universität wien

MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

„Johns, Schmader, and Martens (2005) – a replication“

verfasst von / submitted by

Lorraine Angelika Toni Phillips (BSc)

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie UG2002
Betreut von / Supervisor:	Ass.-Prof. Mag. Dr. Jakob Pietschnig

Introduction

Over the course of a formal education in psychology, every student accumulates a wealth of knowledge pertaining to various psychological effects as they are documented in the relevant textbooks, in particular the ones on social psychology. However, in recent years, it has become more and more apparent that many of the effects the scientific community had already considered established facts of sorts may not be all that set in stone, after all. This concern has come to be known as the widely discussed replication crisis, after several failed attempts at replicating many of these supposedly well-known effects. As such, it is vital to the integrity of the scientific community that it comes to understand and report on what factors have led to this very crisis.

To gain perspective on the matter, one can turn to some of the large-scale replication efforts that have been conducted in recent years. One such effort is the Reproducibility Project: Psychology (RPP) by the Open Science Collaboration (2015). This project comprised one hundred replications of both experimental and correlational studies that had been selected for this purpose from three peer-reviewed psychological journals. The project employed a high-powered design to account for the possibility that the effects – should they exist – may not be as large as originally reported, and thus require a bigger sample for them to be successfully detected. Original materials were used when available, allowing for a more faithful replication. While 97 of the included original studies had revealed significant effects ($p < .05$), this was the case for only 36 of the replications. Significance is not everything, of course, but even when examining the individual effect sizes, a clear pattern emerged: The replicated effects ($r = .19$, $SD = .26$), on average, were around half the size of the ones reported in the original studies ($r = .40$, $SD = .19$). Reviewing these numbers, one can scarcely deny the impact the replication crisis has had on the very foundation of the psychological community – though it should be noted that other sciences, too, have been shown to be subject to this effect: Out of 53 papers which were considered "landmark" studies in the field of cancer biology, only 6 (11%) could be successfully replicated (Begley & Ellis, 2012).

In light of its importance, many have sought to understand the root of the crisis. One of the earliest papers to address this was an essay by Ioannidis (2005), which is now considered foundational to the field of meta-science, because it took a critical look at the scientific process from various angles to discern possible sources of the problem. One of the discussed causes is the so-called publication bias (Ferguson & Brannick, 2012; Turner et al., 2008), which accounts for the fact that it is – or at least, was – predominantly research with significant results that ends up being published in peer-reviewed journals.

The enormous pressure from within the scientific community to publish one's own research appears to be contributing to this phenomenon, because it encourages scientists to deliver “publishable” – meaning, statistically significant – results, so that they may further their personal careers (Fanelli, 2010). The consequence of this is a distortion of the perceived overall results – if nine out of ten papers failed to find the desired effect, but only the one with a significant result is published, the public is likely to overestimate the existence of said effect.

Another contributing factor to the replication crisis can be found in the law of large numbers (Bolthausen & Wüthrich, 2013), which states that the average of a large sample that was randomly selected from a population will be close to the average of said population. Smaller samples may on occasion produce significant results with big effect sizes, but they may not necessarily be an accurate representation of the population as a whole. Just like with publication bias, relying on research with small sample sizes can lead to an overestimation of psychological effects. And if an effect was never truly as impactful as assumed due to previous literature, it should come as no surprise that modern replication efforts with large samples fail to produce evidence of a similar-sized effect.

A suboptimal statistical approach, too, can lead to the overestimation of psychological effects – see below for an example involving ANCOVA models. The consequence of such an overestimation is an inflated effect size in the original report, which may lead to an inadequate power analysis for the replication, making it unlikely to detect the smaller-than-expected effect due to an insufficient sample size.

Also worth noting in the context of the replication crisis are the so-called researcher degrees of freedom (Simmons et al., 2011), which refer to the many different approaches a researcher may take in designing and conducting a scientific experiment. After all, there is often no one correct method for operationalising the relevant outcome or analysing the collected data – rather, there typically exist multiple unique methods that can be chosen from. This choice is ultimately left to the discretion of the conducting researcher, but if they fail to document the specifics of their study with sufficient detail, an attempted replication might fail simply because its scientific approach differs from that of the original study. Furthermore, the original researcher may have chosen a specific method precisely because it – as opposed to others attempted – yielded a significant result, making it less likely for any other method to produce a similar result in replication. This mentality of searching for a method that "works" – meaning it produces a significant effect – has been shown to lead to an increase in false-positive rates in psychological research (Simmons et al., 2011).

Now that some of the underlying causes of the replication crisis have been identified, the question arises of what can be done to mitigate its damage. One approach to this – and the one the present paper will be focusing on – is to put more effort into improving upon the quality of replication projects, so that the scientific community may receive quasi-definitive answers on whether or not a specific effect currently exists within a designated population. To achieve this, the concept of Registered Replication Reports (RRR) was developed, which significantly improves upon simple replications in many ways (Association for Psychological Science, 2021). Each RRR consists of a multitude of smaller, independent replications that all follow the same experimental protocol. This protocol is published in scientific collaboration forums such as the Open Science Framework (OSF) in advance, before any replications begin to take place, and gives a detailed account of how the participating labs are to conduct themselves at every step during their individual replications. This ensures high reliability in the design and that any differences in results may be attributed to genuine fluctuations in the observed effect, rather than to mere discrepancies in scientific procedure. If possible, the authors of the original study are also involved in the planning process and may grant access to their original materials, allowing for an even more faithful replication.

But this well-thought-out design protocol is not the only strength of RRRs – another stems from the sheer number of participating labs. When pooling together the participants of all individual replications, the ultimate analysis benefits from a high-powered design, making it likely to detect even small effects that are present within the overall sample. Additionally, the pooled sample has the potential to be far more culturally diverse than that of any one replication, in the event that the contributing labs are from different countries. As such, the results may be generalised more broadly, and cultural factors can be considered. By conducting a meta-analysis on all the smaller replications, it is also possible to achieve an estimate that is closer to the true effect sizes in the population. Another benefit of RRRs is that they take great care in adhering to open science practices: the protocol, used materials, and analysis scripts are all published online before data collection even commences, and the raw data files are ultimately made available to the public, as well. Because of this, it is easy to track any adjustments to the original protocol that may occur over the course of a project. Also, because all the necessary materials have already been curated for the pre-registration, journals may review the project and decide whether they wish to publish it, regardless of the results it will eventually go on to produce. Once a journal has formally accepted an RRR, this can serve as a great incentive for participating labs: In the case that they are able to acquire a sample whilst adhering to the pre-published protocol, they are guaranteed to be included as co-authors in the following publication. As such, there is none of the usual concern that non-significant results may be perceived as uninteresting and thus not be accepted for a publication.

It is precisely such an RRR project that the present paper seeks to contribute to. The proposing authors of the specific project are Stoevenbelt et al. (2021), and all details on their RRR may be viewed in the OSF (https://osf.io/zwnxy/). The paper they chose to make the subject of this replication effort is *"Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance."* by Johns et al. (2005). This original study was chosen because it had a simple and clear design, making it easy to faithfully replicate, and because it was largely representative for other, similar studies pertaining to stereotype threat – a phenomenon which will subsequently be discussed in greater detail. Another benefit lies in the fact that this study is still relatively recent, meaning the authors are all still accessible and could be involved in the process of creating the RRR protocol. Also, the paper is well-known and thus influential to the field of stereotype threat, with Google Scholar showing 683 citations at the time of this writing.

Before the specifics of the original study by Johns et al. (2005) can be discussed, one must first be made familiar with the concept of stereotype threat itself. While sex differences in the mid-level mathematical performance range are not as prominent, a significant distinction has been identified in higher mathematical performance ranges – specifically, more men than women are represented in this higher range (Halpern et al., 2007). In an attempt to understand why this performance gap exists, many different factors have been considered, resulting once more in a debate regarding the influences of nature versus nurture. On the nature side, biological influences such as hormones and genetics have been discussed (Geary, 2010), while those considering nurture have more so looked to socio-psychological factors for an explanation. One such explanation is the stereotype threat effect, first proposed by Steele and Aronson (1995). The idea behind this effect is that minorities do not perform more poorly on performance tests due to any inherent lack in skill, but rather because of their awareness of existing societal stereotypes regarding their abilities in a specific field. They fear – either consciously or sub-consciously – that if they receive a bad result on a relevant test, it will reflect poorly on other members of their already stigmatised group. This fear of "adding to the problem", so to speak, can cause great internal turmoil, and such anxiety has been shown to be detrimental to more complex cognitive tasks (Hembree, 1990). The same theory can be applied to women's performance in mathematics: A woman might feel enormous pressure to perform well on a maths test, or otherwise contribute to the continued stereotype that all women as a group have no aptitude for mathematics – and this perceived pressure could in turn impede her performance.

