May 2008

A "Rebound Effect" After Stereotype Threat?

Patrick Schnarrenberger
Macalester College, schnarrd@gmail.com

Running Head: STEREOTYPE REBOUND

A "Rebound Effect" After Stereotype Threat?

Patrick Alan Schnarrenberger

Kendrick Brown, Psychology Department

Submitted 5/5/08

Abstract

Two studies investigated a potential cognitive mediator for stereotype threat, a phenomenon whereby the mere threat of confirming a negative stereotype results in a performance deficit. It was hypothesized that people attempt to suppress stereotypes in memory during threatening situations, consuming cognitive resources, but that the suppression is released after the threatening situation has ended. This results in a "rebound effect" and a subsequent increase in stereotyped thought. The experiments failed to find a significant stereotype threat effect when examined individually, but when the data from the experiments were aggregated aggregated, a performance deficit was found. However, because of the failure to find a significant performance deficit in any one experiment, the results to not directly bear on any potential rebound effect.

A "Rebound" Effect After Stereotype Threat?

Stereotype threat is a phenomenon in which the mere threat of confirming a negative stereotype in and of itself results in decreased performance on a given task. This phenomenon is relatively robust, and has been demonstrated in such groups as women taking math tests (Quinn & Spencer, 2001), African Americans taking a verbal portion of the GRE (e.g. Steele & Aronson, 1995), and even White Americans taking the Implicit Attitudes Test (Frantz, Cuddy, Burnette, Ray, & Hart, 2004). Current research makes less clear, however, exactly how stereotype threat occurs. More specifically, the question remains as to what mediating processes cause the threat of confirming a negative stereotype to decrease performance on a given task. The present study seeks to help clarify this issue by determining whether people engage in thought suppression with respect to stereotypes when performing a threatening task.

The earliest research on stereotype threat, conducted by Steele and Aronson (1995), focused on explaining the black / white standardized test gap. It found that if African Americans were told that a test was diagnostic of their verbal abilities, they scored lower on the test than if they were told the test was investigating the psychological factors in solving verbal problems. Steele and Aronson (1995) explained these findings by theorizing that the diagnosticity manipulation primed the participants' awareness of stereotypes positing African Americans' poor mental ability, and that the awareness of these stereotypes placed an extra affective load on these participants that resulted in decreased performance.

Underlying Mechanisms

Subsequent research has attempted to pinpoint the underlying mechanism causing stereotype threat. Following Steele and Aronson's (1995) assumption that stereotype threat causes a burdensome affective load, a significant portion of these studies has focused on the affective underpinnings of stereotype threat. The theory is that aversive affective states, such as increased arousal and anxiety, lead to distraction during the threatening task, and thus,

decreased performance. For example, several studies have suggested that increased anxiety may be related to the stereotype threat phenomenon (Aronson, Lustina, Good, Keough, Steele, & Brown, 1999; Spencer, Steele, & Quinn, 1999). Additionally, manipulations designed to decrease anxiety and arousal seem to moderate the effects of stereotype threat (Martens, Johns, Greenberg, & Schimel, 2002).

However, other studies make the link between affect and stereotype threat less clear. For example, Gonzales, Blanton and Williams (2002) found no difference in self-reported anxiety between conditions. In another study, Brown and Josephs (1999) found no evidence that words related to performance anxiety were more accessible in memory during stereotype threat, implying that anxiety may not strongly mediate the phenomenon. Other studies have yielded similar results (e.g. Oswald & Harvey, 2001, Schmader, 2002, Stone, Lynch, Sjomeling, & Darley, 1999). Thus, the general picture that emerges from the literature regarding an affective mediator for stereotype threat is that negative affect may contribute to decreased performance, but that it is by no means the only factor involved.

Since anxiety does not seem to account fully for the damaging effects of stereotype threat, other research has sought to determine other possible mediators. One possible explanation is that the stereotype negatively impacts an individual's self-confidence to perform the given task and, in a self-fulfilling nature, subsequently disrupts the individual's ability to perform on the task (Rosenthal & Jacobson, 1966). In their original study, Steele and Aronson (1995) provided a test of this hypothesis by administering a word completion task that provided an implicit measure of self-doubt after finishing the threatening task. In support of the performance confidence hypothesis, they found that black participants in the threat condition spontaneously filled in the highest number of doubt-related words. In another study using a similar measure of self-doubt as Steele and Aronson (1995), Stone (2002) induced stereotype threat for golf performance by relating golf to intelligence for African

Americans and natural ability for White Americans. Stone then administered a word completion task for both groups of golfers measuring self-doubt. Both threatened groups produced more doubt-related words than control groups.

