control groups and the subsequent stereotype threat estimates.[26]

VI. Discussion

This study aims to fill an important gap in the stereotype threat literature: namely, the limited number of comprehensive field tests. The majority of prior studies have been conducted in lab settings with small convenience samples of talented undergraduates. Almost all existing "field" studies are either correlational or unlikely to truly reflect natural settings (i.e., conducting tests in a classroom where students are aware of an on-going experiment). I instead quasi-experimentally test for gender-math stereotype threat in a natural setting by exploiting a quirk in the NAEP. Since this assessment has been given to large and nationally representative samples of students from three age cohorts over more than two decades, this allows me to explore stereotype threat in a more rigorous manner than was possible in prior studies.

There is little evidence of traditional stereotype threat in the NAEP. Instead, some primes ("math more for girls") yield null effects while other primes ("math more for boys") yield opposite stereotype reactance effects. This finding is robust across most age cohorts and years and with age-year-block clustering adjustments for test block differences. Heuristic evidence from the NAEP and experimental evidence replicating the results with the NAEP primes (Wei, 2009), confirm that the reactance is from girls performing better and not boys performing worse.

[25] The estimates are all from 1978 and 1982 because these two sample years have several control groups of both types (see Table 1).

[26] One last concern is what the questions on gender perceptions in math really measure, and whether students have an incentive to dishonestly report their preferences. While certainly possible, I argue that it is not a major problem. Since the NAEP assessments are paper-pencil and completely anonymous (i.e., scores are not reported to the student or school because the test does not assess individual proficiency and/or school quality), this blunts any incentive to be dishonest to conform to socially acceptable preferences. Even if it did, I am precisely interested in measuring the effect of socially-driven stereotypes on performance. Finally, with respect to testing stereotype threat, the actual responses do not matter since the "treatment" is simply a prime – all that is required is to get students to think about gender and thereby increase its salience.

The results do not necessarily reject the existence of stereotype threat; instead, they reinforce the idea that framing and priming effects can be important. This is consistent with what a thorough literature review suggests: stereotype threat exists but is complex and subject to many conditions (cf. Shih, Pittinsky, and Ambady, 1999; Dee, 2009). As if acknowledging this, the NAEP designers revised the recent 2004 test by not only trimming the number of non-cognitive background questions but by also combining the questions together in a single section administered at the end of the test.

The lack of stereotype threat is perhaps not too surprising. Laboratories may magnify stereotype decrements by creating more performance anxiety relative to field settings with increased salience of scrutiny and reduced perception of anonymity. Lab participants are also self-selected and may have a particularly stronger desire to satisfy experimenter demand effects.

Other explanations involve domain identification and task difficulty. Lab studies tend to recruit only high math domain-identified participants because this is thought to be a boundary condition. It is nonetheless valuable to understand whether stereotype threat generalizes to broader populations, as with the nationally representative NAEP samples. Still, when highly domain-identified individuals are proxied by students performing in the upper tail of the NAEP test score distribution, traditional stereotype threat is not consistently found. Another necessary condition for stereotype threat is that sufficiently challenging tests are given (Neuville and Croizet, 2007). While the NAEP may be less difficult than some tests given in lab experiments, it is certainly not easy, with test score averages around 60 percent.

A more plausible reason is differences in the framing of primes between the NAEP and most prior lab studies. The NAEP primes are more direct and open-ended: they solicit student opinions on gender stereotypes, rather than using the more typically neutral or rigid primes (i.e.,

indicating gender or claiming that boys outperform girls on math tests). Because the "math more for boys" prime is in the stereotypical direction and open-ended, it may increase the salience of the negative stereotype while empowering girls by allowing them to reject the stereotype.[27] This "saying-is-believing" effect (Higgins and Rholes, 1978) is consistent with the observation that girls tend to more strongly reject gender stereotypes about math than boys in the NAEP and that reactance occurs only when girls are challenged with the traditional stereotype ("math is more for boys").[28] A recent meta-analysis of the stereotype literature reveals that experiments using primes similar to the NAEP primes are much more likely to find reactance (Wei, 2009).

The NAEP findings suggest that stereotype language could shrink the gender gap on math tests. Low stakes exams, such as the NAEP, are thought to better capture human capital attainment; however, that these test scores can be relatively easily altered raises issues about what is actually being measured. With high stakes tests, it has been well-documented that schools (and teachers) face incentives to boost student test scores, for example by giving poor performing students special education status (Jacob, 2005), increasing caloric counts on pre-test

[27] This suggests that girls primed in this way should exert more effort than girls in the control groups. One measure of effort level is the number of questions attempted. The NAEP distinguishes between intentionally omitted questions and questions not reached, based on whether a blank response comes before or after the last non-blank response. Unfortunately, since the math assessment has traditionally been administered via an audio-paced tape, it is assumed that every student attempted all questions.

[28] To confirm this mechanism, we can examine stereotype effects by stated attitude. We expect that girls who strongly disagree with the notion that math is more for boys will exhibit stronger reactance than girls who are either neutral or agree with the stereotype. Because attitudes are only solicited from individuals in the experimental groups, I cannot directly examine this. As an approximate test, I impute a "predicted attitude index" for each student by first running an OLS regression of attitudes on a set of pre-determined characteristics including gender, race, region, parental education, grade, and private school attendance for the experimental group (a separate regression for each age and year using 13 and 17 year-olds from 1986 to 1999), and then use the estimated parameters to predict the attitudes of each student based on background characteristics. Using this continuous predicted attitude index, I split the sample into high agreement and low agreement groups depending on whether a student's attitudes agree or disagree with the gender prime (using the median attitude index value for that student's age and year as the cutoff). I then run the pooled *DID* model for 1986 to 1999 separately by high and low agreement groups and compare the stereotype threat estimates for each group. The results accord with the hypothesized mechanism – there is greater reactance in girls amongst the group more likely to disagree with the "math more for boys" attitude (for 13-year olds, the *DID* estimate is 0.014 for the disagree group, which is significant and 0.009 for the agree group, which is insignificant; the analogous estimates for 17 year-olds are 0.014 and 0.002). I also conducted this exercise by partitioning the samples into quintiles and deciles with similar qualitative conclusions, but estimates are more imprecise due to smaller sample sizes.

meals (Figlio and Winicki, 2005), or even outright cheating (Jacob and Levitt, 2003). It is unclear how stereotype effects translate from the low stakes NAEP to a high stakes environment, but this study provides additional evidence for the susceptibility of test scores to manipulation.