Of course, stereotype threat – like all psychological effects – is subject to the influences of other factors that must also be taken into consideration. One such factor is test difficulty, which has been shown to act as a moderator for stereotype threat (Nguyen & Ryan, 2008). Specifically, threatened minorities performed more poorly than control groups when the administered tests were difficult (d = 1.43), rather than when they were only moderately difficult (d = 1.18). The proposed reasoning for this is that a more difficult test will lead to a higher mental workload, making it easier for stereotype threat to interfere with cognitive performance (Steele & Aronson, 1995).

Another frequently discussed factor is the sex composition of the groups in which testing occurs. While some studies have found no evidence of sex composition acting as a moderator for stereotype threat (Doyle & Voyer, 2016), others have been able to demonstrate that more males being present can lead to poorer performance in women (Inzlicht & Ben-Zeev, 2000). Women's deficits were shown to be proportional to the number of males in a group, and even just a single male being present led to a significant decline in performance. As such, sex composition may have an influence on stereotype threat effects, but further research on the interaction is still needed.

Before discussing stereotype threat in terms of replication, one must first examine the existing evidence on whether this effect can truly help explain the observed sex differences in the mathematical domain. One study evaluated 54 students from Michigan (30 female / 24 male) in a 2x2 experimental design with sex and test characterisation as its factors (Spencer et al., 1999). Each student received two difficult maths tests – one that was meant to establish a threatening setting for the female examinees by introducing the test as one that typically produces sex differences, and a more neutral alternative, where they were told that men and women tend to perform equally on the exam. Subjects were randomly assigned to which test they would complete first. In line with the expectations of stereotype threat, a significant interaction between sex and test characterisation was discovered: While there was no notable difference between men and women in their performance on the neutral test, men significantly outperformed women in the sex-specific version. Many other studies have also demonstrated such a stereotype threat effect (e.g., Stone et al., 1999; Taylor & Walton, 2011), indicating that stereotype threat might well be a suitable explanation for at least some of the observed sex differences.

However, the current state of evidence is not quite as clear as these studies might suggest, because the stereotype threat effect does not seem to be exempt from the previously outlined replication crisis. A meta-analysis (Stoet & Geary, 2012) examined 19 different experimental studies that had been deemed replications of the original stereotype threat effect reported by Spencer et al. (1999). Of these 19 replications, only 55% produced a statistically significant ($p < .05$) stereotype threat effect. Even so, approximately half of the data had been confounded by statistically adjusting for previous maths grades of the test subjects. When only taking the non-confounded studies into account, the success rate of the replications was reduced to a mere 30%. As such, while stereotype threat might still play a role in mathematical sex differences, its influence should be considered with caution, until the time that more definitive research in this area has been conducted.

Assuming that stereotype threat is in fact a contributing factor to women scoring more poorly than men in higher performance ranges, it is vital that the scientific community identifies methods capable of decreasing its influence, thus improving women’s overall performance in mathematics. One study demonstrated that introducing positive female role models can help buffer women’s test performance (Marx & Roman, 2002) – specifically, it was shown beneficial to have a woman, who was presented to the students as a mathematically competent female experimenter, administer the test. Another approach is to place women in an all-female test setting, with some studies finding that even just a few males being present can lead to performance deficits in mathematics for women (Inzlicht & Ben-Zeev, 2000). Yet another study found promising results when asking African American students to think of intelligence as a malleable, rather than a fixed character trait (Aronson et al., 2002), leading to them achieving higher college grade point averages than control groups that did not receive this instruction. Finally, being able to attribute one's internal state of distress to external, situational factors has also been shown to enhance performance (Brown & Josephs, 1999). Examinees might otherwise assume that their anxiety was brought on by their own lack of ability, causing yet further distress. The perhaps most practical application of this is to simply educate women on stereotype threat as a psychological phenomenon, as it is easy to implement and the benefits of doing so could potentially extend beyond the specific test situation.

It is just this approach of stereotype threat education for women in mathematics that Johns et al. (2005) opted to investigate in their original study. While the proposed benefit of such education lies in women being able to attribute their internal distress to stereotype threat, rather than their own lack in arithmetic skill (Brown & Josephs, 1999), it is just as likely that introducing them to the concept of stereotype threat before an exam could merely lead to higher stereotype activation and thus to stereotype-conform behaviour (Wheeler & Petty, 2001). As such, the authors of the original paper deemed further investigation into the matter necessary. For this purpose, they recruited 144 participants from an entry level statistics course, with 117 students (75 female / 42 male) being included in the final analysis. This final sample was comprised of exclusively Caucasian participants, as some prior studies had only been able to find sex differences for this group (Hyde et al., 1990).

Their study followed a 2x3 between-subjects design, with sex and experimental condition as its respective factors. The experimental conditions differed by how the administered arithmetic test was presented to the students, and when they were asked to make note of their own sex. The condition which was meant to be most threatening to female students was the “*Stereotype Threat*” (ST) condition, in which the test was introduced as a quantitative exam, and students were asked to write down their sex before they began solving the items. This was done because previous research had shown that being confronted with one's affiliation to a stigmatised group, as well as being told that a test is diagnostic in a stereotype-relevant domain both lead to poorer performance in said groups (Steele & Aronson, 1995). A second condition – the “*Stereotype Threat Education*” (STE) condition – was conducted in largely the same manner as the ST condition, with the exception that students in this condition received a brief introduction into the stereotype threat effect, and how it might affect them. The goal was to compare women’s performances between the ST and STE conditions, to evaluate whether the implemented stereotype threat education had produced the proposed alleviation effect. However, for this comparison to be meaningful, they first needed to ensure that the manipulation in the ST condition had been successful in inducing the desired effect. For this purpose, a third "*Problem Solving*" (PS) condition was included, in which the exam was characterised as a neutral problem solving task, and participants did not have to make note of their sex until after they had finished working. As such, stereotype threat should not have been induced in the female participants of this control condition, and their performance should have thus been significantly better than that of the women in the ST condition. The study also included two manipulation check items for the students to answer, which allowed for an assessment of how successful the attempted manipulations by condition had been. These items comprised statements that the subjects in STE should more strongly agree with, due to the education they received on stereotype threat. Because they were also included in the present study, the full statements can be found in the "methods" section of this paper.

Using a 2x3 between-subjects factorial analysis of covariance that had test accuracy as its dependent variable and corrected for pre-existing SAT scores, Johns et al. (2005) were able to find an interaction between sex and experimental condition, $F(2, 103) = 5.35$, $p < .01$, $\eta_p^2 = .10$ – it is precisely this interaction effect that is also considered the main to-be-replicated effect of the present paper's replication. Specifically, the results showed that the men outperformed the women, but only in the ST condition, which had aimed to incite stereotype threat in the female participants. Furthermore, it was revealed that in the PS condition, women ($M = .58$) and men ($M = .53$) performed equally well on the test, $F < 1$ (no further test statistics were reported for this comparison), whereas women in the ST condition ($M = .36$) exhibited lower accuracy scores than the men of that condition ($M = .64$), $F(1, 103) = 13.21$, $p < .01$, $d = 1.35$. As such, the results lend support to the theory of stereotype threat. For the STE condition, it was found that the women who had been educated on stereotype threat prior to the exam ($M = .53$) were just as accurate in their performance as the men of said condition ($M = .56$) as well as the women of the PS condition ($M = .58$), both Fs < 1 (no further test statistics were reported for these comparisons). A direct comparison of the women in STE ($M = .53$) to those in ST ($M = .36$) showed that the former significantly outperformed the latter, $F(2, 103) = 5.57$, $p < .01$, $d = 0.82$ – the present replication will also attempt to reproduce this effect, as well as a direct comparison of the women in ST to the women of both other groups, combined. The analysis of the two manipulation check items revealed a significant main effect of condition on both items, with the women in STE more strongly agreeing with the proposed statements, $F(2, 111) = 34.45$, $p < .01$; $F(2, 108) = 6.10$, $p < .01$. As such, it can be concluded that the intended manipulations were successful. These results both support stereotype threat theory and indicate that stereotype threat education could in fact be a suitable method for mitigating the negative effects of stereotype threat, and thus bolster women's overall performance in mathematics.

