Running Head: STEREOTYPE THREAT AND ABILITY BUILDING

Stereotype threat impairs ability building: Effects on test preparation among women in science and technology

Markus Appel

University of Linz, Austria

Nicole Kronberger

University of Linz, Austria, and London School of Economics and Political Science, U.K.

Joshua Aronson

New York University

Preprint. Accepted for publication in the *European Journal of Social Psychology*

We would like to thank Bernad Batinic, Helena Kirchmayr, and Robert Neubauer for their support in conducting the studies. This research was supported by the TEquality project, funded by the state of Upper Austria. Nicole Kronberger received financial support within the Seventh Framework People Program (FP7/2007-2013) of the European Commission (PIEF-GA-2008-220999). This research was further supported by a National Science Foundation grant #BCS-0621356 to Joshua Aronson and support from the Spencer Foundation. Appel and Kronberger contributed equally to this work.

Corresponding author:
Dr. Markus Appel, Associate Professor, Department of Education and Psychology
Johannes Kepler University Linz
Altenberger Strasse 69, 4040 Linz, Austria; Email: markus.appel@jku.at

Abstract

Stereotype threat is an uncomfortable psychological state that has been shown to impair cognitive ability test scores. It is an open question whether and in what ways it affects processes involved in learning and knowledge acquisition. This research examined whether stereotypes also interfere with test preparation among women in the domain of science, technology, engineering, and mathematics (STEM). Study 1 (*N* = 1058) revealed that people are aware of a stereotype portraying women as less proficient in STEM-test preparation than men. Women's note-taking activities were impaired under stereotype threat (Study 2, *N* = 40), particularly when domain identification was high (Study 3, *N* = 79). Moreover, stereotype threat impaired women's performance evaluating the notes of others (Study 4, *N* = 88). Our work thus shows that stereotype threat not only hinders stereotyped individuals' capacity to demonstrate their abilities but also impairs behaviors that develop them.

Stereotype threat impairs ability building: Effects on test preparation among women in science and technology

Negative ability stereotypes can have a detrimental impact on achievements of members of stereotyped groups. Over 300 experimental studies have shown that "stereotype threat," a concern with confirming a negative stereotype, impairs performance on complex intellectual tests (Aronson & McGlone, 2009; Spencer, Steele & Quinn, 1999; Steele, 1997; Steele & Aronson, 1995; Steele, Spencer & Aronson, 2002). Thus far, however, researchers have tended to focus on test performance, and – to a lesser extent – on longer-term effects of stereotype threat on grades (Massey & Fisher, 2005) and identification with academics (e.g., Aronson, Fried & Good, 2002; Osborne, 1997; Steele, 1997). Our aim in the current research is to investigate possible effects of stereotype threat on important components of the learning process. Specifically, we focus on situations in which students prepare and revise for test-taking. We assume that activities necessary to build knowledge in a domain are less efficient when a negative group stereotype is operative. Less effective preparation may be one reason for lower achievement of African Americans, women, or other ability stigmatized group members in important domains. In the following we sketch the place of the present research within the stereotype threat framework.

Stereotype Threat

Stereotype threat is conceived as a state of psychological discomfort that, if sufficiently acute, can impair performance. It is thought to arise when students are confronted with an evaluative situation, in which a stereotype regarding a particular ability is relevant. For example, stereotype threat may occur when a woman who is aware that women are considered inferior to men at math is confronted with a mathematics test (Aronson & McGlone, 2009; Aronson & Steele, 2005; Steele, 1997; Steele & Aronson, 1995). A recent integrative framework suggests that stereotype threat arises, much like cognitive dissonance (e.g., Festinger, 1957; E. Aronson, 1968), from an aversive imbalance between specific cognitive elements that are troubling when considered together (Schmader, Johns, & Forbes, 2008). For example, a female student may enjoy and care about mathematics but, at the same time, be aware of the stereotype that portrays girls as untalented at math. Underperformance due to stereotype threat tends to be most pronounced for participants who are most identified with a domain (e.g., Aronson et al., 1999; Cadinu, Maass, Frigerio, Impagliazzo, & Latinotti, 2003; Keller, 2007; Lawrence, Marks, & Jackson, 2010). Meta-analytic findings suggest that moderate levels of domain identification may be sufficient to elicit stereotype threat in women (Nguyen & Ryan, 2008).

In recent years, a range of reactions and psychological states have been revealed as possible mediators between stereotype threat and performance (e.g., Aronson & McGlone, 2009). These include anxiety (e.g., Blascovich et al., 2001; Bosson et al., 2004), the adoption of performance avoidance goals (e.g., Brodish & Devine, 2009), reduced working memory (Schmader & Johns, 2003; Beilock, Rydell, & McConnell, 2007), and impaired self-control (Inzlicht, et al., 2006; Inzlicht & Kang, 2010). Integrating previous accounts Schmader and colleagues conceived stereotype threat as the interplay of a physiological stress response, increased monitoring of the performance situation, and the regulation of negative thoughts and emotions (Schmader et al., 2008; Johns, Inzlicht, & Schmader, 2008). All three factors consume cognitive resources that are unavailable for whatever cognitive activity a person under stereotype threat undertakes. This perspective is consistent with the frequent finding that performance decrements tend to be pronounced on complex tasks and small – or even reversed – for tasks that require less cognitive processing (e.g., Beilock et al., 2007; Huguet & Régner, 2007; O'Brian & Crendall, 2003; Steele & Aronson, 1995).

Stereotype Threat and test preparation

Most of the research on stereotype threat has focused on abilities that are thought to be developed over long periods of time and which are measured with tests like the SAT or GRE. Although performance on such tests can be enhanced through short-term preparation and training (e.g., van der Molen, te Nijenhuis, & Keen, 1996), such tests are generally thought to tap rather stable abilities. For many academic tasks, however, performance is thought to rely critically upon preparation and study. The preparation for an upcoming achievement test includes a range of activities that demand working memory capacity. Importantly, verbal working memory plays a crucial role in these activities, and verbal working memory resources appear to be especially vulnerable to stereotype threat (Beilock et al., 2007; Schmader et al., 2008).