Although the other major randomized field experiment on stereotype threat with high stakes (for students only) AP tests finds no effects (Stricker and Ward, 2004), more research is needed using stronger stereotype language, as in the NAEP, rather than the subtle "indicate your gender" primes used in that study. Researchers should also seek to exploit natural experiments to test whether stereotype effects generalize to exams that are high stakes for students, schools, and teachers or just for schools and teachers. This then paves the way to explore possible long-term impacts of stereotype threat, for example on education and career choices, as they pertain to understanding the gender gap in labor market outcomes.

References

Allen, N.L., Carlson, J.E., and Zelenak, C.A. (1999). *The NAEP 1996 Technical Report*. NCES 1999-452. National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Allen, N.L., McClellan, C.A., and Stoeckel, J.J. (2005). *NAEP 1999 Long-Term Trend Technical Analysis Report: Three Decades of Student Performance.* NCES 2005-484. National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Altonji, J.G., and Blank, R. (1998). Race and gender in the labor market. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics Volume 3C*. Amsterdam: Elsevier.

Ambady, N., Shih, M., Kim, A., and Pittinsky, T.L. (2001). Stereotype susceptibility in children: effects of identity activation on quantitative performance. *Psychological Science*. *12*, 385-390.

Andreoni, J., and Bernheim, B.D. (2009). Social image and the 50-50 norm: a theoretical and experimental analysis of audience effects. Forthcoming in *Econometrica*.

Aronson, J., Lustina, M.J., Good, C., Keough, K., Steele, C., and Brown, J. (1999). When white men can't do math: necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*. *35*, 29-46.

Aronson, J., Quinn, D.M., and Spencer, S.J. (1998). Stereotype threat and the academic underperformance of minorities and women. In J. Swim and C. Stangor (Eds.), *Prejudice: The Target's Perspective*. New York: Academic Press.

Becker, G.S. (1985). Human capital, effort, and the sexual division of labor. *Journal of Labor Economics*. *3*, s33-s58.

Bitler, M., Gelbach, J., and Hoynes, H. (2006). What mean impacts miss: distributional effects of welfare reform experiments. *American Economic Review*. *96*, 988-1012.

Blascovich, J., Spencer, S.J., Quinn, D., and Steele, C. (2001). African Americans and high blood pressure: the role of stereotype threat. *Psychological Science*. *12*, 225-229.

Blau, F.D., and Kahn, L.M. (2000). Gender differences in pay. *Journal of Economic Perspectives*. *14*, 75-99.

Bohnet, I., and Cooter, R.D. (2005). Expressive law: framing or equilibrium selection? John F. Kennedy School of Government Faculty Research Working Paper RWP03-046.

Burgess, S., and Greaves, E. (2009). Test scores, subjective assessment and stereotyping of ethnic minorities. Working Paper.

Burnham, T., McCabe, K., and Smith, V.L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization*. *43*, 57-73.

Croizet, J.C., and Claire, T. (1998). Extending the concept of stereotype and threat to social class: the intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*. *24*, 588-594.

Croson, R., and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*. *47*, 1-27.

Cullen, M.J., Hardison, C.M., and Sackett, P.R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology*. *89*, 220-230.

Cullen, M.J., Waters, S.D., and Sackett, P.R. (2006). Testing stereotype threat theory predictions for math-identified and non-math-identified students by gender. *Human Performance*. *19*, 421-440.

Danaher, K., and Crandall, C.S. (2008). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology*. *38*, 1639-1655.

Davies, P.G., Spencer, S.J., Quinn, D.M., and Gerhardstein, R. (2002). Consuming images: how television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*. *28*, 1615-1628.

Dee, T.S. (2009). Stereotype threat and the student-athlete. NBER Working Paper No. 14705.

Figlio, D.N., and Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*. *89*, 381-394.

Fryer, R.G., and Levitt, S.D. (2008). An analysis of the gender gap in mathematics. Working Paper, Harvard University.

Fryer, R.G., Levitt, S.D., and List, J.A. (2008). Exploring the impact of financial incentives on stereotype threat: evidence from a pilot study. *American Economic Review*. *98*, 370-375.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: gender differences. *Quarterly Journal of Economics*. *118*, 1049-1074.

Goldin, C. (1994). Understanding the gender gap: an economic history of American women. In P. Burstein (Ed.), *Equal Employment Opportunity: Labor Market Discrimination and Public Policy*. Edison, NJ: Aldine Transaction.

Good, C., Aronson, J., and Inzlicht, M. (2003). Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*. *24*, 645-662.

Gupta, V.K., and Bhawe, N.M. (2007). The influence of proactive personality and stereotype threat on women's entrepreneurial intentions. *Journal of Leadership and Organizational Studies*. *13*, 73-85.

Hanna, R., and Linden, L. (2009). Measuring discrimination in education. NBER Working Paper No. 15057.

Hedges, L.V., and Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*. *269*, 41-45.

Higgins, E.T., and Rholes, W.S. (1978). "Saying is believing": effects of message modification on memory and liking for the person described. *Journal of Experimental Social Psychology*. *14*, 363-378.

Huguet, P., and Regner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology*. *99*, 545-560.

Jacob, B.A. (2005). Accountability, incentives, and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*. *89*, 761-796.

Jacob, B.A., and Levitt, S.D. (2003). Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*. *118*, 843-878.

Johns, M., Schmader, T., and Martens, A. (2005). Knowing is half the battle: teaching stereotype threat as a means of improving women's math performance. *Psychological Science*. *16*, 175-179.