With their RRR, the proposing authors wish to address three core issues they have identified in current stereotype threat literature (Stoevenbelt et al., 2021). The first pertains to publication bias, which they hope to address by achieving an unbiased effect size estimate based on the pooled results of all participating labs. As there are no incentives to produce significant results within this RRR, it should be possible to rule out any distortion of the results due to this specific bias. The second point – which is also directly addressed in the present study – regards concerns on frequently used analysis techniques, specifically, the implementation of ANCOVA models despite inherent data dependencies. This very issue is present in the original study by Johns et al. (2005) as a result of the experimental conditions being randomly assigned at group-level rather than on an individual basis, and will be resolved in the replication by employing a multi-level analysis that statistically accounts for these dependencies. As such, there should be no overestimation of the size of a potentially identified stereotype threat effect (Flore et al., 2019). Lastly, the RRR authors seek to take cultural factors into consideration, and how stereotype threat might present differently depending on the specific country it is observed in. To this purpose, they aim to establish a large, international sample with labs participating from many different countries, so that conclusions may be drawn regarding robustness and generalisability of stereotype threat. The publication of the final RRR (https://osf.io/zwnxy/) is set to occur some time between 2023 and 2024, once a sufficient number of labs have contributed to the overall sample.

For now, however, the present study intends to individually address some of the core concerns surrounding stereotype threat and the replication thereof. In replicating the original study by Johns et al. (2005), the present study seeks to determine 1. whether the found stereotype threat effect can successfully be replicated, 2. if stereotype threat education can truly help reduce the negative influence of this effect and 3. how sex composition may impact stereotype threat. The first in particular is necessary, because while the original study was able to demonstrate a stereotype threat effect for its sample, other studies and replication efforts have failed to do the same (Stoet & Geary, 2012).

Replication criteria and hypotheses

For the present study, the replication criteria of the RRR (Stoevenbelt et al., 2021) have been translated into hypotheses and listed below. Those that are marked with “*” are hypotheses that were not tested in the original study, and as such do not represent a direct replication. Hypotheses 1-4 pertain to the present study’s main analyses of interest – with 1 being considered the overall main, to-be-replicated effect – while 5 and 6 address the two manipulation check analyses.

1. There will be a significant interaction effect between sex and experimental condition for the dependent variable test accuracy.

2. There will be a significant three-way interaction effect between sex, experimental condition and sex composition for the dependent variable test accuracy.*

3. Women in the ST condition will exhibit significantly lower accuracy scores than women in the STE condition.

4. Women in the ST condition will exhibit significantly lower accuracy scores than women in the PS and STE conditions.*

5. There will be a significant main effect for experimental condition on the female participants' responses to the first manipulation check item.

6. There will be a significant main effect for experimental condition on the female participants' responses to the second manipulation check item.

Methods

Sample

The aim set by the authors of the overarching RRR was that each participating lab should produce a sample size of at least 160 test subjects, within which at least 80 needed to have been assigned to the "*Stereotype Threat*" (ST) condition, 40 to the "*Stereotype Threat Education*" (STE) condition and another 40 to the "*Problem Solving*" (PS) condition. Furthermore, data collection was to continue until a minimum of 15% male participation had been achieved. These criteria were chosen because, at the time of defining them, Stoevenbelt et al. (2021) expected at least seven labs to participate in the project. As such, they would end the experiment with a sample of at least $N = 7 * 160 = 1120$ total participants. Power analysis revealed this sample size would be sufficient for a power of 80% to find the desired interaction effect at a size of $\eta_p^2 = .02$, as well as a standardised mean difference of $d = 0.25$ – a full account of these calculations can be found at: https://osf.io/jts3c/. It should be noted that these effect sizes are noticeably smaller than the ones reported in the original study by Johns et al. (2005), so any meaningful effects should be detectable. Also, while the individual sub-samples might not be as large as the pooled one of the final RRR, they are still larger than what the original study reported on: A mere 117 students (75 female / 42 male) were included in the final analysis of Johns et al. (2005).

The initial sample of the present lab – before the application of the following exclusion criteria – consisted of 166 participants (112 female / 54 male), with a male participation quota of 32.53%. As per the guidelines determined by the RRR, all the research subjects were psychology students between the ages of 18 and 25 that had either already completed one statistics course or were currently enrolled in one. These requirements ensured that the participants had at least some affinity with mathematics, as was the case for the sample of the original study, which comprised introductory statistics students. Because of this, the original and the present samples should be as similar as possible regarding factors such as age, student status, and arithmetic ability. Psychology students in particular were chosen due to their curriculum, which comprises a statistics course, and because the University of Vienna requires them to partake in psychological experiments in exchange for course credit, making them easy to access as participants. Additionally, by limiting data collection to one type of study major, it was made more likely that the results of the pilot study (see below) would be representative for the main experiment.

Recruitment and testing took place in Vienna, Austria. Recruitment was almost entirely achieved via the internal “Laboratory Administration for Behavioural Sciences” (LABS) program of the University of Vienna’s Faculty for Psychology, which allowed for the students to participate in the experiment in exchange for university course credit. All 166 participants received this credit. The study itself was advertised as one pertaining to the cognitive aspects of human problem solving abilities, to ensure neutrality from the research subjects. The data can be viewed in its entirety at: https://osf.io/psy7j/.

Following data collection, there were several pre-defined exclusion criteria that needed to be applied to the achieved sample. On these grounds, six participants were excluded for attempting five or less of the mathematical items, and one more was excluded for skipping one of the manipulation check items. Further exclusions would have been made in the event of missing sex specifications, or if the participants had discussed the material / its purpose amongst each other during the experiment, but neither applied to the present sample. Unlike in the original study, the final analysis was not limited to Caucasian participants only. This led to a final sample of 159 participants (107 female / 52 male), with a male participation quota of 32.70%. Table 1 provides a sex-breakup according to condition.

Table 1

*Sex-breakup according to condition for final sample (*N *= 159)*

Condition	Women	Men
ST	57	26
STE	19	17
PS	31	9

Note. ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

Information on the exact ages of the participants was not collected, but they all reported to be within the desired ages of 18 and 25. Out of the 159 final participants, 28 (16 female / 12 male) stated that they belonged to an ethnic minority in Austria.

A total of 19 groups were tested. Table 2 depicts how many participants were in each group, and what the sex composition for each group was.

Table 2

Group descriptives

Group	Condition	Total tested	Female	Male
1	PS	9	6	3
2	STE	10	6	4
3	STE	5	2	3
4	PS	10	8	2
5	ST	5	1	4
6	ST	10	8	2
7	ST	9	7	2
8	ST	10	7	3
9	STE	10	7	3
10	PS	9	7	2
11	PS	10	9	1
12	ST	10	4	6
13	STE	10	4	6
14	ST	10	8	2
15	ST	10	6	4
16	ST	10	7	3
17	ST	9	9	0
18	PS	5	4	1
19	STE	5	2	3

Note. Groups listed in order of data collection; ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

Design and procedure

The design presented here is identical to that of the original study by Johns et al. (2005) – it is a 2x3 between-subjects design with sex and experimental condition serving as its factors, respectively. Randomisation occurred at cluster level, and groups were tested even if all students in one happened to be of the same sex. This was done because there is some evidence that sex composition does not moderate stereotype threat (Doyle & Voyer, 2016), and because dictating a strict sex ratio would have made data collection impractical.

Ultimately, this was only relevant to the inclusion of group 17, which had consisted of nine women but no men. In all other groups, at least one male participant had been present.

As noted above, participants were invited to the study under the false pretence that it was investigating cognitive aspects of the human problem solving ability. Testing occurred in groups between five and ten students, with the minimum being set to ensure that the individuals would feel a certain pressure of being socially evaluated by their peers. The maximum of ten on the other hand prohibited participants from feeling anonymous within a too large crowd. To further establish the competitive, high pressure nature of this setting, it was also vital that the students could see each other while working. This is why the pen-and-paper, in person format was maintained for this experiment, in spite of the at the time of this writing ongoing Covid-19 pandemic.

The interested parties first filled out an online questionnaire, denoting all possible dates that they would be free for testing, and once a timeslot where five or more students could attend had been identified, the finalised invitations were issued. All data collection for the main experiment occurred between August and November of 2020 and was held within spaces provided by the university. During testing, there was strict adherence to the hygiene protocol set by the University of Vienna at the time – masks were worn by all and an appropriate distanced maintained between participants. Additionally, any used writing utensils were disinfected and the rooms aired out between groups. Records of the names and contact information of the participants were kept, so that the other members of a group could have been immediately notified, should one of them have become infected.

Upon entering, students were guided to their designated spaces, where they found a sheet of paper with a brief explanation of the upcoming study – the true purpose of which remained concealed at this point (see Appendix A). The participants of a group were then formally greeted by the same female experimenter in every instance, after which they were given a chance to review and sign their written consent forms (see Appendix B). These included permission to save the collected data for ten years and for it to be anonymously published in the OSF, as well as a reminder that they could abort the experiment at any time without any consequences.