However, although the studies cited above have found potential links between stereotype threat and self confidence, a number of studies have not (e.g. Aronson, Lustina, Good, Keough, Steele, & Brown, 1999, Keller, 2002, Kray, Thompson, & Galinski, 2001, Shih, Pittinsky, & Ambady, 1999). For example, Kray and colleagues (2001) conducted an experiment at an MBA program in which one member of either a male-female or a male-male negotiating dyad were either told that a negotiating exercise was highly diagnostic of important, stereotypically male managerial skills, or that the exercise was simply illustrative of classic buyer-seller bargaining. Additionally, Kray and colleagues administered a self-confidence to the participants across both conditions. Although female participants exhibited a performance deficit in the threat condition relative to the male participants, they showed no corresponding change in confidence levels. Thus, although performance confidence may partially mediate stereotype threat, current evidence suggests that this construct cannot fully account for stereotype threat's damaging effects.

Another body of research has examined possible behavioral mediators for stereotype threat. Some researchers have investigated whether stereotype threat damages performance by causing the target of the threat to change the amount of effort expended in the threatening task. According to this view, stereotype threat would either cause the threatened person to decrease the effort expended, resulting in decreased performance, or drastically increase the amount of effort expended, resulting in overexertion and, ultimately, decreased performance. However, a variety of studies have shown that neither of these theories is a good explanation for the negative effects of stereotype threat (e.g. Aronson et al., 1999, Gonzales, Blanton, & Williams, 2002, Keller, 2002, Keller & Dauenheimer, 2003, Smith & White, 2002).

Another possible behavioral mediator of stereotype threat is self-handicapping, or an individual's attempt to protect the self by either behaving or claiming to have behaved in such a way that a poor performance could be attributed to external circumstances (Leary & Shepperd, 1986). Applied to stereotype threat, self-handicapping implies that threatened individuals may claim more self-handicaps in response to a threatening stereotype. This hypothesis has received partial support in subsequent research. For example, Stone (2002) found that threatened individuals opt not to take practice swings when preparing for a golf task. Additionally, Quinn & Spencer (2001) found that threatened females were less able to formulate strategies for solving math word problems. However, as in the case of the other potential mediators for stereotype threat discussed above, the majority of studies investigating self-handicapping as a potential mediator have provided null results (e.g. Croizet & Claire, 1998, Keller & Dauenheimer, 2003, Kray et al., 2001, Shih et al., 1999).

An additional potential mediator suggested by Smith (2004) that is not easily classified into affective, behavioral, or cognitive categories relates to the adoption of performance-avoidance versus performance-approach achievement goals. Achievement goal research posits that performance expectancies determine the type of achievement goal adopted (Elliot & Church, 1997). A high performance expectancy results in a performance-approach achievement goal, wherein the goal is to demonstrate competence, whereas a low performance expectancy generates a performance-avoidance goal, wherein the goal is to avoid demonstrating incompetence. Several researchers have found that performance-approach achievement goals generally result in positive outcomes, while performance-avoidance goals generally result in negative outcomes (e.g. Barron & Harackiewicz, 2001, Elliot & Church, 1997). Moreover, Smith, Sansone, and White (2007) have found that women with high achievement motivation who undergo stereotype threat spontaneously adopt performance-

avoidance goals. However, the literature linking stereotype threat to achievement goals is still in its infancy, and is unlikely to provide a complete mediator for stereotype threat.

Thought Suppression as a Mediator

Although the above research focuses mainly on the behavioral and phenomenological underpinnings of stereotype threat, another body of research focuses on possible cognitive factors of the phenomenon. For example, Croizet, Després, Gauzins, Huguet, Leyens, and Méot, (2004) presented an adaptation of the Progressive Matrices Test to participants with a perceived reputation for intellectual inferiority. Concurrent with the Progressive Matrices Test, Croizet and colleagues administered several autonomic measures of cognitive load. When the task was presented as a measure of intellectual ability, people belonging to stereotyped groups not only exhibited performance deficits, but also exhibited evidence of greater cognitive load than people from non-stereotyped groups. This study implies a link between a person's available cognitive resources and stereotype threat. In another study investigating the cognitive factors of stereotype threat, Schmader and Johns (2003) found that stereotype threat reduces participants' scores on a working memory task, even though the task was presented as being unrelated to the task triggering the stereotype threat. To explain these results, Schmader and Johns conjecture that cognitive resources might be consumed in suppressing thoughts relating to the target stereotype. Indeed, other research, such as that of Spencer (2003), indicates that targets of a stereotype might often try to suppress thoughts related to that stereotype. This implies that cognitive factors may mediate some of the negative ramifications of stereotype threat. Thus, the extra effort expended in suppressing thoughts related to a stereotype seems to be a reasonable explanation for the working memory deficits elicited in Schmader and Johns' research.