Kahneman, D., and Tversky, A. (1984). Choices, values, and frames. *American Psychologist*. *39*, 341-350.

Keller, J. (2002). Blatant stereotype threat and women's math performance: self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles*. *47*, 193-198.

Keller, J., and Dauenheimer, D. (2003). Stereotype threat in the classroom: dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*. *29*, 371-381.

Kray, L.J., Reb, J., Galinsky, A.D., and Thompson, L. (2004). Stereotype reactance at the bargaining table: the effect of stereotype activation and power on claiming and creating value. *Personality and Social Psychology Bulletin*. *30*, 399-411.

Krueger, A.B. (2003). Economic considerations and class size. *Economic Journal*. *113*, F34-F63.

Lavy, V. (2004). Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment. NBER Working Paper No. 10678.

Levitt, S.D., and List, J.A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*. *21*, 153-174.

Levy, B. (1996). Improving memory in old age through implicit stereo-typing. *Journal of Personality and Social Psychology*. *71*, 1092-1107.

List, J.A. (2006). The behavioralist meets the market: measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy*. *114*, 1-37.

MacPherson, D.A., and Hirsch, B.T. (1995). Wages and gender composition: why do women's jobs pay less? *Journal of Labor Economics*. *13*, 426-471.

McFarland, L.A., Lev-Arey, D.M., and Ziegert, J.C. (2003). An examination of stereotype threat in a motivational context. *Human Performance*. *16*, 181-205.

Mechtenberg, L. (2009). Cheap talk in the classroom: how biased grading at school explains gender differences in achievements, career choices, and wages. Forthcoming in *The Review of Economic Studies*.

Neuville, E., and Croizet, J. (2007). Can salience of gender identity impair math performance among 7-8 years old girls? The moderating role of task difficulty. *European Journal of Psychology of Education*. *22*, 307-316.

Niederle, M., and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics*. *122*, 1067-1101.

Perie, M., Moran, R., and Lutkus, A.D. (2005). *NAEP 2004 Trends in Academic Progress: Three Decades of Student Performance in Reading and Mathematics*. NCES 2005-464. National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Rosenthal, R.W., and Rosnow, R.L. (1969). *Artifact in Behavioral Research*. New York: Academic Press.

Ross, L., and Ward, A. (1996). Naïve realism: implications for social conflict and misunderstanding. In T. Brown, E. Reed, and E. Turiel (Eds.), *Values and Knowledge*, ed. T. Brown, E. Reed, and E. Turiel. Hillsdale, NJ: Lawrence Erlbaum Associates.

Roth, A.E. (1995). Bargaining experiments. In J.H. Kagel and E.R. Alvin (Eds.), *Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Royalty, A.B. (1996). The effects of job turnover and the training of men and women. *Industrial and Labor Relations Review*. *49*, 506-521.

Shih, M., Pittinsky, T.L., and Ambady, N. (1999). Stereotype susceptibility: identity salience and shifts in quantitative performance. *Psychological Science*. *10*, 80-83.

Spencer, S., Steele, C.M., and Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*. *35*, 4-28.

Steele, C.M. (1997). A threat in the air: how stereotypes shape the intellectual identities and performance of women and African Americans. *American Psychologist*. *52*, 613-629.

Steele, C.M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*. *69*, 797-811.

Stricker, L.J., and Ward, W.C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*. *34*, 665-693.

Walters, A.M., Lee, S., and Trapani, C. (2004). Stereotype threat, the test-center environment, and performance on the GRE general test. GRE Board Report No. 01-03R.

Wei, T.E. (2009). Under what conditions? Stereotype threat and prime attributes. Working Paper, Harvard University.

TABLE 1 – NAEP TEST BOOKLET LAYOUT

	Age 9			Age 13			Age 17		
Booklet	*NO*	*PL*	*PR*	*NO*	*PL*	*PR*	*NO*	*PL*	*PR*
1978:									
A		x			x			x	
B		x			x			x	
C		x			x			x	
D		x			x			x	
E			xx		x			x	
F		x			x			x	
G					x			x	
H						xx			xx
I						x			x
J					x			x	
K							x		
L							x		
1982:									
A	x			x				x	
B			xx	x					xx
C		x		x					x
D		x		x			x		
E		x		x			x		
F	x			x			x		
G						xx		x	
H						x	x		
1986-1999:									
A	x				x			x	
B	x					x		x	
C		x				x			x

Notes: NO = no pre-test questions, *PL* = placebo pre-test questions, *PR* = gender prime pre-test questions. For test booklets that fall in the *PR* category, "x" denotes the presence of one prime question, while "xx" denotes the presence of two prime questions. The NAEP booklet designs are identical between 1986 and 1999.

TABLE 2 – SUMMARY STATISTICS AND RANDOMIZATION CHECKS

	Age 9		Age 13		Age 17	
	CM	TM	CM	TM	CM	TM
1978:						
Grade	3.74	3.73	7.72	7.68	10.94	10.94
Male	0.497	0.498	0.495	0.514	0.487	0.487
Black	0.138	0.138	0.131	0.131	0.118	0.118
Parents College Ed	0.238	0.229	0.259	0.251	0.324	0.322
Attend Private School	0.116	0.089	0.090	0.100	0.053	0.069
Northeast Region	0.222	0.241	0.220	0.241	0.231	0.226
4+ Reading Materials	0.373	0.351	0.554	0.546	0.631	0.622
No. of observations	12,323	2,429	19,410	4,799	22,297	4,459
Joint F-Test	0.94		1.31		1.79 [a]	
(p-value)	(0.54)		(0.21)		(0.04)	
1982:						
Grade	3.69	3.69	7.69	7.72	10.92	10.90
Male	0.492	0.497	0.500	0.511	0.485	0.493
Black	0.143	0.143	0.138	0.138	0.125	0.125
Parents College Ed	0.307	0.269	0.322	0.317	0.320	0.311
Attend Private School	0.145	0.085	0.105	0.110	0.089	0.069
Northeast Region	0.212	0.203	0.238	0.236	0.241	0.237
4+ Reading Materials	0.354	0.360	0.476	0.462	0.573	0.572
No. of observations	9,993	2,045	11,768	3,990	12,170	4,149
Joint F-Test	1.97 [b]		0.82		0.97	
(p-value)	(0.03)		(0.67)		(0.50)	
1986-1999:						
Grade	3.65		7.62	7.63	10.84	10.81
Male	0.494		0.498	0.495	0.492	0.491
Black	0.157		0.150	0.151	0.145	0.152
Parents College Ed	0.410		0.416	0.420	0.410	0.426
Attend Private School	0.137		0.094	0.086	0.066	0.091
Northeast Region	0.216		0.219	0.219	0.224	0.228
4+ Reading Materials	0.340		0.494	0.501	0.578	0.584
No. of observations	37,611		12,219	24,190	12,036	11,749
Joint F-Test	--		1.07		1.39	
(p-value)			(0.38)		(0.14)	