Following this, an audio file was played with the explicit instruction that the students please listen carefully to it. This audio instruction is one of the main priming methods of the study and varies in its content depending on the experimental condition – which, as stated, was randomly assigned to the groups before each session. A full account of these instructions translated into German (see below for details on the translation method) can be found in Appendix D-F, but will be briefly summarised here, also. All three versions began with a male confederate introducing himself to the students as a fictitious “Dr. Hamelang”, which gave the impression that they were – if indirectly – being evaluated by a male with a professional background in psychology. For the ST condition, the upcoming test was explicitly framed as one that assesses a person's inherent mathematical ability, and it was stated that the results would be used to establish performance standards for both men and women. This was intended to remind the female participants of the stereotypes surrounding women’s mathematical performance, and to highlight the pressure of one's own performance in this test potentially reflecting negatively on other women. This pressure would then, according to stereotype threat theory (Steele & Aronson, 1995), cause them to do poorly on the test. In the STE condition, the exam was also characterised to be mathematical in nature, however this was followed by a short explanation of stereotype threat as a psychological phenomenon, and how it could affect the women in the room. Female participants were instructed to keep in mind that any distress they experienced could possibly be due to stereotype threat, rather than their own lack of skill. The intent was that they would then attribute their anxiety during the test to stereotype threat, and possibly perform better because of this (Brown & Josephs, 1999). Finally, the students in the PS condition were told that they were about to undergo a simple problem solving task that did not function as a diagnostic tool for any kind of ability, merely intended to observe the underlying processes of human problem solving. Without any added pressure, students were expected to perform better in this condition, as well. All groups were told that they would receive feedback after completing the test, incentivising them to do their best.

After the audio instructions had been heard by all, the experimenter gave a brief explanation of the upcoming exam, and how the students should conduct themselves during it. They were told to avoid guessing, and that they should at least try to see every single item through to completion before moving on to the next. If there were no questions, participants in the ST and STE conditions were asked to note down their sex at the top of the exam sheet. Students in the PS condition, however, were not confronted with this information until after they had completed the exam, as part of an additional set of questionnaires. The purpose of this was to make the female students explicitly aware of their sex and the stereotypes that accompany it, so that it could influence them during the test in the ST and STE conditions. In addition to this, the title of the exam sheet in the ST and STE conditions read "*Quantitative Exam*", while that of the PS condition read "*Problem Solving Task*". While the audio instructions had already introduced them as such, putting it in writing was meant to make it yet more salient to the examinees.

The following arithmetic test comprised 30 items, for which the participants were given 20 minutes to attempt to complete them. Results were noted on a response sheet that was provided to them by the experimenter, to conserve paper waste by allowing for the test booklets to be reused. Once the time was up, the exam sheets were collected and additional questionnaires distributed. The questionnaires were identical across all conditions, with the exception that those used in the PS condition also collected information on the participants' sex, because this step had previously been skipped.

The students could complete these questionnaires in their own time, but were asked to remain seated until all had been debriefed on the true nature of the study. This debriefing occurred both orally and in form of a written statement that the participants could take home with them (see Appendix C). It included a brief explanation of stereotype threat as a psychological phenomenon, as well as some details on the overarching RRR project and its purpose. It was revealed that the GRE items and test setting were deliberately engineered to be as stressful as possible. They were told that the items were purposefully difficult, and that they were usually expected to be completed in 35, rather than the much shorter 20 minutes they were given. As such, they need not feel inadequate if they had not been able to solve many of them. Additionally, an e-mail address was provided for the occasion that any more questions should surface after their departure.

As a final step, the students were thanked for their time, and the corresponding amount of course credit was granted through the internal "LABS" system of the University of Vienna. The experiment in its entirety – from greeting to dismissal – typically took between 30 and 45 minutes.

Materials

GRE arithmetic items

To assess the mathematical performance of women under the three opposing experimental conditions, a set of arithmetic items from the American Graduate Record

Examinations (GRE) was chosen for this study. The GRE is a standardised test used primarily in the United States and Canada, which serves as a measure of general academic ability and can be an admissions requirement for graduate schools. The arithmetic portion of this exam which has been adapted for the present study consists of 30 short single choice items, the topics of which range from algebra to geometry to statistics. Each item provides the examinee with five possible answers to choose from, with only ever exactly one option being the correct response. In the actual GRE-examinations, students are given 35 minutes to complete these items. However, because the goal of this study was to create a stressful, high pressure working environment for the participants, the allocated time was reduced to a mere 20 minutes for the described experiment.

Evidence has shown that test difficulty can act as a moderator for stereotype threat (Nguyen & Ryan, 2008). The cited study found that stereotype-threatened minorities performed more poorly than non-threatened minorities on highly difficult cognitive tests, more so than they did on only moderately difficult ones. The theory behind this is that the already high mental workload of a difficult test makes it easier for a threatening stereotype to cause further interference. On the other hand, a too difficult test could cause a floor effect where all students score similarly lowly on the exam, leaving the collected data with too little variance to allow for meaningful analysis. This desired “sweet spot” for the employed test difficulty can vary greatly depending on factors such as the intended population, culture, and even the specific university or academic discipline. Taking this into account, all participating labs of the RRR project conducted their own short pilot study to determine the appropriate difficulty level for each lab.

For each pilot study, 15 students from the desired population – see above for a detailed description – undertook the arithmetic portion of the experiment, which only included the GRE items. While the main experiment contained some additional questionnaires (see below) and manipulation via audio instructions, these were not administered in the pilot, because its sole purpose was to determine whether the implemented items were too difficult for the available psychology students within the desired age range. All tests were introduced in the same way as in the PS condition of the main experiment – as a neutral problem solving task that is not indicative of any inherent mathematical ability. The pilot sessions were held on an individual basis, because this was easiest to organise and the socially competitive setting of the main experiment was not a requirement, due to no attempt being made at inducing stereotype threat in the research subjects. The students were also given 30 minutes to complete the test, rather than the 20 provided in the main experiment.

For the statistical analysis, the mean success rate of the participating students was calculated in R. The analysis script for this procedure was provided by the RRR project (https://osf.io/hq2b4/). A cut-off value of 33% was chosen – should the pilot students on average score below this number, an easier version of the GRE-items would need to be implemented for the main experiment. Similar percentages have already proven successful in other studies (20%, Davies et al., 2002; 35%, Jamieson & Harkins, 2009). Both the regular and easy items can be found in the OSF (https://osf.io/9y73x/) and were for this study translated into German by means of the forward / backward translation method by two participating labs that recruited German-speakers with a third one acting as arbiter. Both versions are largely identical, with 25 items being shared between them, and each only containing five unique items per difficulty. It should be noted that these items were not the same as the ones used in the original study by Johns et al. (2005), but rather the result of a large-scale pilot conducted by the authors of the RRR, which had aimed to establish representative GRE items that would be suitable across different countries. Details on this pilot project may also be viewed at https://osf.io/k27sp/.

The 15 psychology students for the pilot study of the present lab were all recruited directly from the immediate social circle of the experimenter, and the tests were carried out individually from July 3 to 8, 2020. This sub-sample comprised nine female and six male students. The statistical analysis revealed an item mean of $M = .36$ and an internal item consistency of $\alpha = .83$. Because this meant that the students had on average answered 36% of the items correctly, the regular version of the GRE-items was maintained for the main experiment.

Additional questionnaires

The supplementing questionnaires were all provided by the RRR project and were to be included in each individual lab assessment, so that additional research questions may later be examined with this data at a meta level, for example via sensitivity analyses. For this paper, only the sex information, answers to the manipulation check items and previous knowledge of stereotype threat were included in the statistical analysis, so the individual questionnaires will only be briefly described here. A full account of them can be found at:

https://osf.io/4x6jp/, and they were translated into German by the means mentioned above for this specific study. Slight adjustments were made during translation, mostly to account for the country in which data collection was to occur – Austria. The questionnaires will be presented here in the order that they were administered to the students. Note that unlike in the original study and the American labs of the RRR, no SAT or ACT scores could be collected from the predominantly European sample. Original materials were obtained from Johns et al. (2005) for the three manipulation recordings as well as for all the following questionnaires marked with an "*".

Sex and ethnic minority status

As noted above, only the participants in the PS condition were asked for their sex information as part of the supplemental questionnaires. The options provided were "*male*", "*female*", and "*other*". However, all students regardless of groups were asked whether they thought they belonged to an ethnic minority in Austria ("*yes / no*") – a question, which was not included in the original study, but which is relevant because students may experience stereotype threat not only due to their sex, but based on their ethnic minority status, also.

Anxiety & anxiety attribution questionnaires (Johns et al., 2005)*

Next, participants were confronted with 11 items intended to assess their perceived anxiety level during the experiment. Items included statements such as "*I felt anxious.*", to which the students could answer on a 7-point Likert scale ranging from 1 "*not at all*" to 7 "*very much*". A second questionnaire asked them how much they thought specific factors – "*time pressure*", "*sex-specific stereotypes*", and "*lack of skill or knowledge*" – contributed to their anxiety during the test. Once more they were able to respond between 1 "*not at all*" and 7 "*very much*". Anxiety was also assessed in the original study. The internal consistency of the anxiety questionnaire (referring to the first out of the two described here) was calculated to be $\alpha = .88$ in the present sample, with alpha-if-deleted ranging from $\alpha = .86$ to $\alpha = .89$.