The current research addresses preparation and training, that is, activities prior to the actual test-taking situation through which people intend to get ready for a test. We focus on behavior that typically takes place days and weeks prior to an examination. Our specific focus here is on note-taking and assessing the quality of notes. In everyday student life as well as in professional contexts, external memory aids such as notes are crucial in the communication and processing of information. Students take notes in order to record information that they will need to learn at a later date. However, the result of taking notes is much more than the production of a passive "external" information store; note-taking itself is part of the memorization process and results in the creation of a form of "internal" storage (Kiewra, 1987). Furthermore, the taking of notes seems to ease the load on the working memory and thereby helps people resolve complex problems. The quality of student notes predicts later test performance (Peverly et al., 2007); thus, factors that impede note-taking will also impede the potential knowledge measured in a subsequent ability test.

Note-taking requires distinguishing important from unimportant information and often takes place under time pressure. This makes note-taking a highly resource-consuming activity (Piolat, Olive, & Kellogg, 2005). Furthermore, for successful test preparation, the quality of learning materials must be assessed. For example, students often need to evaluate the quality of lecture notes written by other students or the quality of information found on the internet. If clear markers of high quality are unavailable, substantial cognitive effort is required to determine the suitability of information sources as learning material (e.g., Just & Carpenter, 1992). In sum, both taking notes and evaluating the quality of learning materials are important activities that consume substantial amounts of cognitive resources, and thus, we assume that the quality and efficiency of these activities can be vulnerable to resource depletion induced by stereotype threat.

Study overview

Study 1 was designed to investigate the existence of stereotypes relevant to knowledge acquisition and our population's awareness of it, a crucial prerequisite for stereotype threat (Steele, 1997). We examined stereotypes about women preparing for an examination in the STEM (Science, Technology, Engineering, and Mathematics) domain, predicting that our female participants believe that others view women as less able than men to learn in STEM domains. If so, this perceived group stereotype may elicit stereotype threat in relevant situations. Study 2 addressed the effects of the stereotype on female students' note-taking. We predicted that stereotype threat would result in lower quality notes. Study 3 replicated and extended the note-taking results, taking into account individual differences in domain identification. Study 4, investigated women's evaluation of other people's notes. We assumed that women under stereotype threat would have more difficulty distinguishing adequate from inadequate notes than would women in two reduced stereotype threat conditions.

Study 1

Study 1 was designed to assess stereotypes about knowledge acquisition in the engineering and natural sciences domain. We also investigated perceptions of shared beliefs regarding knowledge acquisition in general.

Method

Participants. Participants were 1058 German adult volunteers (580 women) between the ages of 16 and 75 (*M* = 30.3 years; *SD* = 11.7). They were recruited through a market research panel and participated online in their homes.

Stereotype Awareness. In order to investigate awareness of the stereotype positing female deficiency in STEM domains, participants were asked a question about what most people believe, irrespective of their own opinion (“What does the majority think? As compared to men (women), women (men) are better at learning natural sciences/engineering content”). A complementary second question asked about learning in general (“What does the majority think? As compared to men (women), women (men) are better at learning in general”). The order of questions as well as whether the questions were worded for men or for women was varied and randomly assigned. Participants were asked to indicate the majority’s agreement with each statement on a 7-point scale, with -3 indicating *completely disagree* and +3 indicating *completely agree*. Participants subsequently answered additional questions that were unrelated to the current study.

Procedure. Participants were invited to take part in the study through e-mail. The questions were presented online and were accessed by the participants via the web browser of their home computers. The software *EFS-survey* was used to collect the data. It further examined potential repeat responders through IP protocols (cf. Gosling, Vazire, Srivastava, & John, 2004).

Results

We tested the hypothesis that there is public awareness of a negative stereotype about female learners positing that they do not learn as well as their male counterparts in STEM domains. We did not expect to find agreement with the statement alleging that women are worse than men at learning in general. Data were recoded so that high scores represent higher ascribed learning proficiency for men. On average, participants indicated that most people assume a male learning advantage in the STEM domain ($M = 1.28$, $SD = 1.46$). This mean score is different from the scale mean zero, $t(1057) = 28.6$, $p < .001$. These judgments did not depend on participants' sex (male: $M = 1.24$, $SD = 1.46$; female: $M = 1.32$, $SD = 1.47$, $t(1056) = -0.9$, $p > .35$). Thus, both men and women report awareness of the STEM learning stereotype. In addition, participants reported that when it comes to learning in general, most people assume women are better learners than men, with $M = -1.01$ ($SD = 1.38$), which is significantly different from the scale mean zero, $t(1057) = 23.7$, $p < .001$. Women have a somewhat stronger sense of this difference ($M = -1.13$, $SD = 1.38$) than men ($M = -0.86$, $SD = 1.37$), $t(1056) = 3.1$, $p < .01$, $d = 0.19$.

Conclusion: Study 1

Study 1 confirmed our assumption that there are gender-related stereotypes which are relevant not only to testing situations but also to learning activities. Although women are expected to be good learners in general, they are expected to be less proficient than men in learning the traditionally male fields of the natural sciences and engineering. Having established the existence of this gender stereotype, Studies 2, 3, and 4 investigated its potential effects on different forms of test preparation in the STEM domain.

Study 2

Effective note-taking is an important activity in the course of ability building (Peverly et al., 2007). Study 2 was designed to examine the hypothesis that students under stereotype threat take lower-quality notes. Based on the results of Study 1, which highlighted general awareness of a gender stereotype about learning STEM contents, women were expected to perform worse when their gender had been made salient as compared to a condition in which the stereotype is invalidated.

Method

Participants. Forty female participants were recruited at an Austrian university and received 10 Euros for participation. All were students between 19 and 43 years of age (M = 24.1 years; SD = 4.6). The experiment was introduced as a study of computer-based learning. Due to the focus of this university on science, engineering, and economics, at least moderate domain identification with STEM domains among the participants was presumed.