Notes: [a] If "other race" is excluded, F-Test (p-value) is 1.30 (0.22). [b] If "attends private school" is excluded, F-Test (p-value) is 1.32 (0.21). CM = control group mean; TM = treatment group mean; F-statistics are from a joint F-Test that the treatment and control group covariates significantly differ from each other. 19 variables were tested in total (some variables omitted to conserve space), including grade, gender, race, parental education, private school attendance, geographic region of residence, and number of reading materials in the household. Data from 1986 to 1999 are pooled together. Student sampling weights are applied.

TABLE 3 – DIFFERENCES-IN-DIFFERENCES ESTIMATES OF STEREOTYPE THREAT, FULL SAMPLE

	Dependent Variable = Test Score (proportion correct) ; [M = 0.58 , SD = 0.18]					
DID (female x treat coefficient):	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Panel A [Equation 2]						
Any Prime	0.0032 (0.0017)* [0.0017]*	0.0028 (0.0018) [0.0021]	0.0028 (0.0021) [0.0020]	0.0024 (0.0020) [0.0026]	0.0023 (0.0052) [0.0025]	0.0019 (0.0049) [0.0030]
Panel B [Equation 3]						
"Math More For Boys" Prime	0.0073 (0.0020)** [0.0021]**	0.0057 (0.0022)** [0.0025]**	0.0081 (0.0023)** [0.0024]**	0.0062 (0.0024)** [0.0030]**	0.0074 (0.0053) [0.0024]**	0.0086 (0.0051)* [0.0028]**
"Math More For Girls" Prime	-0.0038 (0.0023) [0.0023]	-0.0024 (0.0025) [0.0026]	-0.0066 (0.0027)** [0.0030]**	-0.0044 (0.0027) [0.0035]	-0.0016 (0.0042) [0.0026]	-0.0043 (0.0043) [0.0032]
F-test of difference in primes:						
Age-Year-Block Clusters	17.20**	8.32**	24.61**	12.72**	3.08*	7.12**
Primary Sampling Unit Clusters	15.73**	6.91**	17.31**	7.17**	9.88**	12.85**
Age-Year Interactions with Covariates	x	x	x	x		
Student Sampling Weights	x		x		x	
Covariates	x	x			x	
Number of Observations	207,127	207,127	207,637	207,637	207,127	207,637

Notes: For all specifications, the dependent variable is the NAEP math test score (fraction of questions correctly answered). The right-hand-side variables are a dummy for female, treatment, a female-treatment interaction term, age-year dummies, and when indicated, covariates (geographic region, parental education, race, grade, and private school attendance) and/or age-year dummies interacted with all variables except for the female-treatment interaction term – see equations 2 and 3 for details. Treatment group students are those who received the relevant gender prime pre-test questions, while control group students are those who received placebo or no pre-test questions. Only the female-treatment interaction coefficients, which capture the differences-in-differences effects (*DID*), are reported above. The female main effect is significant and about -0.02. Standard errors are reported below each *DID* estimate. Parentheses correspond to clustering standard errors by age-year-block, while brackets correspond to primary sampling unit clusterings. F-tests comparing the "math more for boys" and "math more for girls" primes are also displayed for both age-year-block and primary sampling unit clusters. See text for discussion. * = significant at 10% level. ** = significant at 5% level.

TABLE 4 – DIFFERENCES-IN-DIFFERENCES ESTIMATES OF STEREOTYPE THREAT, BY AGE

	Dependent Variable = Test Score (proportion correct)					
	Age 9 (M = 0.58 , SD = 0.19)		Age 13 (M = 0.57 , SD = 0.17)		Age 17 (M = 0.59 , SD = 0.19)	
DID (female x treat coefficient):						
Panel A [Equation 2]						
Any Prime	-0.0019 (0.0059)	-0.0020 (0.0063)	0.0029 (0.0028)	0.0013 (0.0033)	0.0043 (0.0022)*	0.0050 (0.0026)*
Panel B [Equation 3]						
"Math More For Boys" Prime	-0.0019 (0.0059)	-0.0020 (0.0063)	0.0091 (0.0043)**	0.0087 (0.0047)*	0.0058 (0.0021)**	0.0072 (0.0020)**
"Math More For Girls" Prime			-0.0027 (0.0034)	-0.0055 (0.0033)	0.00027 (0.0030)	-0.00037 (0.0048)
F-test of difference in primes:			7.90**	12.50**	4.17**	2.82*
Covariates	x		x		x	
Number of Observations	64,206	64,401	76,116	76,376	66,805	66,860

Notes: Regressions are conducted for each age group noted, pooling across all sample years. For all specifications, the dependent variable is the NAEP math test score (fraction of questions correctly answered). The right-hand-side variables are a dummy for female, treatment, a female-treatment interaction term, year dummies, year dummies interacted with all covariates except for the female-treatment interaction term, and when indicated, covariates (geographic region, parental education, race, grade, and private school attendance) – see equations 2 and 3 for details. Student sampling weights are included in all regressions. Treatment group students are those who received the relevant gender prime pre-test questions, while control group students are those who received placebo or no pre-test questions. Only the female-treatment interaction coefficients, which capture the differences-in-differences effects (*DID*), are reported above. The female main effect is significant and about -0.02 for 13 year-olds and -0.04 for 17 year-olds. Standard errors in parentheses below each *DID* estimate are clustered by year-block. F-tests comparing the "math more for boys" and "math more for girls" primes are also displayed. The age 9 samples were not given the "math more for girls" prime, thus the estimates for "math more for boys" and "any prime" are equivalent. See text for discussion. * = significant at 10% level. ** = significant at 5% level.