Manipulation checks items*

Of unique importance to the overall experiment are the two manipulation check items that were included in the study. These were imbedded in two smaller sets of items and are necessary in order to determine whether the manipulations via the differing audio instructions per condition had indeed been successful. Both items were also included in the original study.

The first set of items asked the participants what factors they thought the researcher believed would cause people to do poorly on the administered test. The three factors discussed here were "*time pressure*", "*sex-specific stereotypes*", and "*anxiety*". The students were able to rate each of these on a 7-point Likert scale, ranging from 1 "*do not agree*" to 7 "*agree*". The item of interest here is the one pertaining to sex-specific stereotypes, because one would expect the participants in the STE condition to more strongly agree with this item, based on the manipulation they received in form of the audio instruction.

The second set of items asked the participants 1. how they expected men and women to do relative to each other and 2. how they thought the researcher expected men and women to do relative to each other on the test. The students could answer on a 7-point Likert scale, ranging from 1 "*men will do better than women*" to 7 "*women will do better than men*", with 4 being a neutral "*men and women will do the same*". The critical item is the one regarding the beliefs of the researcher, because students in the STE condition – had they indeed listened to the audio instruction explaining the stereotype threat effect – should have likely thought that the researcher would expect the men to outperform the women.

<u>Perceived test difficulty*</u>

Here, students were asked to simply rate how difficult they found the arithmetic items on a 7-point Likert scale, ranging from 1 "*extremely easy*" to 7 "*extremely difficult*". This item was also included in the original study.

<u>Domain Identification Measure (Smith & White, 2001)</u>

In this questionnaire, students were asked to rate items based on how strongly they identified with the mathematical domain. Nine items were presented, comprising statements such as "*Mathematics is one of my best subjects.*". The participants were able to respond on a 5-point Likert scale, ranging from 1 "*strongly disagree*" to 5 "*strongly agree*". This scale was not included in the original study. Cronbach's alpha for this questionnaire was $\alpha = .91$, with alpha-if-deleted ranging from $\alpha = .88$ to $\alpha = .91$.

Stereotype awareness 1*

Here, the students were presented with two items intended to assess their awareness level for societal stereotypes. The first item asked them whether they – disregarding their personal beliefs – thought that people held stereotypes that men and women differ in mathematical ability. Students were able to answer on a 7-point Likert scale, ranging from 1 ”*yes, men are stereotyped to be better than women*” to 7 “*yes, women are stereotyped to be better than men*”, with 4 serving as a neutral “*no, there are no stereotypes*”. The second item repeated this question, however instead of referencing “people” in general, it specifically inquired about people in Austria. Only the first question was included in the original study.

Gender Identification Scale (Luhtanen & Crocker, 1992; Schmader, 2002)

This questionnaire was used to assess how strongly participants identified with their respective sex. The four items consisted of statements such as “*Being a woman / man is an important part of my self-image.*” and could be answered on a 5-point Likert scale ranging from 1 “*strongly disagree*” to 5 “*strongly agree*”. This scale was not included in the original study. Cronbach’s alpha for this questionnaire was α = .80, with alpha-if-deleted ranging from α = .74 to α = .76.

Ethnic minority stereotypes

Participants that had previously stated they belonged to a minority group in Austria, were now asked to rate on a 7-point Likert scale whether they thought academic stereotypes existed for said group. Possible answers ranged from 1 “*Yes, my group is stereotyped to have lower performance*” to 7 “*Yes, my group is stereotyped to have higher performance*”, with 4 being a neutral “*There are no academic stereotypes for my group*”.

Stereotype awareness 2 (Inzlicht & Kang, 2010)

While general awareness for societal, sex-specific stereotypes was already addressed in a prior item, these items asked participants for what domains specifically they thought such stereotypes exist. The three items included statements such as “*Cultural stereotypes indicate that men are better than women in* ____”, to which the students could fill out the blank space either with 1 “*verbal skills*” or 2 “*mathematical skills*”. These items were not included in the original study.

Perceived purpose of the study and stereotype threat awareness*

This last section asked the participants in an open response format for their own theories on what the purpose of the experiment could have been, as had also been done in the original study. For the finalised data sheet, which may be viewed at: https://osf.io/psy7j/, these responses were translated into English by the experimenter. Additionally, using a set of rules provided by the RRR project (https://osf.io/efa5q/), the responses were coded as either 0 "*Participants did not write down the true nature of the study*" or 1 "*Participants did write down the true nature of the study*".

Finally, the students were asked whether they had ever been or currently were enrolled in a course on social psychology ("*yes / no*"), and if they had already been familiar with the concept of stereotype threat ("*yes / no / don't know*"). The last two items are new to the replication.

Ethical concerns

Most of the ethical concerns going into this study were due to the high pressure, socially competitive work environment that needed to be established for the experiment to function as intended, as well as the fact that recruitment occurred under false pretences. This could have potentially left many of the students – especially the female ones in the ST condition – feeling rather inadequate regarding their abilities during and after the test.

To address these negative effects, an extensive debriefing was implemented, both orally and on a separate sheet the participants were able to take home with them (see the "design and procedure" segment of this paper for more details on its content). Students were also made aware in their consent forms that they could end the experiment at any time, free of consequence, and it was decided to not offer any incentives other than course credit – such as actual monetary compensation – so the students would not feel coerced into participating.

Other health concerns stemmed from the at the time of this writing ongoing Covid-19 pandemic, but as was already outlined above, all necessary steps to protect the participating students were taken here.

The experiment, as it was performed in this specific lab, was approved by the ethics committee of the University of Vienna (reference number 00501), as well as the overall RRR by Tilburg University's Ethics Review Board (reference number EC-2017.104a).

Statistical analysis

All R scripts used in the presented data analysis were provided by Stoevenbelt et al. (2021) and can be found at: https://osf.io/c9pm5/. These scripts remained largely unchanged, though some alterations needed to be made in order for the code to function as intended – such as adapting variable names to the current data file (https://osf.io/psy7j/), or making adjustments for the fact that only a single data set is being analysed in the present study, rather than the pooled data of many labs that the scripts were originally written for.

The scripts included a set of pre-defined exclusion criteria that were preliminarily applied to the data – for a full account of these criteria and how many participants were excluded on their basis, see the above sample description.

Internal consistency / reliability (Cronbach's alpha) was calculated for the GRE items, and classical test theory analysis performed to assess item difficulties and item-rest correlations. All results here pertain only to the regular version of the test, as no participants received the easy version in this sample. Cronbach's alpha was also calculated for some of the additional questionnaires (see above).

For the score calculation of the GRE items, correct answers were coded as 1 and incorrect / missing / multiple answers were coded as 0. Because the interaction of sex and experimental condition on the dependent variable test accuracy is considered the primary to-be-replicated effect, accuracy scores were calculated for all participants. This was achieved by dividing a participant's number of correctly answered items by their number of attempted items – with attempted items being any items where the student had marked at least one answer as correct. For the RRR project, all raw total scores will need to be equated with the chained linear equating method (Kolen & Brennan, 2014) to account for the fact that some labs will have used the regular and others the easy version of the test. Because all participants in this sample received the regular version, this step was skipped, and all scores are reported as such.

In the original study by Johns et al. (2005), an ANCOVA model was implemented despite the data's inherent cluster dependencies. This is a direct violation of the independence assumption, which has been shown to lead to overestimations of both size and robustness of stereotype threat effects (Flore et al., 2019). To avoid these complications, a two-level random-intercept multilevel analysis was chosen instead for both the RRR as well as the present study, with the experimental groups at level 2 and the individual students at level 1.

The present paper addresses the influence of sex and experimental condition, as well as their interaction, on test accuracy. As such, the independent variables were sex (between subjects; at level 1) and condition (between subjects; at level 2), with accuracy – see above for the calculation method – serving as the dependent variable and the testing groups as a level-two clustering variable. Sex was coded as either 1 for "*female*" or 0 for "*male*". For the three differing experimental conditions, two orthogonal Helmert contrasts (Cohen et al., 2003) were created: The first compares women in the ST condition to women in both other conditions, the second compares women in STE specifically to women in PS. To achieve balance for the design of the first contrast, contrast coefficients were calculated relative to sample size. Table 3 depicts how each experimental condition is coded for each contrast, respectively.