With respect to thought suppression, previous studies have shown that suppressing thoughts actually makes those thoughts more accessible in memory. Wegner and Erber (1992)

demonstrated that attempting to suppress a target word in a high-stress situation resulted in longer reaction times in naming colors and more responses matching that target word when the word is primed than in a control condition. To explain these results, they proposed a model, termed the ironic process model, whereby the mind utilizes two simultaneous processes to suppress thoughts, one controlled and one automatic. The automatic process scans the contents of working memory to detect traces of the unwanted thought, which would indicate a lapse in mental control. If such a thought is detected, the controlled process, which is intentional and limited by available working memory, seeks to replace this thought with a suitable distracter thought. The implication of this model is that at any given time, a person attempting to suppress a thought should be highly sensitized to the presence of the unwanted item. Thus, when high-stress situations trigger a large cognitive load, the controlled process becomes subsequently less able to replace unwanted thoughts with distracter items and unwanted thoughts become even more accessible than if the person did not engage in thought suppression at all.

This theory has since been applied to stereotype suppression. The reasoning goes that since there is a present social norm prohibiting the explicit expression of stereotyped attitudes, people who have stereotypes made salient to them have an incentive to consciously suppress these thoughts. However, according to the ironic process theory, people attempting to suppress stereotyped thoughts who are under a disruptive mental load should have these thoughts hyperaccessible in working memory, and thus when these people discontinue their thought suppression effort, they should experience a "rebound effect" where they are hyper-aware of the unwanted stereotype. Consistent with this model, Macrae and colleagues (1994) show that when people try to suppress stereotyped thoughts, they subsequently express more negative stereotyped thoughts than people in a control condition. Thus, stereotype suppression may actually lead to more stereotyped thinking than may otherwise occur.

These results have important implications for people in situations where stereotype threat is triggered. First, they imply that if a person engages in thought suppression during the specific task, as long as this task is sufficiently difficult and uses sufficient cognitive resources, suppression efforts should use cognitive resources that would normally be allocated to the task at hand, resulting in decreased performance. However, if the task is not sufficiently difficult, there should be no stereotype threat effect. Consistent with this prediction, O'Brien and Crandall (2003) have found that when women are administered simple math tests, which presumably do not use sufficient cognitive resources to conflict with the stereotype suppression process, they do not exhibit the stereotype threat effect. However, when the difficulty is scaled up, increasing cognitive load, performances in the stereotype threat condition decrease with respect to controls. Second, the results imply that people in a threat condition engaging in thought suppression should experience a rebound effect after the thought control process is relaxed, and thus stereotyped thoughts should be hyperaccessible after the threat condition has ended. This rebound effect could represent post-test ruminations about either the test itself or the individual's performance on said test. However, the reason finding a rebound effect after stereotype threat is important is that it would re-emphasize the value of considering the cognitive approach when devising interventions to reduce or eliminate the stereotype threat effect. Additionally, although some of the theories regarding the underlying mechanisms for stereotype threat discussed above appear to have more empirical support than others, it is important to emphasize that these theories are not mutually exclusive. The overall picture that emerges from the literature is that no one factor fully mediates stereotype threat. Instead, the threat of confirming a negative stereotype about one's group triggers a cascade of changes at multiple levels of analysis, including, but not necessarily limited to, the affective, behavioral, phenomenological, and cognitive. These changes then combine and interact, producing a performance deficit. In a phenomenon as

complex as stereotype threat, multiple psychological approaches are not only beneficial, but necessary for devising effective interventions. Thus, this study attempted to use the fine grain of analysis provided by the cognitive approach to describe one of the proximate factors responsible for the destructive effects of stereotype threat with the ultimate goal of allowing more targeted interventions at the cognitive level.

In service to the above goal, the present study sought to provide a conclusive link between stereotype threat and thought suppression, and subsequently further elucidate the means by which the cognitive mediators of stereotype threat operate. First, I will attempt to show that females engaging in a threatening math task engage in thought suppression during the task, resulting in a "rebound effect", where unwanted stereotypes are hyperaccessible after the completion of the task. Second, I hope to show that when this thought suppression is eliminated, the negative effects of stereotype threat are alleviated.

Experiment 1

Experiment 1 was designed to establish a firm link between stereotype threat and thought suppression. Participants engaged in a math task that has been framed as either having been shown to produce gender differences or as having been shown to be neutral with respect to gender. Each participant, regardless of how the math task was framed, will complete two lexical decision tasks designed to measure the activation of stereotypes related to women, one during the math task and one afterwards. If the threat manipulation is effective and thought suppression does play a role in stereotype threat, I expect to see two main patterns of results. First of all, participants in the threat condition should score lower on the math task than participants in the control condition. Secondly, participants in the threat condition should show relatively more stereotype activation after the math task is completed as compared to during the task because of the rebound effect. Participants in the control

condition should show a similar amount of stereotype activation both during and after the math task.