FIGURE 1 – PERCEPTIONS IN RESPONSE TO "MATH MORE FOR BOYS THAN GIRLS" DATA POOLED FROM 1978-1999 [29]

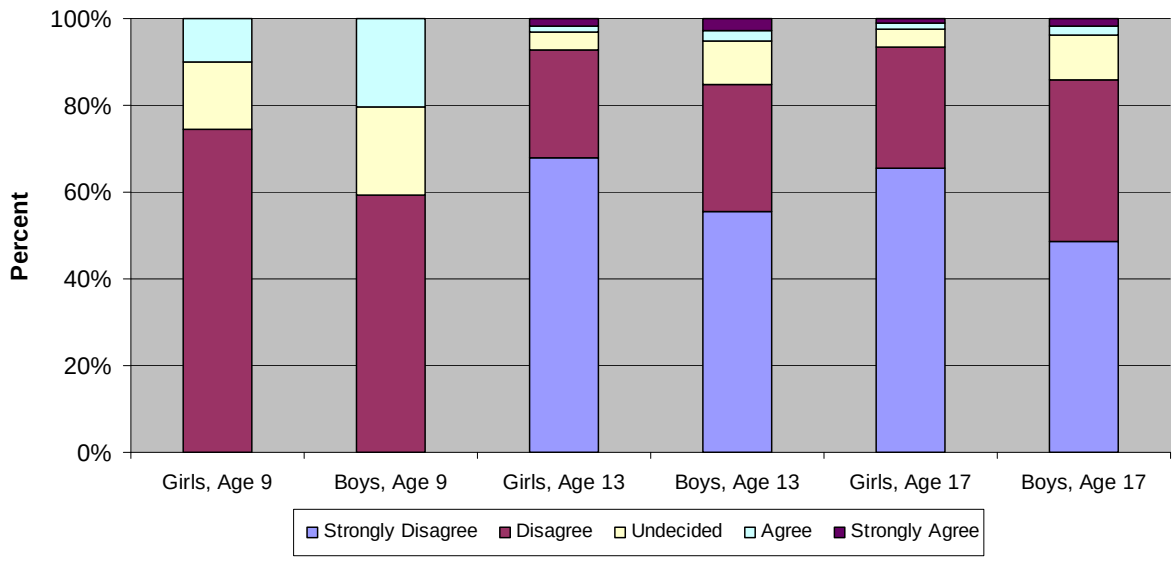
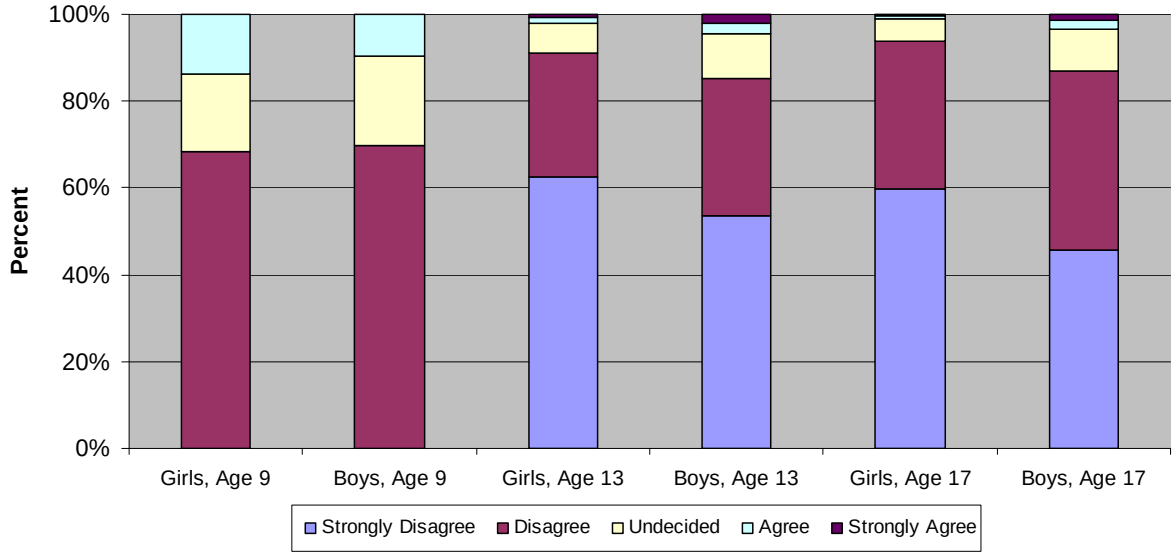


FIGURE 2 – PERCEPTIONS IN RESPONSE TO "MATH MORE FOR GIRLS THAN BOYS" DATA POOLED FROM 1978-1999



[29] For Figures 1 and 2: the "strongly disagree" and "strongly agree" choices were not in the original question for the age 9 cohorts.