Stereotype threat conditions. Participants were told that the first task dealt with text comprehension. The participants were randomly assigned to read one of two texts, which were presented as scientific news articles, and were instructed to answer four questions about that text. Both of the texts and their related questions were adapted from Dar-Nimrod and Heine (2006) and translated into German. The first text, titled “Women’s body in art; Women’s unique experience”, highlighted gender differences; however, it did not explicitly mention achievement-related characteristics. This text represents the *standard stereotype threat* condition in Dar-Nimrod and Heine’s studies and was intended to activate achievement stereotypes about women. The title of the alternative text was “There are no gender differences in mathematical and natural sciences abilities, researchers say.” This article, which represents the *no gender difference condition*, reported that large-scale studies found no gender differences and was meant to invalidate achievement stereotypes related to the subsequent note-taking task.[1]

Note-Taking. A list of ten keywords related to the STEM domain, which included terms such as *vulcanization* or the *photo-electric effect*, was provided. The participants were instructed to look up these keywords in an online encyclopedia and to write notes using word-processing software. Participants were asked to produce notes that were "most helpful for anyone wanting to use these notes as a basis for learning these terms." Two independent instructional psychologists (both blind to the experimental treatment) rated the quality of the notes, evaluating each of the ten topics on four 5-point-scales ranging from *intelligible* to *non intelligible*, *leaves many questions open* to *does not leave open questions*, *I could explain the topic to somebody else based on the notes* to I *could not explain the topic to somebody else based on the notes*, and *correct* to *incorrect*. To estimate inter-rater reliability, a two-way mixed intraclass correlation coefficient (ICC) was calculated for each of the four quality measures for all ten topics. Thereby the model tested for absolute agreement, which is a comparatively strict test for inter-rater concordance. Averaged across the ten topics, intraclass correlations for the four quality measures suggest satisfactory inter-rater agreement (.87 for "intelligible", .81 for "open questions", .84 for "could explain" and .81 for "correct"). For further calculations, the mean values provided by the two raters were used.

Procedure and design. The experimental sessions included groups of two to five participants.[2] Each participant was seated in front of a computer in a University computer lab. A male experimenter provided the participants with a booklet that contained the instructions for the tasks, the gender-activating news feature or the stereotype-invalidating news feature, and a brief introduction to the note-taking task. Subsequently, the participants activated the screen where an empty word-processing document and the link to an online encyclopedia were visible. The participants were asked to save the word-processing document, and to label the file with "w" for their gender (*w*eiblich in German) together with a number that was written on their paper booklet.

This booklet contained the list of the ten keywords. The time limit for the complete note-taking task was ten minutes, which was meant to exert the sort of time pressure that occurs during actual note taking (e.g., during class). After completing the note-taking task, participants worked on additional tasks unrelated to the present study, and were thanked, debriefed, and paid.[3] The design of this experiment was a one factorial between-subjects design with random assignment of participants to the stereotype-threat condition.

Results and Discussion

Participants produced notes on as few as three to as many as ten topics (*M* = 6.60, *SD* = 2.02); only three participants managed to extract information for all ten topics. Students in the standard stereotype threat condition worked on about the same number of keywords (*M* = 6.42, *SD* = 2.12) as those in the no gender difference condition, *M* = 6.76, *SD* = 1.98, *t*(38) = 0.5, *p* > .50.

Reflecting previous results on note-taking (Kiewra, 1987), quality of notes was positively related to note quantity, as indicated by the total number of words written, *r* = .64, *p* < .001. In order to address the prime hypothesis of Study 2, the quality of notes created under the two experimental conditions was compared. Participants under stereotype threat produced notes of lower quality: Average scores on the four-item quality measure, with higher scores indicating higher quality, were lower for the standard stereotype threat group (*M* = 2.13, *SD* = 0.32) than for the no gender differences group (*M* = 2.33, *SD* = 0.23), *t*(38) = 2.3, *p* = .03, *d* = 0.75. Thus, Study 2 provided evidence that female students' ability to transfer STEM information into high-quality notes is impaired under conditions that elicit stereotype threat, suggesting that their test preparation can suffer when they are confronted with threat-inducing cues.

The design of Study 2 was adapted from previous research (Dar-Nimrod & Heine, 2006); however, it is faced with nontrivial limitations. First, Study 2 does not include a neutral or control condition, instead contrasting only a no gender differences condition with a standard stereotype threat condition. Second, the manipulation used is an exception in the stereotype threat literature and can be criticized for being open to multiple interpretations. For example, one may argue that the connection between reading a text on women's bodies in the arts and a gendered perception about STEM learning is rather indirect. Given these limitations, Study 3 attempted to replicate and extend Study 2's results using a well-established stereotype threat manipulation.

Study 3

Study 3 was conducted to replicate the note-taking effect with an experimental treatment that provided more easily interpretable results. Moreover, we varied the note-taking procedure by applying an auditory stimulus in a paper-and-pencil setting. In addition, we were interested in determining whether domain identification influenced the effects on note-taking. Previous research on stereotype threat and test-taking showed that moderate to high levels of domain identification increased the detrimental effects of stereotype threat (e.g., Aronson et al., 1999; Cadinu, et al., 2003; Keller, 2007; Lawrence, et al., 2010). For example, when confronted with the stereotype that Asians are better at math, high math-identified White males were more likely to underperform than White males who were less domain-identified (Aronson et al., 1999). Less domain-identified students tended to perform even better in the stereotype threat condition compared to the control condition (see also Keller, 2007). Thus, our additional aim was to examine whether domain identification played a similar role in a note-taking task.

Method

Participants. Eighty-five female students participated for course credit for an undergraduate psychology class at a German university. Six participants did not recall the instruction correctly and were therefore excluded from further data analyses. The remaining 79 participants were between 19 and 42 years old (*M* = 22.29 years; *SD* = 3.41). Unlike Study 2, this study took place at a university with a broad academic profile, which ensured variance in STEM domain identification.

Domain identification. People who are identified with a domain are likely to be motivated to achieve; they will spend more time and effort on that particular domain, and as a consequence they can be expected to perform well. Osborne and Walker (2006), for example, found that students' identification with academics prospectively predicted their grade point average. Researchers, as a consequence, have used achievement (such as test scores) as an indicator of domain identification (e.g. Cullen, Hardison, & Sackett, 2004). Although domain identification has both a competence and an importance component (Keller, 2007; Steele, 1997), in fact, the two components will be confounded because students tend to identify with domains in which they can excel, and they will acquire competence in domains that they find important. In the following analyses we therefore use the most recent grade in physics as an indicator of domain identification. The grades ranged from 1 (very good) to 5 (not sufficient) and were recoded so that higher scores represent higher domain identification.

Stereotype threat conditions. Participants were randomly assigned to one of two conditions. All participants received a booklet, which introduced the study's topic as "learning in the STEM-domain". In the stereotype threat condition, participants were told that the purpose of the study was to examine why men perform better in such tasks than women. Subsequently, participants were ask to indicate their gender. In the control condition any reference to gender was omitted. This group was informed that the purpose of the study was to examine differences between universities (see Appendix). Participants in this condition were asked to indicate their age instead of their gender.

Note-Taking. "World of Physics" is a website hosted by the German Federal Ministry of Education and Research and the German Physical Society in order to disseminate information about physical topics to the general public. Part of this website includes podcasts produced to introduce recent scientific progress. We presented participants with a 15-minute podcast from this website. The main topic of this podcast was the formation of supernovae. The participants were asked to take notes that would be most helpful for themselves as well as for fellow students when preparing for an exam.