Table 3

Contrast coding for statistical analysis

Condition	Contrast 1	Contrast 2
ST	-1	0
STE	$N_{STE} / (N_{STE} + N_{PS})$	-1/2
PS	$N_{PS} / (N_{STE} + N_{PS})$	1/2

Note. Contrast 1 = ST vs. STE and PS; Contrast 2 = STE vs. PS; ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

This analysis began with an "empty" model without predictors that had random intercepts for experimental clusters. Next, sex was included as an independent variable, then the two condition contrasts, and finally the interactions between sex and each contrast (*sex * condition contrast 1, sex * condition contrast 2*).

Chi-square tests were incrementally conducted to assess whether adding further predictors led to significantly better model fit, and the Bayesian Information Criterion (*BIC*) of each model was also calculated to this purpose. For the replication to be successful, the full model would have to exhibit the lowest *BIC* – thus indicating best model fit – and there would need to be a significant ($p < .05$) interaction effect between sex and the first contrast, which compares women in ST to the women of both other conditions.

While evidence for the moderating effect of sex composition on stereotype threat has so far been mixed (Doyle & Voyer, 2016; Inzlicht & Ben-Zeev, 2000), further investigation into the matter was pursued here by including sex composition per experimental cluster in a secondary confirmatory analysis. First, a sex composition variable was created for each group by dividing the number of female participants by the number of total participants in that group. Next, a dummy variable was created, where 0 was coded as "*more women than men / as many women as men*" and 1 was coded as "*more men than women*". This dummy was intended to function as the predictor variable for sex composition in the final analysis.

However, the variance of the dummy within the given sample was ultimately calculated to be rather low ($\sigma^2 = 0.17$) – a reflection of the fact that most groups had more women than men partaking in the experiment. Because there might otherwise not be enough variance in the predictor variable for the analysis to be meaningful, it was decided that the raw sex mean scores per group (number of female participants in group / number of total participants in group) would instead be used as the variable for sex composition.

The secondary analysis started out with a model only consisting of the simple effects of sex, experimental condition, and the new variable for sex composition. In a next step, the two-way interactions of these variables (*sex * condition contrast 1, sex * condition contrast 2, sex * sex composition, sex composition * condition contrast 1, sex composition * condition contrast 2*) were included, and finally, their three-way interactions (*sex * condition contrast 1 * sex composition, sex * condition contrast 2 * sex composition*). As with the first main analysis, chi-square tests were incrementally calculated and the *BIC* of all three models compared to assess best model fit. The original study did not account for sex composition, and as such, there are no specific effects to be replicated here. As a final step, the *BIC* of the best fitting models of both main analyses were compared to one another to determine which model overall best explained the data.

Finally – although this could by design not be included in the present study – the authors of the RRR will add the conducting lab as a third-level grouping variable and perform a third multilevel analysis with it, to evaluate the possible influence of cultural differences across the varying countries the individual replications took place in.

Aside from the main analyses, which focused on the interaction effect of sex and condition, two standardised mean differences (Cohen's *d*) and their 95% confidence intervals were also calculated to determine whether there was a notable difference in accuracy scores for women depending on their experimental condition. The first effect is a direct replication of the original study by Johns et al. (2005) and compares women's accuracy in the ST condition to that of women in the STE condition. The second compares women in the ST condition to the combined performance of women in the other two conditions. Stoevenbelt et al. (2021) deemed it suitable to combine the STE and PS conditions, because previous studies had not found any notable contrast of test performance across differing low-threat conditions such as these (Doyle & Voyer, 2016). While this measure was not included in the original study, the authors of the RRR intend to include these effect sizes in a random-effects meta-analysis across all participating labs.

For the manipulation check items, two additional multilevel analyses were conducted with the responses to said items serving as their dependent variables, respectively. The goal was to confirm whether the intended manipulation by condition had indeed been successful – after all, participants in the STE condition should already be familiar with the importance of stereotype threat to the experiment and answer accordingly. The first item, which asked the students to what extent they thought the researcher expected sex-specific stereotypes to cause poor performance during the test, was recoded for this purpose: It now ranges from 1 "*strongly agree*" to 7 "*strongly disagree*". This was done so that the direction remained consistent between the two analyses, because a lower score would now be the expected outcome across both items for students in the STE condition. For the conditions, two dummies were created – the coding for these, however, notably differs from the contrast coding of the main analyses, as demonstrated in Table 4. Both multilevel analyses included both dummies as their independent variables. Unlike in the original study, men were excluded from these analyses because the theory suggests the manipulation should not have affected them (Steele, 1997). Random intercepts were included for each experimental cluster.

Table 4

Dummy coding for manipulation check analyses

Condition	Dummy 1	Dummy 2
ST	1	0
STE	0	0
PS	0	1

Note. STE always coded as 0 because it is the relevant reference group; ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

Several possible sensitivity analyses were proposed by Stoevenbelt et al. (2021), such as excluding the non-Caucasian students for the American samples – as was the case in the original study by Johns et al. (2005) – or by including the SAT, sex- or domain-identification scores as a possible covariate. For the present study, participants that claimed to have prior knowledge of stereotype threat were, exploratively, excluded to assess whether the critical interaction effect could possibly be more pronounced in less biased groups. It was considered to also exclude all students that had already undergone a course on social psychology, however the remaining sample size would have been reduced to a mere N = 16.

Results

Descriptives

Because accuracy on the GRE items – calculated as the number of correct items over the number of attempted items – is the main dependent variable of the experiment, Table 5 serves to illustrate the mean accuracy scores of the sample per condition. Mean scores are reported both for the male and female participants of each condition separately.

Table 5

*Means and standard deviations of accuracy scores per condition and sex (*N *= 159)*

Condition	Women	Men
ST	.63 (.22)	.70 (.18)
STE	.70 (.17)	.66 (.21)
PS	.64 (.22)	.64 (.24)

Note. Accuracy scores = number of correct items / number of attempted items; ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

The results show that women in the STE condition on average have slightly higher accuracy scores than the women in the other two conditions, with accuracy in the ST and PS conditions being almost identical. The following multilevel analyses and calculated standard deviations will help determine the significance of these mean differences.

While not included in the main analyses, means and standard deviations were also calculated for some of the additional items / questionnaires and are presented in Table 6.

Table 6

*Means and standard deviations of additional items / questionnaires (*N *= 159)*

Item / Questionnaire	*M (SD)*
Perceived test difficulty	4.18 (1.06)
Anxiety questionnaire	3.26 (1.09)
Gender Identification Scale	3.26 (0.93)
Domain Identification Measure	3.19 (0.89)
Stereotype awareness 1	1.85 (0.79)
Stereotype awareness 1 (Austria)	1.97 (0.84)

Note. See the "materials" segment for a description of the response ranges for each item / questionnaire.

Reliability and classical test theory analysis

In terms of reliability, the arithmetic items exhibited an internal consistency of $\alpha = .72$, which is generally considered to be within an acceptable range (Blanz, 2015). Even when calculating Cronbach's alpha in the absence of each individual item, no one item seemed to have significantly impaired the overall consistency of the test – the highest alpha-if-deleted was $\alpha = .74$ in the case of the removal of item 5. A full list of these "what-if" reliabilities may be viewed in Table 15 of Appendix G. The (biserial) item-rest-correlations ranged from -.01 to .88.

Item difficulties ranged from .00 to .80. When reviewing the full list, which can also be found in Table 15 of Appendix G, it becomes clear that the low solution frequencies of the items presented later in the test are caused by a speed-power problem: Specifically, while the items are all designed to be equally difficult, the administered exam focused on the speed with which participants could solve them, making it unlikely that all examinees would be able to attempt all questions – in particular those towards the end of the test. As such, it is the very first item that exhibits the lowest difficulty / highest solution frequency (.80), and the last five items all show an overall mean score of .00, as barely anyone even got that far in the test. For the most part, the item difficulties follow this pattern of decline, with the exception of some items that the participants on average seemed to just inherently find easier / more difficult.

Manipulation check analyses

As mentioned above, this analysis was applied to a sub-set of the collected sample, which only included the female participants (*N* = 107).

Table 7

*Means and standard deviations of women's response to both manipulation check items by condition (*N *= 107)*

Condition	Item 1 (recoded)	Item 2
ST	3.47 (1.48)	2.91 (0.91)
STE	1.90 (1.41)	2.74 (0.81)
PS	4.07 (1.81)	3.03 (1.08)

Note. ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

Table 8

*Fixed regression coefficients and standard deviations for the multilevel analyses of both manipulation check items (*N *= 107)*

Parameters	Item 1 (recoded)	Item 2
Dummy 1	1.58 (0.41)***	0.18 (0.29)
Dummy 2	2.17 (0.45)***	0.32 (0.32)

Note. * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; dependent variable = response to relevant item; Stereotype Threat Education Condition coded as 0 in both dummies (reference group).