FIGURE 3 – QUANTILE TREATMENT EFFECTS OF STEREOTYPE THREAT
POOLED SAMPLE, "MATH MORE FOR BOYS" PRIME [30]

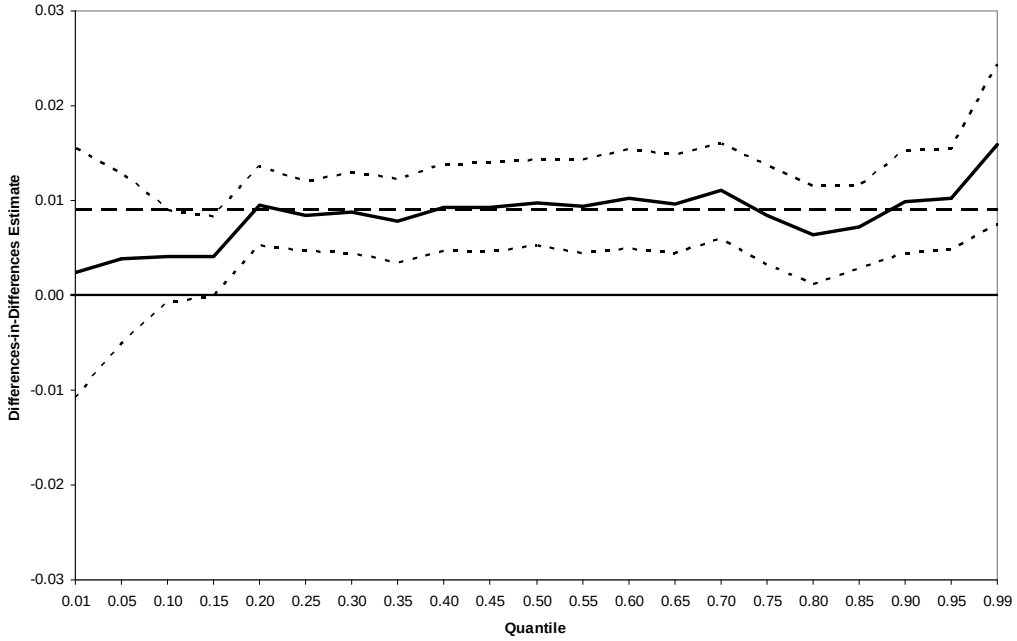
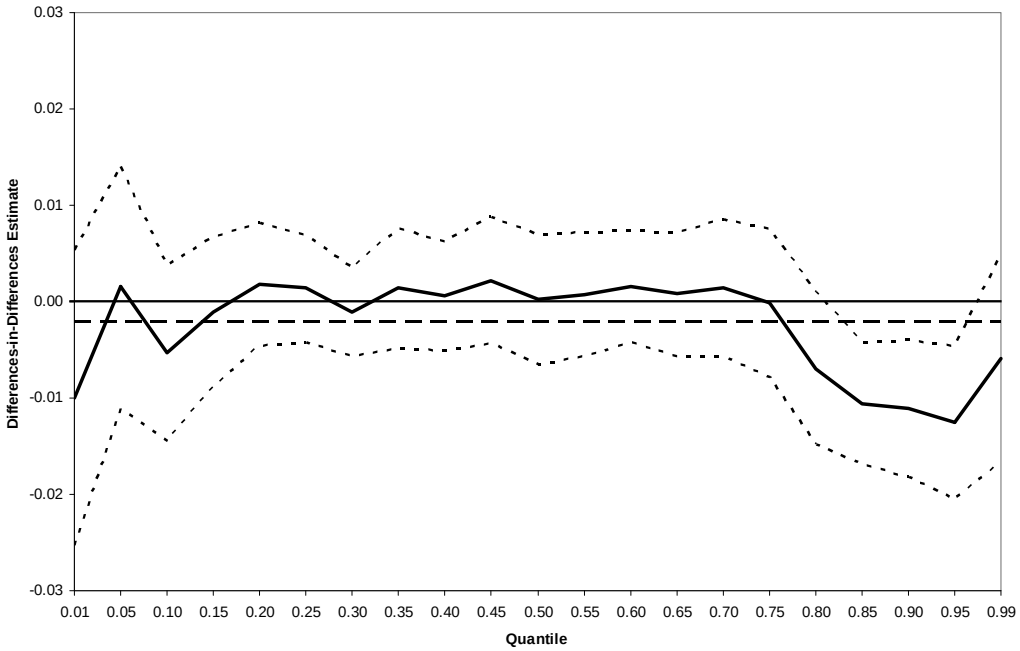


FIGURE 4 – QUANTILE TREATMENT EFFECTS OF STEREOTYPE THREAT
POOLED SAMPLE, "MATH MORE FOR GIRLS" PRIME



[30] Figures 3 and 4 plot coefficients (dark solid line) from the female-treatment interaction term in a quantile regression model with NAEP test score (fraction correctly answered) as the dependent variable, as well as main effects for female, treatment, and a full set of covariates (geographic region, race, parental education, grade, and private school attendance) on the right-hand-side for each indicated quantile (see equation 2). Dotted lines provide bootstrapped 95% confidence intervals clustered by age-year-block. Dashed line shows the mean OLS impact.