Two independent raters blind to the experimental treatment coded the quality of the notes on four 5-point-scales identical to the scales used in Study 2. To estimate inter-rater reliability, a two-way mixed intraclass correlation coefficient (ICC) was calculated for each of the four quality measures. Intraclass correlations for the four quality measures suggest good inter-rater agreement (.81 for "intelligible", .92 for "open questions", .78 for "could explain" and .82 for "correct"). For further calculations, the mean values provided by the two raters were used.

In addition, the original text was divided into 31 content units. For each unit a maximum number of points was determined, so that the higher the number the more details were mentioned. The two independent raters evaluated each set of notes for all 31 units. A two-way mixed intraclass correlation coefficient (ICC) was calculated for each of the sections to estimate inter-rater reliability, testing for absolute agreement. The average reliability of the mean ratings was .79, which indicates satisfactory inter-rater agreement. The number of details mentioned was summed up and the mean values of the total score of the two raters were used for further calculations.

Results and Discussion

The notes in the two experimental conditions were examined with regard to their overall quality as well as the details reproduced. Both measures were positively associated (r = .68, p <

.001) and both were correlated with the number of words written (quality rating, $r = .33$, $p < .01$; details reproduced, $r = .55$, $p < .001$).

We assumed that participants in the stereotype threat condition would produce notes of lower quality than participants in the control condition, at least when their domain identification was high. We conducted a moderated regression analysis to determine the influence of stereotype threat and domain identification. Overall quality scores were regressed on the experimental condition (control = -1, stereotype threat = 1), domain identification (continuous, *z*-standardized), and the product of both predictors. Neither the main effect for domain identification ($B = -.12$, $SE_B = 0.08$, $\beta = -.16$, $p = .16$) nor the main effect for the experimental treatment were significant ($B = .07$, $SE_B = 0.08$, $\beta = -.10$, $p = .38$). However, consistent with our assumptions, a significant interaction was found ($B = -.19$, $SE_B = 0.08$, $\beta = -.26$, $p = .03$, $\Delta R^2 = .06$). Additional analysis examined the impact of stereotype threat for participants who reported a high degree of domain identification (one standard deviation above the sample mean) and participants who reported a low degree of domain identification (one standard deviation below the sample mean, see Figure 1).

| Figure 1 |

In these comparisons, the impact of the treatment was strongest for participants who reported a high degree of domain identification ($B = -0.26$, $SE_B = 0.12$, $\beta = -.36$, $p = .03$, $\Delta R^2 = .05$). For participants who reported a low degree of domain identification, no significant treatment effects were observed ($B= 0.12$, $SE_B = 0.11$, $\beta =.16$, $p = .30$, $\Delta R^2 = .01$). We found a significant negative relationship between domain identification and note quality in the stereotype threat group (B = -0.30, SE_B = 0.13, β = -.42, p = .02, ΔR^2 = .07). Domain identification was not significantly related to the quality of the notes in the control condition (B = 0.07, SE_B = 0.10, β = .10, p = .49, ΔR^2 = .01).

A parallel analysis was conducted with the number of details found in the participants' notes as the criterion variable. There was no significant main effect of treatment (β = .01, p > .90); there was, however a trend-significant main effect of domain identification, B = -1.35, SE_B = 0.71, β = -.22, p = .06, ΔR^2 = .04. Importantly, we again found the expected interaction between treatment and domain identification, B = -1.71, SE_B = 0.71, β = -.27, p = .02, ΔR^2 = .07. Further inspection of the data revealed a trend that for participants who reported a high degree of domain identification (+ 1 SD), stereotype threat was negatively related to the number of details reproduced, B = -1.66, SE_B = 1.00, β = -.27, p = .10, ΔR^2 = .03, and a reversed trend for participants who reported a low degree of domain identification (- 1 SD), B= 1.76, SE_B = 0.97, β =.28, p = .07, ΔR^2 = .04. Domain identification was negatively related the number of details in the stereotype threat condition, B = -3.06, SE_B = 1.09, β = -.49, p < .01, ΔR^2 = .10. However, domain identification was unrelated to the number of details in the control condition (B = 0.35, SE_B = 0.09, β = .06, p = .69, ΔR^2 = .00).

Extending the results of Study 2, we found that a standard stereotype threat instruction impaired female students' ability to take notes about STEM-related information. This effect was restricted to students who were identified with the STEM domain, as indicated by above-average physics grades. This result is consistent with the literature on stereotype threat and test-taking which considers domain identification an important precondition for the experience of stereotype threat's negative effects.

Study 4

In everyday learning contexts, taking notes from lectures or from books or websites is an essential part of knowledge acquisition and a key prerequisite to success in school examinations and standardized tests. In Studies 2 and 3 we demonstrated that stereotype threat has an impact on note-taking of presented materials. Everyday learning, however, also consists of searching for relevant materials, and students must choose from a diverse range of sources that differ substantially in content quality. For example, lecture notes borrowed from one classmate may be more comprehensive than the notes from another, and websites may differ greatly in the adequacy of the information they present. The ability to identify high- and low-quality information is a key component of efficient information processing at school and at work. Therefore, in Study 4 we investigated the impact of stereotype threat on the ability to distinguish between high- and low-quality notes. Moreover, we examined the impact of the experimental treatment both on participants' judgment certainty and on the self-evaluation of their performance.

Method

Participants. Female participants of different majors were recruited at an Austrian university. The study was introduced as an investigation of "computer-based learning in the context of engineering and natural sciences". Only students with average or above-average grades in STEM subjects were accepted. Eighty-eight female students between the ages of 18 and 33 (M = 22.7 years; *SD* = 2.8) participated in the study. Students received 7 Euros for their participation.

Stereotype threat conditions. Participants were randomly assigned to read one of three different texts that supposedly introduced the study's purpose. As in Study 3, one text described the study as an investigation of the reasons why men have better learning abilities in STEM domains than women (stereotype threat condition) and one text gave no information about gender differences (control/neutral condition). The third text emphasized that, although men outnumber women in most STEM study programs, standardized tests indicate that men have worse *learning* abilities in math and science (counter-stereotype condition, Appendix). These introductions were adapted from previous studies by Aronson et al. (1999) and Beilock et al. (2007).