Method

Participants

The participants consisted of 45 female Macalester college students of ages 18 to 21, who were told that they were being recruited for a study researching gender differences in math ability. Of these, 33 participated for course credit, while 12 were entered into a random drawing for a $50 gift certificate to local merchants. Additionally, the data from 5 participants were excluded from the analysis, 4 because they were non-native English speakers and 1 because the debriefing revealed that she had guessed the true motive for the experiment.

Materials

There were three primary materials for this experiment, namely one 45-minute math section from the general Graduate Register Examination (GRE) consisting of a total of 29 questions and two separate lexical decision tasks, which were used as measures of stereotype activation. The GRE math section was divided into approximately equal halves, with one extra question assigned to the first half because the experimenter perceived that the first half was the least difficult. Following the work of several other researchers (e.g. Macrae, Bodenhausen, & Milne, 1995, Mussweiler, 2006, Parrott, Zeichner, & Hoover, 2006), the lexical decision tasks each consisted of 10 words and 10 non-words. Of the 10 words, 5 were target words consisting of adjectives highly associated with stereotypically female traits (e.g. *romantic*, *gentle*, *emotional*), while 5 were filler words consisting of adjectives irrelevant to stereotypical female traits (e.g. *healthy, foreign, straight*). The stereotypically feminine words were obtained through a pre-test in which 20 female participants were asked to rate 51 personality words on a scale from 1 to 9 as to their masculinity and femininity. The 10

highest-scoring feminine words with mean femininity scores of at least 7 were selected for use in the two lexical decision tasks. Lists of the non-words, filler words, and target words are given in the Appendix.

Procedure

Each participant was randomly assigned to either the threat or no-threat condition. In both conditions, the participant entered the testing room, after which the (male) experimenter asked the participant to sign a consent form and fill out some basic demographic information. In the threat condition, this demographic information contained a question asking the gender of the participant, which served to make gender a salient characteristic. The experimenter then explained that the participant was going to take the first half of a 40-minute math test. Halfway through the math test, the participant would also complete a short verbal task to "cleanse the palette" before completing the second half of the math test. Additionally, the experimenter explained in the experimental condition that the math test had been previously shown to produce gender differences, while in the control condition, the experimenter described the test as having been shown to be "gender neutral." The experimenter then administered one half of a 45-minute GRE math section. After 20 minutes, the experimenter collected the first half of the test and led the participant to a separate chamber, where the participant completed the first lexical decision task. Upon completion of this lexical decision task, the participant was led back to the original testing chamber, where the experimenter then administered the remaining half of the GRE math section. After 20 minutes, the experimenter collected the final portion of the math test. However, before the participant left the testing chamber, the experimenter explained that he had a "friend" for whom he was doing a favor. The experimenter then explained that this "friend" needed more data, and was using a verbal task very similar to the one the participant had taken earlier, and the experimenter subsequently asked whether the participant would be willing to complete this second verbal

task. If the participant agreed, she was led to a third room to complete the second lexical decision task, after which she was probed for suspicion and then thoroughly debriefed. If the participant did not agree, the experimenter proceeded directly to the debriefing. No participants refused to complete this second verbal task.

Results

Manipulation Check

The math test score data were first analyzed to ensure the threat manipulation was effective. The answer to one question in the first half of the math test was found to be ambiguous, and answers to that question were consequently excluded from the analysis. The resultant mean test score of participants in the no-threat condition was 16.05, while the mean test score in the threat condition was 14.70, which is slightly lower than the test score in the no-threat condition. A one-tailed independent groups t-test, however, reveals that this difference was not significant, $t(38) = 1.26$, $p = ns$. This implies that the threat manipulation was not effective in generating a stereotype threat effect. The proceeding analyses, therefore, should be considered with this in mind.

Stereotyped Thought Data

These data consist of response times to lexical decision tasks. The tasks each contained 10 non-words, 5 stereotypically female target words, and 5 stereotype-unrelated filler words. Each lexical decision task could be administered either during or after the math test.

Inaccurate responses were first excluded from the data, and the resulting response times were submitted to a logarithmic transformation. These data were then grouped by time, which could have been during or after the math test, and threat level, which could have been high or low. Means and standard deviations for each participant's data were then calculated according to these groups, and individual response times more than 3 standard deviations

from the participant mean were excluded from the analysis. Participant means that then consisted of only 2 response times were also excluded. The preceding steps resulted in less than 5% elimination of response time data, and resulted in zero excluded participant means. The grand means calculated by time and condition of the resulting mean participant reaction times are presented in Table 1. Note that while I used the logarithmically transformed reaction times for our analyses, only the untrasformed means are presented in the table.