FIGURE 5 – TEST PERFORMANCE BY TREATMENT AND CONTROL CONDITION
DATA POOLED FROM 1978-1999

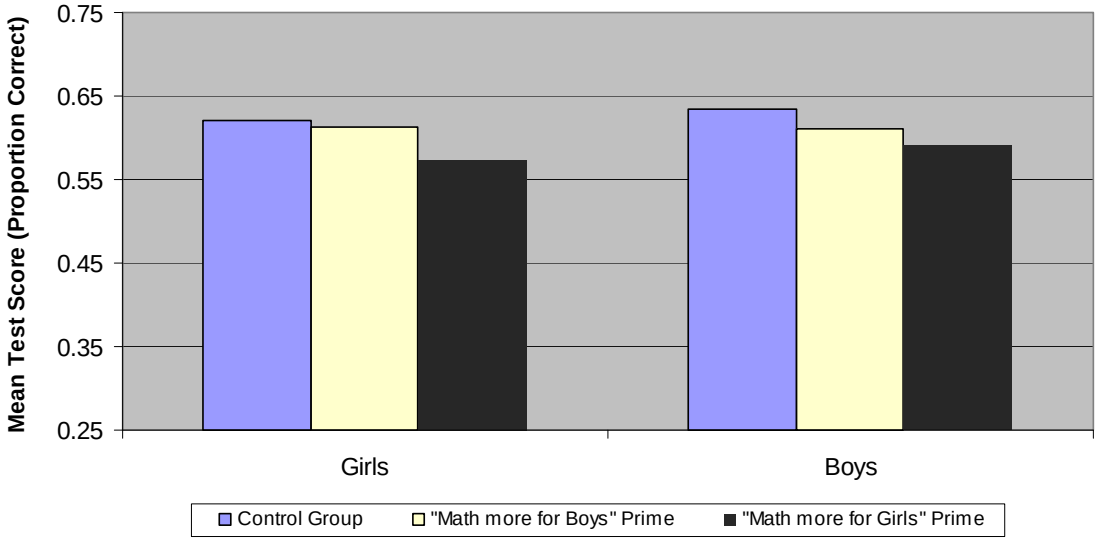
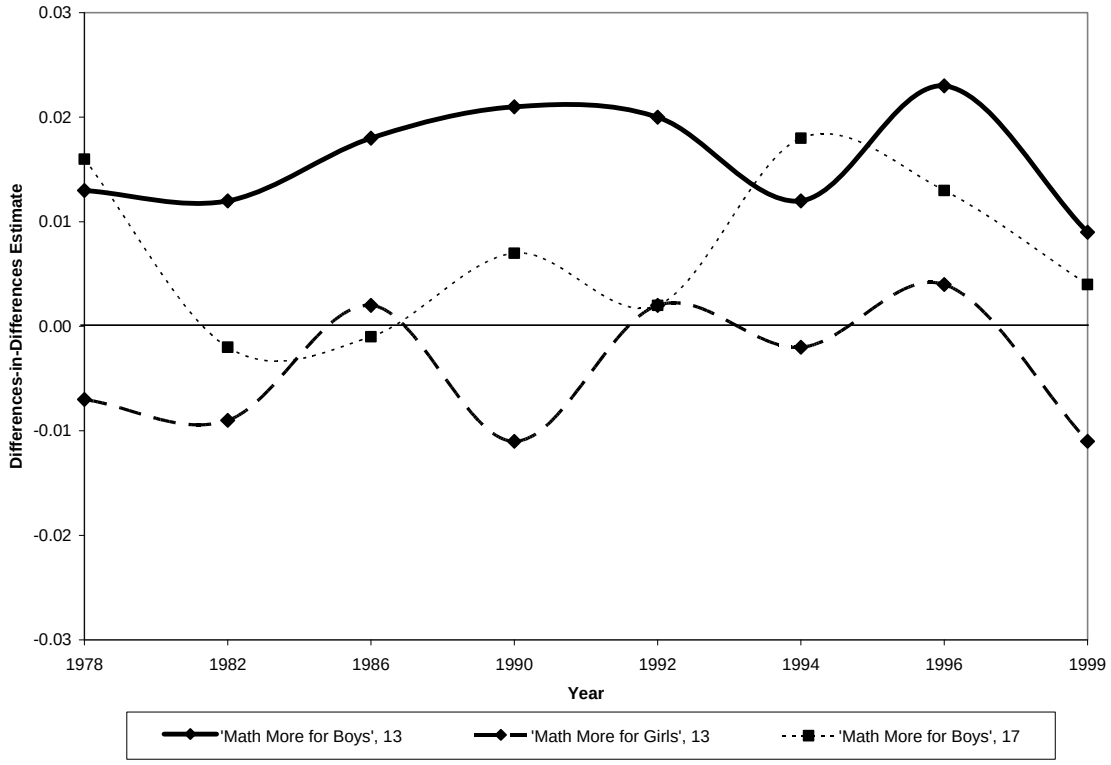


FIGURE 6 – STEREOTYPE EFFECT TRENDS BY TYPE OF PRIME AND AGE [31]



[31] Figure 6 plots coefficients from the female-treatment interaction term in a model with NAEP test score (fraction correctly answered) as the dependent variable, as well as main effects for female, treatment, and a full set of covariates (geographic region, race, parental education, grade, and private school attendance) on the right-hand-side for each separate age, year, and prime noted.

APPENDIX TABLE 1 – PRIMING AND PLACEBO QUESTIONS

	Age 9	Age 13	Age 17
Placebo:			
1.	Do you feel it is important to learn about money?	How often do you do math homework?	Have you ever studied math through computer instruction?
2.	Do you like to learn how to measure with a ruler?	How often do you see your teacher do math on the board?	Do you feel that computer knowledge helps one get a better job?
3.	Do you feel it is easy to learn to weigh objects on a scale?	How often do you work ahead in your math book?	Do you feel that computers are suited for doing monotonous tasks?
4.	Do you like to play mathematics games?	How often do you use a mathematics textbook?	Do you feel that computers store instructions and information?
5.	Do you like listening to your teacher explain math?	Do you feel it is useful to get help from your teacher on math?	How often do you take mathematics tests?
6.	Do you feel that it helps to use a mathematics book?	How important or not is social studies?	How often do you do math problems that are not assigned?
7.	Do you like or dislike physical education?	Do you feel it is important to be able to estimate answers to problems?	How often do you work math problems at the board?
8.	Would you like to work at a job using math?	Do you like to do proofs?	Do you feel that math helps a person think logically?
Prime:			
1.	Do you feel that math is more for girls than for boys? (Strongly disagree, Disagree, Undecided, Agree, Strongly agree)		
2.	Do you feel that fewer men have logical ability than women? (Strongly disagree, Disagree, Undecided, Agree, Strongly agree)		
3.	Do you feel that mathematics is more for boys than for girls? (Strongly disagree, Disagree, Undecided, Agree, Strongly agree)		