Stimuli. Four encyclopedia entries related to STEM topics were presented along with notes that were supposedly taken by students in order to summarize the texts. Whereas the encyclopedia entries about the four keywords (*Doppler effect*, *neutron stars*, *vulcanization*, and *Newton's laws)* were the same in all conditions, the notes were systematically manipulated. The main elements of the encyclopedia article were either correctly represented in the summary (high-quality notes) or they were missing or incorrect (low-quality notes). The validity of the quality manipulation was supported by an independent evaluation by postgraduate students of physical science. For each of the four topics, the high quality summary was evaluated as better than the low quality summary (with effect sizes of Cohen's d = 2.0, d = 2.7, d = 0.9, and d = 3.5).[4] Furthermore, the author of each summary was presented to be a male (*Maximilian*, *Lukas*) or a female student (*Johanna*, *Laura*). For each of the four topics, notes were created for all four possible combinations of gender and note quality. Employing a Latin square-design, all four gender and quality combinations were presented in a balanced order while the order of the topics remained fixed. All participants were asked to examine four encyclopedia entry-note combinations. Thus, each set of stimuli consisted of one high-quality summary attributed to a male student, one high-quality summary attributed to a female student, one low-quality summary attributed to a male student, and one low-quality summary attributed to a female student. The stimuli sets were randomly assigned to the participants.

Quality judgment. The participants' main task was to judge the quality of the summaries Participants were asked to rate on a 7-point Likert scale, with 1 indicating *not at all* and 7 indicating *very much*, how *comprehensible*, *helpful*, *correct*, and *ideal* they found each set of notes to be. Note that these four adjectives are similar to the ratings that were employed in the quality judgments by two independent raters in Studies 2 and 3. However, unlike the previous studies, it was the participants' job to evaluate the summaries. The scores for the four adjectives were averaged. The Cronbach's alpha of the quality judgment scale amounted to α = .92.

Certainty and subjective performance ratings. After evaluating each note, the participants indicated how certain they felt in their judgments using a 7-point Likert scale ranging from *not at all certain* (1) to *very certain* (7). Four additional items assessed the participants' overall satisfaction with their performance in the note judgment task ("I did well on this task"). The four items went with a seven-point scale. Cronbach's alpha of the scale was α = .79.

Procedure and design. The experiment was conducted in a computer lab at the university. Each experimental session included two to eight participants. The participants were seated in front of a computer, expecting an experiment on collaborative learning. The experimenter made a mock phone call to another university, supposedly making arrangements for the collaborative learning session soon to follow. All material was presented by the computer, which was also responsible for randomization. After a general introduction, the respondents indicated their demographics, including their gender. Subsequently, initial information about the supposed study was given, including our experimental treatment. Then the encyclopedia entry about the Doppler effect was presented along with one of the four student summaries and the four summary-evaluation items. Next, the participants were asked how confident they felt about their summary evaluation. This procedure was repeated for the remaining three encyclopedia entries. In a final section, participants' satisfaction with their own performance was assessed. At this point, the participants believed that the collaborative learning session was their next task; however, the experiment ended here. The participants were thanked and debriefed.

The experiment followed a 3 x 2 x 2 mixed design with the stereotype-threat manipulation (stereotype activation vs. neutral vs. counterstereotypic information) as a between-subjects factor and the summaries' quality (low vs. high) and the summary authors' gender (female vs. male) as repeated measures.[5]

Results and Discussion

Quality judgments. In this experiment, our main dependent variable was the respondents' judgments about the summaries, which varied in quality and author gender. We expected a poor distinction of high- versus low-quality summaries by participants who experienced stereotype threat. A GLM with repeated measures on two of the factors (author's gender and note quality) and nonrepeated measures on the stereotype threat factor was calculated. As expected, high-quality summaries were judged as more adequate ($M = 5.18$; $SD = 0.82$) than low-quality summaries ($M = 4.15$; $SD = 1.28$). This main effect was significant in the GLM, F (1, 85) = 39.4, $p < .001$, $\eta^2 = .32$.

| Figure 2 |

The core assumption guiding experiment 4 was that stereotype threat would reduce women's ability to distinguish high from low quality notes. As expected, the main effect of summary adequacy was qualified by an interaction with the stereotype threat factor, F (2, 85) = 3.6, $p < .05$, $\eta^2 = .07$. The results are depicted in Figure 2. No other main effect or interaction effect reached statistical significance, all Fs < 1, all ps > .25. Of note, in all three conditions, the author's gender did not influence the quality judgments.

Subsequently, differences between low-quality and high-quality texts were analyzed for the three experimental conditions separately. The simple main effects indicate that participants in the counter-stereotype condition were able to distinguish between low-quality and high-quality notes, F (1, 85) = 25.8, $p < .001$, $\eta^2 = .23$, as were participants in the neutral/control condition, F (1, 85) = 26.0, $p < .001$, $\eta^2 = .23$. Under stereotype threat, however, the actual quality of the notes did not significantly influence the quality judgments, F (1, 85) = 2.0, $p = .16$.

Subjective task performance and certainty. Although participants in the stereotype threat condition performed worse when evaluating the quality of the summaries, the experimental groups did not differ in their subjective task performance ratings, which were obtained at the end of the study. Women in the stereotype threat condition rated themselves as effective (M = 4.23, SD = 1.23) as those in the counter-stereotype group (M = 4.23, SD = 1.25) and the neutral group (M = 4.44, SD = 1.08), F (1, 84) = 0.4, $p = .75$. Likewise, participants in the stereotype threat condition did not differ in their certainty ratings in comparison to participants in the other experimental groups. After each quality rating, participants were asked to indicate their subjective certainty in giving the judgment. The participants felt more certain when judging high-quality notes than when judging low-quality notes, irrespective of the author's gender (female, low quality, M = 4.06, SD = 1.11; male, low quality, M = 3.97, SD = 1.10; female, high quality, M = 4.42, SD = 0.98; male, high quality, M = 4.35, SD = 0.98; F [1, 85] = 15.9, $p < .001$, $\eta^2 =$ 0.16). Neither the between-subjects factor nor any of the interactions reached statistical significance (all Fs < 1.1). In sum, when the negative group stereotype was activated, women failed to distinguish between low- and high-quality information. However, this inefficiency did not affect reported performance satisfaction or judgment certainty, suggesting that the women were not aware of the detrimental effects caused by stereotype threat.

It should be noted that the women in the counter-stereotype condition did not differ from women in the neutral/control condition in their ability to differentiate high-quality from low-quality notes. At this point, the meaning of this finding cannot be fully clarified. One possible explanation is that women in the counter-stereotype condition are so aware of the existing gender stereotype that they do not fully believe the instruction that women do better than men on these tasks. Future studies should provide additional data to control for this possibility.