As can be seen from the table, the reaction times for the high and low threat conditions do not appear to differ significantly. A 2 (threat) x 2 (time) repeated measures mixed ANOVA confirms this fact, with no significant main effect for threat, $F(1, 38) < 1$, $p = ns$. However, there does appear to be a difference for time, a fact which is confirmed through the above repeated measures mixed ANOVA, $F(1, 38) = 6.835$, $p < .05$. There is no significant interaction between threat and time, $F(1, 38) < 1$, $p = ns$. Thus, all participants appear to have responded faster to the stereotypically female words following the math task, and this pattern did not vary according to threat condition.

Discussion

The main hypotheses of this study were that participants who undergo stereotype threat would experience more stereotyped thought after the threatening task than during, and additionally that these participants would experience a similar amount of stereotyped thought as non-threatened participants during the task, but more stereotyped thought afterwards. Because of this experiment's failure to find a significant effect for stereotype threat, and because the main hypotheses of this experiment are predicated upon finding effects for stereotype threat, the results of this experiment do not clearly support or refute any of these hypotheses. However, a few explanations for why the experiment failed to find a significant stereotype threat effect do deserve some examination.

The first, and perhaps most obvious, explanation for the failure of the stereotype threat manipulation was that the manipulation was simply ineffective, and therefore failed to produce a measurable difference in performance between the two groups. However, in light of the fact that the threat manipulation used for this experiment has been successfully used several times in the stereotype threat literature (e.g. Spencer et al., 1999, O'Brien & Crandall, 2003), this explanation seems a bit premature unless no other convincing one exists.

Another potential explanation for the failure of the threat manipulation is that stereotype threat was present in the current experiment, as evidenced by the difference in mean test scores between groups, but that effect was not large enough to be detected given the number of participants used in the experiment. Given the fact that this experiment, after the elimination of 5 participants, only used 20 participants per between-subject condition, and given the fact that previous research has used closer to 30 or more subjects per between-subject condition (e.g. Marx & Stapel, 2006, O'Brien & Crandall, 2003, Davis, Aronson, & Salinas, 2006), it seems likely that this may have played a role in the failure of the threat effect to reach significance. Indeed, a post-hoc power analysis indicates that given the effect size observed, the study would have required approximately 80 participants to achieve significance, lending further credence to this interpretation.

One final potential explanation for the failure of the threat manipulation is that stereotype threat was present in both the threat and the no-threat conditions. Prior to participating, each participant believed that the experiment was investigating potential gender differences in math ability, a fact which may have been enough to evoke stereotype threat in and of itself. Although the math test was described as being "gender neutral" in the control condition, this may not have been enough to eliminate the threat of confirming the negative stereotype about women's math ability.

Clearly, the reasons for the failure of the stereotype threat manipulation are ambiguous. However, the picture becomes even less clear upon consideration of the lexical decision data. Because lexical decision latencies were significantly shorter after the math test, these data seem to suggest that thought suppression was present in both the threat and no-threat groups. However, because no firm link has yet been established between thought suppression and stereotype threat, it would not be logically sound to take this as disambiguating evidence to sort through the competing explanations for the failure of the threat manipulation. Indeed, there may be other reasons for this difference in lexical decision latencies, since a number of participants mentioned that the second half of the math test seemed harder than the first. This is corroborated through a paired-group t-test, which reveals that participants scored marginally significantly lower on the second half of the math test, $t(38) = 1.863$, $p = .07$. This suggests that test difficulty was confounded with time. It should be noted that although originally, the two halves of the math test consisted of unequal numbers of questions, this t-test is still legitimate because the answers to one question in the first half of the math test were excluded because of ambiguity, as noted in the results section above. Since O'Brien & Crandall (2003) have found that increasing test difficulty tends to exaggerate the difference in performance on math tasks among women in threat and no-threat groups, one could posit that the more difficult half of the math test led to greater stereotype activation, which led to the difference in lexical decision latencies. This explanation for the lexical decision latencies cannot be easily dismissed.

In sum, due to both problems in the experiment methodology and an insufficient sample size, the results of Experiment 1 are ambiguous, and further investigation is needed to clarify the data.

Experiment 2

The purpose of Experiment 2 was to disambiguate the competing interpretations of Experiment 1. Because Experiment 1 failed to find a significant difference in test scores despite using a threat manipulation that has been established in previous literature (e.g. Spencer et al., 1999, O'Brien & Crandall, 2003), one goal of Experiment 2 was to increase the power of the experiment through larger sample sizes across conditions. The problem of a lack of a significant threat manipulation will hopefully resolve itself with the resultant increase in statistical power.