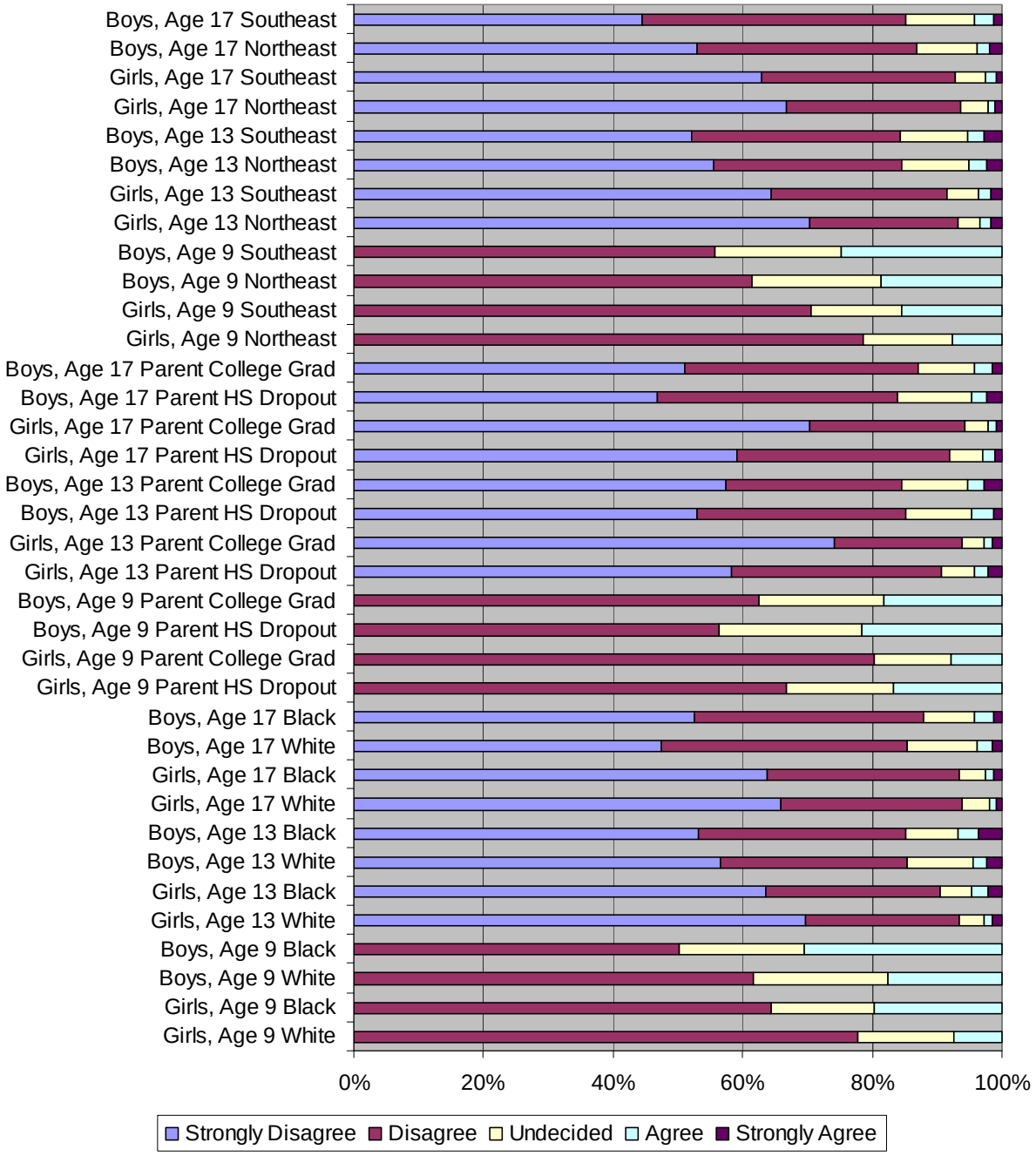
Notes: The lists above provide examples of the types of placebo and priming questions used in the NAEP long-term trend data.

APPENDIX TABLE 2 – DIFFERENCES-IN-DIFFERENCES ESTIMATES OF STEREOTYPE THREAT, EXCLUDING CONTROL GROUPS WITH NO PRE-TEST QUESTIONS

	Dependent Variable = Test Score (proportion correct) ; [M = 0.58 , SD = 0.18]			
DID (female x treat coefficient):				
Panel A [Equation 2]				
Any Prime	0.0040 (0.0018)** [0.0020]**	0.0035 (0.0018)* [0.0023]	0.0043 (0.0021)** [0.0024]*	0.0041 (0.0020)** [0.0028]
Panel B [Equation 3]				
"Math More For Boys" Prime	0.0078 (0.0017)** [0.0022]**	0.0062 (0.0020)** [0.0026]**	0.0092 (0.0018)** [0.0026]**	0.0075 (0.0021)** [0.0031]**
"Math More For Girls" Prime	-0.0032 (0.0023) [0.0026]	-0.0021 (0.0026) [0.0028]	-0.0052 (0.0028)* [0.0033]	-0.0029 (0.0030) [0.0038]
F-test of difference in primes:				
Age-Year-Block Clusters	16.55**	8.17**	23.32**	11.54**
Primary Sampling Unit Clusters	15.37**	7.12**	16.43**	6.78**
Student Sampling Weights	x		x	
Covariates	x	x		
Number of Observations	153,739	153,739	154,226	154,226

Notes: For all specifications, the dependent variable is the NAEP math test score (fraction of questions correctly answered). The right-hand-side variables are a dummy for female, treatment, a female-treatment interaction term, age-year dummies, age-year dummies interacted with all variables except for the female-treatment interaction term, and when indicated, covariates (geographic region, parental education, race, grade, and private school attendance) – see equations 2 and 3 for details. Treatment group students are those who received the relevant gender prime pre-test questions, while control group students are those who received *placebo pre-test questions only*. Only the female-treatment interaction coefficients, which capture the differences-in-differences effects (*DID*), are reported above. Standard errors are reported in parentheses below each *DID* estimate. Parentheses correspond to clustering standard errors by age-year-block, while brackets correspond to primary sampling unit clusterings. F-tests comparing the "math more for boys" and "math more for girls" primes are also displayed for both age-year-block and primary sampling unit clusters. See text for discussion. * = significant at 10% level. ** = significant at 5% level.

APPENDIX FIGURE 1 – PERCEPTIONS IN RESPONSE TO
"MATH MORE FOR BOYS THAN GIRLS" FOR SELECTED SUBGROUPS
DATA POOLED FROM 1978-1999 [32]



[32] The "strongly disagree" and "strongly agree" choices were not in the original question for the age 9 cohorts.