General Discussion

The impact of stereotype threat on ability building

African American students tend to score lower on general cognitive tests and women tend to score lower on tests in the science, technology, engineering, and mathematics-domain, even if stereotype threat during test taking is accounted for. Stereotype threat during testing does not explain all of the variance between ethnic groups or men and women. The aim of this paper was to shed light on a largely overlooked source of variance: stereotype threat during preparation and learning tasks. In Study 1 we demonstrated that there is indeed general awareness of a gender-related stereotype in the context of STEM learning, which is a precondition for the potential occurrence of stereotype threat (Steele, 1997; see also Maass, D'Ettole, & Cadinu, 2008). In our subsequent experiments we tested the effects of stereotype threat on test preparation, and more specifically on note-taking and note evaluation. As expected, the quality of test preparation was impeded in the stereotype threat condition. If stereotype threat also impairs learning activities (at least among those who are domain identified), then, over time, targets not only will demonstrate impaired test performance but will actually learn content in less efficient ways as well. Gradually the knowledge gaps between targets and nontargets will widen.

The present studies thus add to the evidence that stereotype threat is not only a phenomenon which impacts on ability measurement but impedes the acquirement of ability and knowledge. In one of the rather rare preceding studies that addressed the impact of stereotype threat on task preparation, Stone (2002) investigated preparatory behavior in a non-academic domain. In two experiments, European Americans were instructed that an upcoming golf-putting task assessed natural athletic ability (stereotype threat, European American background is associated with low natural athletic ability). European Americans who received a low-threat instruction or Hispanic Americans served as the control group. All participants were allowed to practice the forthcoming golf-putting task. As expected, participants under stereotype threat practiced less than the control participants, given that their self-worth was closely related to their performance. Thus, stereotype threat affected preparation necessary for good test results.

Stone's (2002) and our data on preparation complement previous findings on stereotype effects prior to test-taking, including task choice, long-term aspirations, and domain identification. Women under stereotype threat chose to work on verbal items as compared to math items (Davies, Spencer, Quinn, & Gerhardstein, 2002; Study 2), and women who saw stereotypic TV ads preferred a submissive role over a leadership role in an upcoming problem solving task (Davies, Spencer, & Steele, 2005). Another way to avoid experiencing the discomfort of stereotype threat is to choose simple rather than complex and challenging tasks. In one study, when girls had the choice between an easier, an appropriate, or a very challenging task, girls who thought the tasks prompted mathematical abilities more often chose the easier problems to solve than did girls in the control group (Good & Aronson, 2001, in Aronson, 2002). Extending stereotype threat to long-term aspirations, Davies and colleagues (2002; Study 3) asked women about their educational and vocational preferences. Those women who saw the gender-stereotyped TV ads had less interest in quantitative domains (e.g., mathematics, engineering, or physics), and preferred more verbal domains (e.g., communications, or authoring novels). This reaction may reflect weakening ties between the self-concept and the stereotyped STEM domain. The situational detachment of self and domain may serve as a way to lower the cognitive inconsistency that elicits stereotype threat. Hence, situational disengagement may reduce stereotype threat. As a consequence, stereotyped individuals who temporarily distance themselves from a task show higher persistence and motivation (Nussbaum & Steele, 2007). In the long run, however, episodes of situational disengagement may lead to a chronic detachment of the self from the domain, i.e., disidentification (Crocker, Major, & Steele, 1998; Steele, 1997; Steele et al., 2002), which is likely to result in low achievement in respective fields.

Limitations and research perspectives

Whereas the majority of research on stereotype threat in the last few years has focused on mediating and moderating variables of stereotype threat during test-taking, we focused on the scope of stereotype threat effects. Some (Cullen, Waters & Sackett, 2006; Sackett et al., 2004; Stricker & Ward, 2004) have questioned the real-life validity of stereotype threat. We examined stereotype content in the learning domain (Study 1) and investigated stereotype-threat effects in task-preparation activities (Studies 2-4). Our data suggest that when the power of the stereotype threat as an explanation for group differences is discussed, its detrimental effects during times of task preparation and revision need to be considered. We built on previous work regarding mediation which was not at the heart of our present studies. Future research may profit from an explicit consideration of mediation processes in the learning context, including measures of cognitive load, motivation, and emotion (e.g., Sweller, 1988; Pekrun, Elliot, & Maier, 2006; cf. Ryan & Ryan, 2005).

The results of Study 1 suggest that stereotype threat may be a relevant factor in male learning and ability building as well. In this study our mixed sample of participants indicated that men are perceived as less efficient than women in general knowledge acquisition. Although male performance was not a focus of the research presented here, this finding is relevant as this negative stereotype may impair boys' preparatory activities and performance in non-STEM domains. Male students' underperformance in school achievement relative to their female peers has troubled the public and attracted scholarly attention (e.g., Conger & Long, 2010). However, no such gender differences, or even better scores for male students than female students, are reported for standardized tests (Halpern et al., 2007). Stereotype threat could play a crucial role in explaining this paradox. Stereotype threat may impair female performance in test-taking situations when negative achievement stereotypes are activated regarding domains such as STEM, political science, or history (e.g., McGlone, Aronson, & Kobrynowicz, 2006). Stereotype threat is a mechanism that affects male majority members like all other groups when a relevant negative stereotype is activated (e.g., Aronson et al., 1999; Koenig & Eagly, 2005). If, as our data indicate, male students are perceived as less effective in preparation activities and are ascribed less effort, their test preparation or homework is likely to be impaired. An extra pressure not to fail may lower working memory capacity and may increase avoidance motivation at times of preparation and revision, which impedes effective knowledge building. As a result, male students may deliver lower-quality homework and perform worse on school achievement tests than female students and than their basic aptitude would predict. To support these considerations on stereotype threat's effects on male students during times of knowledge building, more research is needed on stereotypic beliefs regarding not only ability but effort as well.

References

Aronson, E. (1968). Dissonance Theory: Progress and problems. In R.P. Abelson, E. Aronson et al. (Eds.), *Theories of Cognitive Consistency: A Sourcebook* (pp. 5-27). Chicago: Rand McNally.

Aronson, J. (2002). Stereotype threat: Contending and coping with unnerving expectations. In J. Aronson (Ed.), *Improving academic achievement* (pp. 279-301). San Diego: Academic Press.

Aronson, J., Fried, C. & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38*, 113-125.