The other major goal of Experiment 2 was to resolve the experimental confound of test difficulty and time. This way, differences in lexical decision latencies could be successfully attributed to differences in when the latencies were measured.

The predictions of Experiment 2 are identical to those of Experiment 1, namely that participants in the threat conditions should score lower on the math test than participants in the control condition, participants in the threat condition should show greater stereotype activation after the test than during, and that participants in the control condition should show a similar amount of stereotype activation during and after the test.

Method

Participants

The participants consisted of 46 female Macalester college students of ages 18 to 21 who were told that they were being recruited for a study researching gender differences in math ability. Of these, 15 participated for course credit, while 31 were entered into a random drawing for a $50 gift certificate to local merchants. Additionally, the data from 4 participants were excluded from the analysis, 2 because they were non-native speakers of English and 2 because they had been alerted as to the goal of the experiment prior to participation.

Materials

The materials for Experiment 2 were identical to those in Experiment 1 except that the order of the questions used in the GRE math section for Experiment 1 was randomized. Groups of questions that referred to a common graph or data set were randomized in blocks. One question from the test was also eliminated because of potential ambiguity. Finally, the test was administered to a group of 4 participants to ensure both that each half of the math test required approximately the same amount of time and that each half was perceived as being equally difficult.

Procedure

The procedure for Experiment 2 was identical to that in Experiment 1 with two exceptions. First, the order of administration for the two halves of the math test was counterbalanced across participants. Second, both to compensate for the elimination of the one ambiguous question and to increase overall test difficulty, participants were only allowed 15 minutes for each section of the math test.

Results

Manipulation Check

As in Experiment 1, the math test scores were first analyzed to ensure that the threat manipulation was effective. First, because of the large amount of variability in the test scores, any scores more than two standard deviations from the mean were excluded from the analysis. This resulted in the exclusion of three participants from the analysis.

The subsequent mean test score for the threat condition is 14.95, which is slightly lower than 16.68, the mean test score in the no-threat condition. A one-tailed independent groups t-test reveals, however, that this difference only approaches significance, $t(37) = 1.50$, $p = .143$. This implies that the threat manipulation was only partially successful in generating stereotype threat.

Stereotyped Thought Data

The stereotyped thought data consisted of reaction times to lexical decisions on 10 non-words, 5 stereotypically female target words, and 5 stereotype-unrelated filler words. Additionally, each lexical decision task could be presented either during or after the threatening math test.

Inaccurate responses were first excluded from the data, and the resulting responses were submitted to a logarithmic transformation. As in Experiment 1, the responses were then grouped by time, which could have been during or after the math test, and threat level, which could have been high or low. Means and standard deviations were then calculated according to these groups, and individual response times more than three standard deviations from the participant means were excluded from the analysis. Any resulting participant means that consisted of only two individual response times were also excluded from the analysis. The preceding steps resulted in the elimination of less than 5% of the total data, and the elimination of zero participant means. The grand means calculated by time and condition of the resulting mean participant reaction times are presented in Table 2. Note that while we used logarithmically transformed data to perform the analysis, Table 2 shows only untransformed data.

As can be seen from the table, reaction times for all conditions do not appear to differ significantly. A 2 (threat) x 2 (time) repeated measures mixed ANOVA confirms this fact, as there is no significant main effect for threat, $F(1, 39) < 1$, $p = ns$; time, $F(1, 39) < 1$, $p = ns$; or for the interaction between threat and time, $F(1, 39) < 1$, $p = ns$. Thus, participant response times did not significantly vary according to time or condition.

Discussion

Experiment 2 was designed with the twin goals of eliminating some of the competing interpretations of the data of Experiment 1 and increasing Experiment 1's overall experimental power. Experiment 2 succeeded on the first count, eliminating the experimental

confound of difficulty with time with respect to the lexical decision tasks, but failed on the second, as Experiment 2's sample size offered virtually no increase over that of Experiment 1. Additionally, the marginal significance of the stereotype threat effect in Experiment 2, combined with the universally non-significant results the lexical decision data, render it exceedingly difficult to draw firm conclusions based the results of Experiment 2.

First, although this experiment failed to find evidence of thought suppression during the math task, one may not conclude from this that thought suppression does not play a role during stereotype threat, for much the same reasons that one could not draw this conclusion based on the results of Experiment 1. Given the fact that the difference in test scores between the threat and no-threat groups was only marginally significant, the possibility remains that an increase in either statistical power or the power of the threat manipulation could result in a corresponding increase in the ease with which a potential rebound effect could be detected via the twin lexical decision tasks. Thus, as in Experiment 1, the results of this experiment do not bear directly upon the main hypotheses of this study, even though the experiment found evidence of a trend of lower test scores in the threat condition.