The first manipulation check item asked participants on a 7-point Likert scale to what degree they thought the researcher expected sex-specific stereotypes would lead to poor performance on the test. Table 7 shows that women in the STE condition did in fact score lower on this first item than did the women of the other two conditions – meaning they more strongly agreed with the proposed statement. A multilevel analysis was conducted to determine the significance of this difference, with Table 8 demonstrating a significant main effect of both dummy variables for experimental condition on item response. As such, it can be concluded that women in STE agreed to the proposed statement at a significantly higher rate than the women in ST and PS.

The second manipulation check item asked participants to rate on a 7-point Likert scale how they thought the researcher expected men and women to do relative to each other on the test, with a lower score indicating a preference towards the male students. While the women in the STE condition exhibited the lowest averaged score for this item according to Table 7, the observed difference across all conditions is rather minimal. Once more, a multilevel analysis was conducted to determine the significance of these minor differences. Table 8 demonstrates that there was no significant main effect of either condition dummy on item response. The conclusion follows that the experimental condition had no significant influence on how strongly women agreed to this second statement.

Confirmatory analyses

Sex-condition interaction

The first and main multilevel analysis started out with an "empty" random-intercept model, then included sex as an independent variable, after that the two contrasts for condition – one comparing ST to both other conditions, and one comparing STE to PS –, and lastly, the interactions of sex and condition. Table 9 depicts model comparison tests for the four proposed models, as well as the *BIC* of each model.

Table 9

*Model comparisons for multilevel analysis 1 (*N *= 159)*

Model	Log-likelihood	*BIC*	χ^2	*df*	*p*
empty	27.00	-38.79			
+ sex	27.40	-34.52	0.798	1	.372
+ conditions	27.73	-25.05	0.668	2	.716
+ interactions	28.62	-16.69	1.784	2	.410

Note. Dependent variable = accuracy; sex coded as 1 for "*female*", 0 for "*male*".

Based on these results, it can be deduced that it is the "empty" model with random intercepts that best explains the data. This is evidenced by the "empty" model having the lowest *BIC* out of all four models, as well as by the fact that none of the incremental chi-square tests were significant. Adding sex, condition, or their interaction effects as predictors did not better explain participants' accuracy scores. To further illustrate this, the results of the "full" model – including sex, condition, and their interaction – are more closely examined in Table 10. Standardised regression coefficients (Betas) were calculated separately for each predictor and included in the table.

Table 10

*Results for the "full" model (*N *= 159)*

Parameters	Unstandardised coefficients	Standardised coefficients (β)
Sex	-0.01 (0.04)	-0.02
C1: ST vs STE and PS	-0.03 (0.04)	-0.12
C2: STE vs PS	-0.03 (0.08)	-0.04
Sex * C1	0.06 (0.05)	0.18
Sex * C2	-0.04 (0.10)	-0.05

Note. * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; dependent variable = accuracy; ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

Results show that none of the predictor variables were significant in this model, meaning no significant main effect of sex, either condition contrast, or their interactions was discovered for the dependent variable accuracy. Both the fact that none of the proposed independent variables were significant, as well as the small Beta-coefficients of those predictors (Acock, 2014), are further evidence that an "empty" model without predictors best explains the data.

As an exploratory analysis, the just described model comparisons were repeated under the exclusion of any participants that claimed to have already been aware of the stereotype threat effect. For this purpose, a total of 122 students were excluded, leaving this sub-sample with 37 participants (24 female / 13 male).

Table 11

*Model comparisons for multilevel analysis 1 under added exclusion criteria (*N *= 37)*

Model	Log-likelihood	*BIC*	χ^2	*df*	*p*
empty	4.31	2.22			
+ sex	4.31	5.83	0.003	1	.955
+ conditions	5.12	11.42	1.632	2	.442
+ interactions	9.52	9.84	8.797	2	.012*

Note. * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; dependent variable = accuracy.

As shown in Table 11, adding the interactions of sex and condition as predictor variables lead to a significant improvement in fit when compared to the model only including the main effects of sex and condition. As was the case for the full sample analysis, however, it is the “empty” model that overall exhibited the lowest *BIC* and thus the best fit out of all four models. The conclusion follows that, even in this sub-sample, neither the proposed interaction effect nor the main effects of sex or condition explain the data better than a model with no predictors.

Sex composition

The secondary multilevel analysis began with a random-intercept model that only contained the main effects of sex, condition, and a variable for sex composition by group. A next step added two-way interactions between these predictors and, finally, three-way interactions were included.

Table 12

*Model comparisons for multilevel analysis 2 (*N *= 159)*

Model	Log-likelihood	*BIC*	χ^2	*df*	*p*
Main effects with sex composition	28.00	-20.51			
+ two-way interactions	29.61	1.60	3.234	5	.664
+ three-way interactions	29.82	11.33	0.403	2	.818

Note. Dependent variable = accuracy.

The model comparison tests presented in Table 12 demonstrate that, out of the three, the model with the main effects only exhibited the best fit. This is evidenced by it having the lowest *BIC*, as well as none of the chi-square tests being significant. As such, there did not appear to be any significant two- or three-way interaction effects.

Table 13

*Results for the "main effects + sex composition" model (*N *= 159)*

Parameters	Unstandardised coefficients	Standardised coefficients (β)
Sex	-0.02 (0.04)	-0.04
C1: ST vs STE and PS	0.01 (0.02)	0.02
C2: STE vs PS	-0.02 (0.05)	-0.03
Sex composition	-0.00 (0.01)	-0.07

Note. * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; dependent variable = accuracy; ST = Stereotype Threat; STE = Stereotype Threat Education; PS = Problem Solving.

Table 13 presents the results for this "main effects + sex composition" model, once again also including its standardised regression coefficients (Betas). While this was the best fitting model out of the three, it can be seen here that none of the included predictors were significant, and the Beta coefficients are all rather small in size (Acock, 2014). As such, sex composition did not seem to have a significant influence on the achieved accuracy scores, and there were also no significant two- or three-way interactions.

Table 14

*Comparison of "main effects" and "main effects + sex composition" models (*N *= 159)*

Model	Log-likelihood	*BIC*	χ^2	*df*	*p*
Main effects	27.73	-25.05			
+ sex composition	28.00	-20.51	0.534	1	.465

Note. Dependent variable = accuracy.

The direct comparison shown in Table 14 of the "main effects (for sex and condition)" model to the one including sex composition shows that the basic "main effects" model has the better fit out of the two – further illustrating that the data holds no evidence of sex composition having a significant influence on test accuracy. Furthermore, across both confirmatory analyses, the "empty" model with no predictors whatsoever remains the one with the lowest *BIC* of -38.79 and thus the highest overall fit for the presented sample.

Standardised mean differences

The standardised mean difference for women in the ST compared to women in the STE condition was d = -0.35 [-0.88; 0.19]. This indicates a small effect (Cohen, 1988) for women in STE outperforming the women in ST, though it appears to be a non-significant effect, due to the value of "0" falling within its 95% confidence interval.

For the comparison between women in ST to the women of both other conditions combined, the standardised mean difference was estimated to be d = -0.16 [-0.54; 0.23]. The trivial effect size (Cohen, 1988), along with the 95% confidence interval once more containing the value "0", indicates that there was no significant – or otherwise meaningful – difference in accuracy scores between the women in ST and those of the two control conditions.

Discussion

To address the ambiguity currently surrounding the topic of stereotype threat, the present study's main goal was to evaluate whether a stereotype threat effect could successfully be replicated when adhering to the same design and protocol as the original study by Johns et al. (2005). The overall results of the conducted experiment, however, provide no evidence of such an effect existing within the analysed sample.

According to stereotype threat theory, the ST condition should have induced stereotype threat in the participating women by reminding them of the stereotypes surrounding their sex and the domain of mathematics, thus causing them to experience high performance pressure and, as a result, do poorly on the test (Steele & Aronson, 1995). In contrast, the women in the other two neutral conditions (STE and PS), as well as the men across all conditions, should not have been hindered in their performance.

Because only one specific sex in one specific condition should have been negatively impacted by stereotype threat, there would have needed to be a significant interaction effect between sex and experimental condition for the results to support stereotype threat theory. This was not the case, however, because out of several model comparisons, a model with no predictors whatsoever best explained the observed differences in accuracy scores. In other words: Including the interaction between sex and condition as a predictor did not help to better account for differences in the students' performance, ruling out a stereotype threat effect for the current sample. This conclusion is also supported by a closer look at the model which included all relevant predictors – sex, condition, and the interaction of the two: None of the predictors were significant in this model, and their effect sizes ranged from $\beta = -0.12$ to 0.18. While there are no unanimous benchmarks within the scientific community for how to interpret standardised regression coefficients, none of these predictors appear to be impactful enough in size to be meaningfully considered in spite of their lacking significance (Acock, 2014). Adding the simple effects of sex and condition as predictors also failed to explain the data better than a model with no predictors. This demonstrates that – in addition to there being no observable stereotype threat effect – the differing experimental conditions had no notable influence on participants' performance, and furthermore, that there were no sex differences across all conditions.