Aronson, J., Lustina, M., 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.

Aronson, J., & McGlone, M. S. (2009). Social identity and stereotype threat. In T. Nelson (Ed.), *Handbook of Stereotyping, Prejudice, and Discrimination Research* (p. 153-178). New York: Psychology Press.

Aronson, J. & Steele, C.M. (2005). Stereotypes and the fragility of human competence, motivation, and self-concept. In C. Dweck & E. Elliot (Eds.), *Handbook of Competence and Motivation.* New York: Guilford.

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

Bosson, J. K., Haymovitz, E. L., & Pinel, E. C. (2004). When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety. *Journal of Experimental Social Psychology, 40*, 247-255.

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spill over. *Journal of Experimental Psychology: General, 136, 256-276.*

Brodish, A. B. & Devine, P. G. (2009). The role of performance-avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal of Experimental Social Psychology, 45*, 180-185.

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology, 33*, 267-285.

Conger, D. & Long, M.C. (2010). Why are men falling behind? Gender gaps in college performance and persistence. *Annals of the American Academy of Political and Social Science, 627*, 184-214.

Crocker, J., Major, B., & Steele, C. M. (1998). Social Stigma. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 504-553). New York: McGraw Hill.

Cullen, M. J., Hardison, C. M., & 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*(2), 220-230.

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

Dar-Nimrod & Heine (2006). Exposure to scientific theories affects women's math performance. *Science, 314*, 435.

Davies, P. G., Spencer, S. J., Quinn, D. M., & 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.

Davies, P. G., Spencer, S. J., & Steele, C. M. (2005). Clearing the air: Identity safety moderates the effects of stereotype threat on women's leadership aspirations. *Journal of Personality and Social Psychology, 88*, 276-287.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford: Stanford University Press.

Gosling, S. D., Vazire, S., Srivastava, S. & John, O. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*, 93-104.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The Science of Sex Differences in Science and Mathematics. *Psychological Science in the Public Interest, 8*(1), 1-51.

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

Inzlicht, M., Aronson, J., Good, C., & McKay, L. (2006). A particular resiliency to threatening environments. *Journal of Experimental Social Psychology*, *42,* 323–336.

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*, 365-371.

Inzlicht, M. & Kang, S. K. (2010). Stereotype threat spillover: How coping with threats to social identity affects, aggression, eating, decision-making, and attention. *Journal of Personality and Social Psychology*, 99, 467-481.

Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General, 137*, 691-705.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review,98,* 122–149.

Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology, 77*(2), 323-338.

Kiewra, K. A. (1987). Notetaking and review: The research and its implications. *Instructional Science, 16*, 233-249.

Koenig, A. M., & Eagly, A. H. (2005). Stereotype threat in men on a test of social sensitivity. *Sex Roles, 52,* 489-496.

Lawrence, J. S., Marks, B. T., & Jackson, J. S. (2010). Domain identification predicts black students' underperformance on moderately difficult tests. *Motivation and Emotion, 34*, 105-109.

Maass, A., D'Ettole, C., & Cadinu, M. (2008). Checkmate? The role of gender stereotypes in the ultimate intellectual sport. *European Journal of Social Psychology, 38*, 231-245.

Massey, D. S., & Fischer, M. J. (2005). Stereotype Threat and academic performance: New data from the national longitudinal survey of freshman. *The DuBois Review: Social Science Research on Race*, 2, 45-68.

McGlone, M. S., Aronson, J., & Kobrynowicz, D. (2006). Stereotype threat and the gender gap in political knowledge. *Psychology of Women Quarterly, 30*, 392-398.

van der Molen, H.T., te Nijenhuis, J., & Keen, G.W. (1996). The effects of intelligence test preparation. *European Journal of Personality, 9*, 43-56.

Nguyen, H. H., & Ryan, A. M. (2008). Does stereotype threat affect cognitive ability test performance of minorities and women? A meta-analytic review of experimental evidence. *Journal of Applied Psychology, 93*, 1314-1335.

Nussbaum, A. D., & Steele, C. M. (2007). Situational disengagement and persistence in the face of adversity. *Journal of Experimental Social Psychology, 43*, 127-134.

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

Osborne, J. W. (1997). Race and academic disidentification. Journal of Educational Psychology, 89, 728-735.

Osborne, J. W., & Walker, C. (2006). Stereotype Threat, Identification with Academics, and Withdrawal from School: Why the most successful students of colour might be most likely to withdraw. *Educational Psychology, 26*(4), 563-577.

Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology, 98*, 583-597.

Peverly, S. T., Ramaswamy, V., Brown, C., Sumowski, J., Alidoost, M., & Garner, J. (2007). Skill in Lecture Note-taking: What Predicts? *Journal of Educational Psychology, 99,* 167-180.

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology, 19,* 291–312.

Rheinberg, F., Vollmeyer, R. & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [QCM: A questionnaire to assess current motivation in learning situations]. *Diagnostica, 47,* 57-66.

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist, 40,* 53-63.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *American Psychologist, 59*, 7-13.

Schmader, T., & Johns, M. (2003): Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85*, 440-452.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115,* 336-356.

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. M. (1997): A threat in the air: How stereotypes shape intellectual ability and performance. *American Psychologist, 52*, 613-629.

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

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 34* (pp. 379-440). San Diego, CA: Academic Press.

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

Stricker, L. J., & 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.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.

Endnotes

[1] The original English-language versions of the texts can be found at http://www.sciencemag.org/cgi/content/full/sci;314/5798/435/DC1).

[2] Along with our female participants, a smaller number of male learners were invited in order to make gender more salient. Prior research has shown that the presence of male participants enhances the effect of the experimental treatment for women (Inzlicht & Ben Zeev, 2000). Mixed-gender group composition was used in both experimental conditions.

[3] In addition, the material included a self-report measure of situational motivation (Questionnaire on Current Motivations, QCM, Rheinberg, Vollmeyer, & Burns, 2001). Results of the motivation self-reports were unrelated to the experimental treatment. This reflects previous research which showed that self-reported feelings and motivations are unaffected by stereotype threat (e.g., Stone, 2002).

[4] Post-graduate physical science students ($N = 15$), blind to the experimental manipulation, received a booklet that contained each encyclopedia article along with the high- and the low-quality summary (in counterbalanced order). They performed the four-adjective quality rating that was also used in Experiment 4 (*comprehensible*, *helpful*, *correct*, and *ideal*). Paired-samples t-tests revealed that the evaluations were significantly more positive for high vs. low quality summaries, *t*s > 2.5, *p*s < .05. In addition the science students were to choose which of both summaries was preferable. Separate analyses of the four topics revealed that 12 or more of the 15 evaluators opted for the high quality summary (exact binomial test, *p*s < .05).