As for the reason that this experiment found only a trend in the direction of stereotype threat, rather than a significant difference, the three primary explanations investigated in Experiment 1, and the conclusions drawn from those investigations, still hold. Thus, it still seems premature to draw the conclusion that the threat manipulation was simply ineffective given the fact that it has been used previously in the literature (e.g. Spencer et al., 1999, O'Brien & Crandall, 2003). However, it remains possible both that claiming that the math test in the control condition was "gender neutral" was insufficient to fully eliminate the threat of confirming the negative stereotype about women's math ability and that Experiment 2 simply offered too little experimental power to detect the threat effect. Once again, a post-hoc power analysis lends some credence to the low experimental power interpretation, indicating

that around 80 participants were required to achieve significance given the observed effect size.

Finally, it is interesting to note that concurrent with the elimination of test difficulty as an experimental confound with time in Experiment 2, the differences in lexical decision latencies between the two lexical decision tasks disappeared. Although additional differences in methodology between Experiment 1 and 2 do not permit us to make a direct comparison between the two experiments, this fact still seems to suggest that more difficult tasks lead to greater stereotype activation, an idea to which we shall return in the general discussion.

General Discussion

The original motivation for this study was to establish a role for thought suppression during stereotype threat, which would further highlight the importance of the cognitive perspective in explaining the proximate processes underlying the stereotype threat phenomenon. Although both experiments in this study failed to sufficiently demonstrate a stereotype threat effect, making it difficult to speak to the original intent of the study, pooling the test score data from both experiments partially alleviates this problem. The pooled data result in a mean test score of 14.83 for the threat condition and 16.36 for the no-threat condition. A one-tailed, independent groups t-test reveals that this difference is marginally significant, $t(77) = 1.79$, $p = .082$. It is worth emphasizing that the procedures across these two experiments were different and so pooling the results of the two experiments is not strictly valid. However, changes to the experimental procedure across experiments would serve to increase the overall variability in the pooled test scores, which would render significance harder to achieve. Thus, the marginal significance of the overall mean test scores in the threat and no-threat groups provides compelling evidence that this study's difficulty in demonstrating stereotype threat lay not in a quirk of procedure or an idiosyncratic participant pool, but was rather an artifact of insufficient statistical power.

The stereotyped thought data takes on a new light in consideration of the above conclusion. Although the null results exhibited in Experiment 2 are still ambiguous in that they themselves may also be due to insufficient statistical power, the knowledge that stereotype threat may have been present below the experiments' respective detection thresholds lends some credence to the hypothesis that stereotype activation occurs at a similar (low) level during both the threatening and the non-threatening tasks. If true, this would imply that thought suppression does not play a major role in driving stereotype threat, though, as I have already argued, I am reluctant to draw this conclusion on the basis of such impoverished data.

However, the truly interesting result of this study, as mentioned in Experiment 2, was obtained almost by accident. This is the fact that increasing the difficulty of one half of the math test led to greater stereotype activation after that half, while eliminating the difference in difficulty between the two halves of the math test eliminated the disparity in stereotype activation. One must be cautious when interpreting these results, as several confounding factors vary systematically with the difficulty disparity as a result of differences in experimental methodology, and besides which, any explanation for these results will be decidedly post-hoc. However, these results do tentatively suggest a link between stereotype activation and difficulty, a link which is consistent with the thought suppression hypothesis. As O'Brien and Crandall (2003) have found, test difficulty exaggerates the stereotype threat effect, which could result from the extra cognitive resources required for more difficult tests coupled with the resources required to suppress stereotypical thoughts. Thus, more difficult tests result in a rebound effect upon completion because they trigger thought suppression, while less difficult tests do not exhibit a rebound effect because they do not trigger thought suppression. However, this explanation assumes that the participants in both experiments performed in this study treated each test as separate entities. In other words the explanation

assumes that, upon completion of the first half of the math test, the participant perceived that the math context was over, as the participants presumably did when they were taken to complete the "second study". Additionally, even if one takes this interpretation at face value, we still do not know the direction of causality between stereotype activation and the increased performance deficit exhibited during more difficult tasks. Essentially, greater stereotype activation could cause the increased performance deficit or vice-versa.

Unfortunately, we are left with essentially the same question at the end of this study as we had at the beginning, namely whether thought suppression plays a role during stereotype threat, as evidenced by a rebound effect after the threatening task. The evidence presented in this study is ambiguous at best, and the experiments performed demonstrated insufficient power to draw any meaningful conclusions. The study may have accidentally uncovered a link between stereotype activation and test difficulty, but because it was not directly testing for such a link, the interpretation of these data remains ambiguous. In the eyes of this experimenter, it remains a distinct possibility that thought suppression plays a role in stereotype threat. However, determining this fact with any certainty will be a task for future research.