[5] Three additional two-item self-report measures on motivational states (confidence, task-engagement, and concern) were applied before the main task and at the end of the experiment. However, self-reported motivation failed to mediate the treatment effect as motivation was unrelated to indexed note quality discrimination scores.

Figure caption

Figure 1. Note-taking quality as a function of stereotype threat and domain identification (Study 3)

Figure 2. Participants' quality judgments as a function of note quality and stereotype threat (Study 4)

Figure 1.

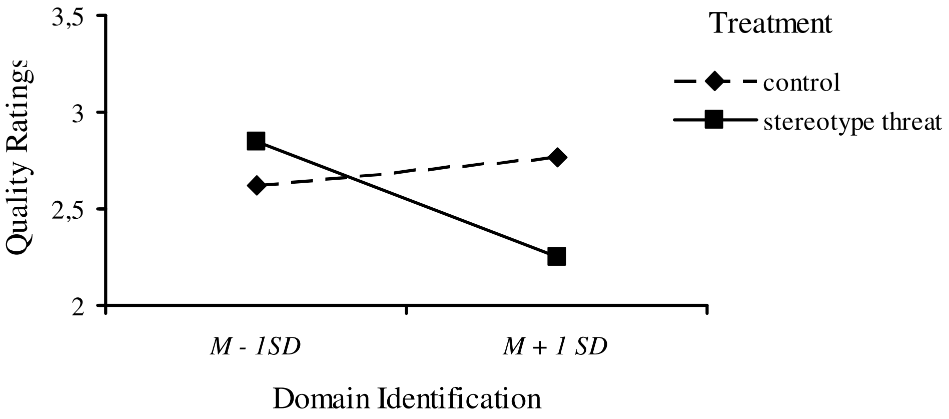
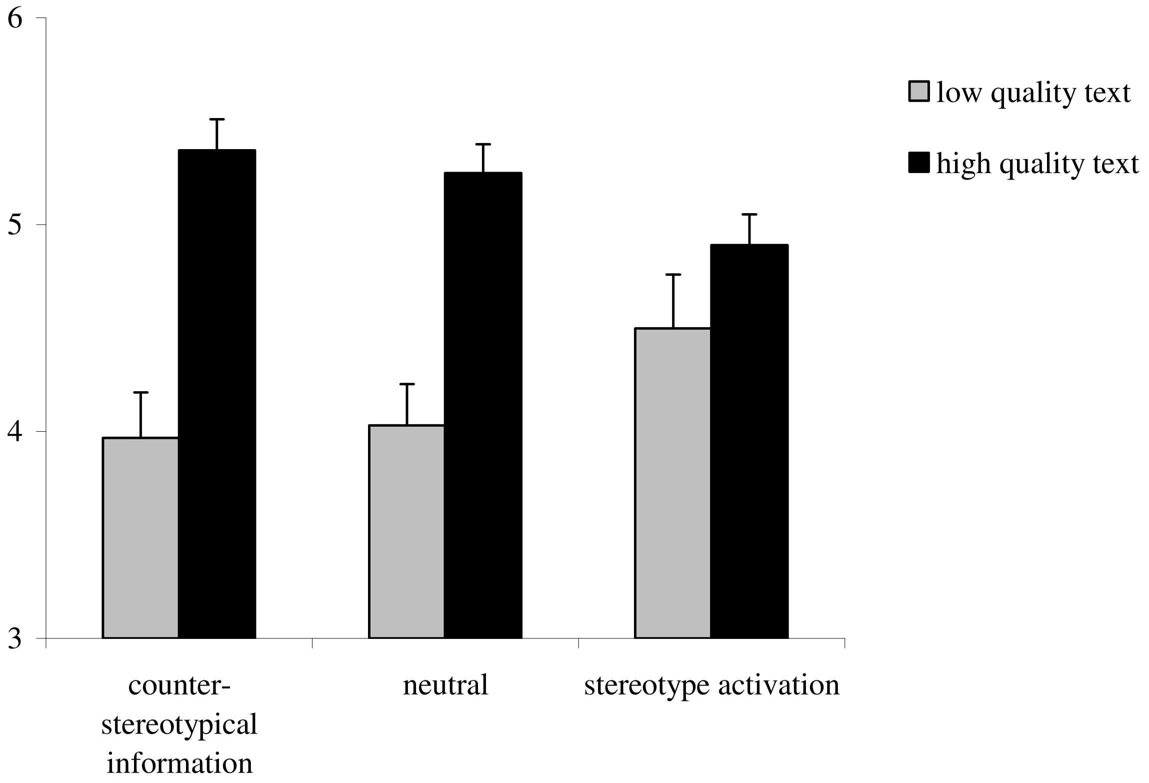

3,5
3
2,5
2
Quality Ratings
M - 1SD
M + 1 SD
Domain Identification
Treatment
control
stereotype threat

Figure 2.


6
5
4
3
counter-stereotypical information
neutral
stereotype activation
low quality text
high quality text

Appendix: Stereotype Threat Manipulation in Experiments 3 and 4

The instructions were adapted from Aronson et al. (1999) and Beilock et al. (2007) and they were presented in German.

All participants read (Experiments 3 and 4):

"We are interested in learning and information processing in the science domain for a reason. As you probably know, math- and science-related cognition is crucial to performance in many important subjects in college. Yet surprisingly little is known about the mental processes underlying cognitive abilities in math and science. This research is aimed at better understanding what makes some people process information in the math and science domain better than others."

Control group also read (Experiments 3 and 4):

"Your performance today in the science domain will be compared to students from other places of study."

Stereotype threat group also read (Experiments 3 and 4):

"As you also may know, at most programs of study in the fields of mathematics, science and engineering, male students outnumber female students. A good deal of research indicates that men outperform women on standardized tests that assess learning abilities in math and science. But thus far, there is not a good explanation for this. The research you are participating in is aimed at better understanding these differences."

Counter-stereotype group also read (Experiment 4):

"As you also may know, at most programs of study in the fields of mathematics, science and engineering, male students outnumber female students. However, differences in performance seem to decrease. A good deal of research indicates that men obtain lower scores in standardized tests that assess *learning* abilities in math and science (!). But thus far, there is not a good explanation for this. The research you are participating in is aimed at better understanding these differences."