References

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & 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.

Barron, K. E. & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology*, 80(5), 706-722.

Brown, R. P. & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology*, 76(2), 246-257.

Croizet, J., Després, G., Gauzins, M., Huguet, P., Leyens, J., & Méot, A. (2004). Stereotype Threat Undermines Intellectual Performance by Triggering a Disruptive Mental Load. *Personality and Social Psychology Bulletin,* 30(6), 721-731.

Croizet, J. & 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(6), 588-594.

Elliot, A. J. & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 72(1), 218-232.

Frantz, C. M., Cuddy, A. J. C., Burnette, M., Ray, H. & Hart, A. (2004). A Threat in the Computer: The Race Implicit Association Test as a Stereotype Threat Experience. *Personality and Social Psychology Bulletin*, 30(12), 1611-1624.

Gilbert, D. T. & Hixon, J. G. (1991). The Trouble of Thinking: Activation and Application of Stereotypic Beliefs. *Journal of Personality and Social Psychology*, 60(4), 509-517.

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin*, 28, 659–670.

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, 11(5), 365-371.

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(3-4), 193-198.

Keller, J. & Dauenheimer, K. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*, 29(3), 371-381.

Kray, L. J., Thompson, L., & Galinsky, A. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology*, 80(6), 942-958.

Leary, M. R. & Shepperd, J. A. (1986). Behavioral self-handicaps versus self-reported handicaps: A conceptual note. *Journal of Personality and Social Psychology*, 51(6), 1265-1268.

Macrae, N., Bodenhausen, G. V., Milne, A. B., & Jolanda Jetten, J. (1994). Out of Mind but Back in Sight: Stereotypes on the Rebound. *Journal of Personality and Social Psychology,* 67(5), 808-817.

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effects of self-affirmation and emotional expression on women's math performance. *Journal of Experimental Social Psychology*, 42(2), 236-243.

Monteith, M. J., Spicer, C. V., & Tooman, G. D. (1998). Consequences of stereotype suppression: Stereotypes on AND not on the rebound. *Journal of Experimental Social Psychology,* 34(4), 355-377.

O'Brien, L. T. & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, 29(6), 782-789.

Oswald, D. L. & Harvey, R. D. (2001). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology: Developmental, Learning Personality, Social*, 19(4), 338-356.

Quinn, D., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues*, 57(1), 55-71.

Rosenthal, R. & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, 19(1), 115-118.

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology*, 38(2), 194-201.

Schmader, T., & Johns, M. (2003). Converging Evidence That Stereotype Threat Reduces Working Memory Capacity. *Journal of Personality and Social Psychology,* 85, 440-452.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10(1), 80-83.

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review*, 16(3), 177-206.

Smith, J. L, Sansone, C., & White, P. H. (2007). The stereotyped task engagement process: The role of interest and achievement motivation. *Journal of Educational Psychology*, 99(1), 99-114.

Smith, J. L. & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles*, 47(3-4), 179-191.

Spencer, S. (2003, February). Media images and stereotype threat: How activation of cultural stereotypes can undermine women's math performance. Paper presented at the annual meeting of the Society for Personality and Social Psychology, Universal City, CA.

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

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797-811.

Stone, J. (2002). Battling doubt by avoiding practice: The effects of stereotype threat on self-handicapping in white athletes. *Personality and Social Psychology Bulletin*, 28(12), 1667-1678.

Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology*, 77(6), 1213-1227.

Wegner, D. M., & Erber, R. (1992). The hyperaccessibility of suppressed thoughts. *Journal of Personality and Social Psychology,* 63(6), 903-912.

Table 1

Mean reaction times (ms) in Experiment 1 for each condition during and after the math task.

	Threat	No-Threat	Total
During	546.24	530.66	538.06
After	511.33	507.09	509.10
Total	528.78	518.88	

Table 2

Mean reaction times (ms) in Experiment 2 for each condition during and after the math task.

	Threat	No-Threat	Total
During	538.68	533.45	536.13
After	534.81	548.07	541.28
Total	536.75	540.76	

Appendix 1: Target, filler, and non-words used for the lexical decision tasks

Target words: Gentle, expressive, romantic, warm, intuitive, caring, affectionate, sensitive, tender, emotional

Filler words: Lethal, hurried, healthy, foreign, straight, bumpy, hungry, smelly, receptive, instinctive

Non-words: Opern, unsive, poweep, stalom, wotard, rhinde, twampte, ghwurche, freuste, kningly, slockish, throod, sprinked, vaipish, kwoizlet, cleique, zursty, thrighbed, gheemel, skelicee