Because the achieved results do not support stereotype threat theory, one must conclude that the H0 of the present study's first hypothesis, which states that there is no interaction effect for sex and experimental condition, needs to be maintained. Because this interaction effect is considered the main to-be-replicated effect of the RRR, it follows that the present study was not able to successfully replicate the results found by Johns et al. (2005). This is consistent with the meta-analysis by Stoet and Geary (2012), where only 55% of 19 stereotype threat studies were able to reproduce the relevant effect. These results should also be considered in the context of the replication crisis, with stereotype threat being one of many psychological effects that recent replication efforts – such as the RPP (Open Science Collaboration, 2015) – have struggled to successfully replicate.

When directly comparing the accuracy scores of women in the ST to women in the STE condition, a standardised mean difference of d = 0.35 was determined, with the women in STE outperforming those in ST. This represents a small effect size (Cohen, 1988), which could indicate that 1. the women in the ST condition experienced stereotype threat, thus negatively impacting their accuracy and 2. the women in the STE condition may have benefited from the education on stereotype threat as a psychological concept and performed better for this reason. This small effect, however, is lacking in statistical significance, and the effect found by Johns et al. (2005) for this same comparison was far larger (d = 0.82). Such a decline in effect size is consistent with other failed replication efforts (Open Science Collaboration, 2015). As such, it can be concluded that the present study failed to replicate the effect of the original study, and that the H0 for the third hypothesis – that the women of these two conditions performed equally – must be maintained. The results neither support stereotype threat theory, nor an alleviating effect of stereotype threat education on women's mathematical performance.

The second standardised mean difference was assessed for women in ST vs. the women of both other conditions, combined, which resulted in d = 0.16 for the women in STE and PS outperforming those in ST. This difference also failed to be statistically significant, and its effect size is furthermore considered negligible (Cohen, 1988). These results indicate that women performed equally across all conditions, and as such, that the H0 of the fourth hypothesis is to be maintained. This is in line with the results of the first confirmatory analysis, which also failed to demonstrate a main effect of experimental condition for the dependent variable accuracy, and overall supports an absence of stereotype threat in the observed sample. This second comparison was not included in the original study, thus not representing a direct replication. The authors of the RRR will later use the effect sizes of all contributing labs to calculate a random-effects meta-analysis for the women in ST vs. those in the other two conditions.

To better understand the apparent absence of stereotype threat in the present sample, one must consider any factors which may help account for why the results differ so drastically from those reported in the original paper by Johns et al. (2005).

This includes potential limitations of the present study that could have impeded its ability to successfully detect an existing stereotype threat effect, such as the effectiveness of the chosen manipulation method. To control for this, two manipulation check items were added to determine whether the intended manipulation by experimental condition had in fact been successful. For the first item, where participants were asked to what degree they thought the researcher expected sex-specific stereotypes to cause students to do poorly on the exam, the multilevel analysis revealed a significant main effect of experimental condition on the women's response to this item. The mean responses showed that women in the STE condition more strongly agreed with the above statement. This is in line with expectations for a successful manipulation, because the women in STE had actively been made aware of stereotype threat as a psychological concept and would thus likely expect sex-specific stereotypes to at least partially be the focus of the experiment. The analysis of the second manipulation check item, however, where participants were asked how they thought the researcher expected men and women to do on the exam relative to each other, demonstrated no significant main effect of experimental condition on the relevant item response. The mean responses also follow this pattern, with women in STE, ST, and PS all exhibiting similar mean scores. As such, this second analysis would suggest that the manipulation had perhaps not been so successful after all, because the women in STE should have believed the researcher to expect the men to outperform the women due to the negative influence of stereotype threat.

These results indicate that the H0 of the present study's fifth hypothesis, which states that there is no main effect of experimental condition on the response to the first manipulation check item, can be discarded in favour of the H1. For the sixth hypothesis, however, the H0 must be maintained, because no main effect of condition on response was found for the second manipulation check item. As such, the manipulation check results of Johns et al. (2005) can only be considered partially replicated – they had found a significant main effect of condition on both manipulation check items.

The fact that the intended manipulation appears to have only partially been successful could mean that all results of the present study need to be interpreted with due caution. However, it could also speak to a fundamental flaw in the design of the manipulation check items: Having been clued into the fact that the experiment pertains to stereotypes and their threatening influence on women, it is easy for the women in STE to deduce that the researcher expects sex-specific stereotypes to be of great importance to the study and answer the first item accordingly. The second item, however, does not relate quite so directly to stereotype threat – it does not mention sex-specific stereotypes in any way, only, how the researcher expects men and women to perform relative to each other on the test. As such, the women in STE might not have been actively considering sex-specific stereotypes when reading this second question, and thus simply have defaulted to men and women performing equally in the researcher's eyes. Additionally, if they were considering stereotype threat, they may have caught on to the fact that the initial audio instruction was meant to mitigate its negative effect on women and expect men and women to perform equally for this reason. In light of these factors, as well as the fact that the first manipulation check analysis did produce a significant main effect of experiential condition on item response, one likely need not entirely discount the success of the attempted manipulations. However, the possibility that they were not as effective as intended should be kept in mind for all further discussion.

While the manipulation check analyses – at least the first one – confirmed that the participants had listened to the audio instructions and were thus able to deduce the researcher's true intent in the STE condition, they do not control for the possibility that stereotype threat could have existed in all conditions, independent of any active manipulation. Describing what is clearly a set of mathematical items as a neutral problem solving task in the PS condition, and asking the students to write down their sex only after they are done working, might not have been enough to make the situation less threatening for the women. As such, there would be no main effect of condition for the dependent variable accuracy – with women's accuracy being consistent across all conditions –, but stereotype threat would still exist within the sample and in each of the conditions. If this were the case, however, there should have at least been a main effect of sex, with the men surely outperforming the women in the event of the latter being negatively influenced by stereotype threat. Because it was the “empty” model that best fit the data, this explanation seems rather unlikely.

Another consideration is that – with all participants of the present sample being psychology students – the women may have already been made aware of stereotype threat and thus, through the benefits of stereotype threat education, were no longer negatively impacted by it. To account for this possibility, an exploratory sensitivity analysis was conducted where all participants from the sample that stated to have already had knowledge of stereotype threat were excluded. With this sub-sample (N = 37), the first confirmatory analysis and its model comparisons were repeated, to assess whether the absence of stereotype threat remained robust under these conditions. Results show that even in this subset of participants, the "empty" model best explained the data, further confirming the absence of a stereotype threat effect in the present study.

Of course, possible limitations of the present study are not the only factors that may have led to this outcome. One theory is that there has simply been a genuine decline in stereotype threat effect over time. The original study was published over a decade ago at the time of this writing, and many of the relatively young participants of the present study would have grown up in an educational environment where sex-specific stereotypes in mathematics were likely not as prevalent as they had once used to be (Eagly et al., 2019). It has been suggested that the women of this generation might not fear to confirm a negative stereotype for their sex with their own poor performance, simply because they feel that such a stereotype no longer exists. On the other hand, the recent shift in how women are perceived in modern academia could have just as easily led to an increase in stereotype threat: After all, the stakes have never been so high to consolidate the fact that women too can excel in the sciences, now that they are – if slowly, and with some trade-offs – finally beginning to receive recognition for it (Huang et al., 2020). The modern woman could conceivably fear taking away from all the progress that has already been made by other, academically more successful women. As such, it cannot – without further dedicated research – be definitively concluded that societal changes in stereotype perception have led to a decline in stereotype threat.

Another possible explanation is that stereotype threat manifests differently in different countries, and might simply not be as prevalent in Austria as it was in the American sample of the original study. Furthermore, Johns et al. (2005) only included Caucasian participants in their final analysis, because some prior studies had only been able to find sex differences for this group (Hyde et al., 1990), while no such criterion was implemented for the present study. As such, future research should aim to investigate these cultural aspects of stereotype threat.

The authors of the RRR project, for example, will perform sensitivity analyses to compare Caucasian-only-samples to the more heterogeneous pooled sample, and to generally assess how the results may differ depending on the country of the participating lab.

Assuming that the effect – to whatever extent it may exist – has remained consistent over the years and perhaps even cultures, further reasons for the failed replication could be found in the methodological differences between the present study and that of Johns et al. (2005). For one, an ANCOVA model was employed for the original analysis, despite the inherent cluster dependencies of the data that resulted from participants being randomly assigned to experimental condition at group-level. This violation of the independence assumption could have led to an overestimation of both size and robustness of the discovered stereotype threat effect (Flore et al., 2019). If the effect was never as impactful as assumed, it follows that the present study – which made use of a multilevel model to account for data dependencies – would not be able to provide similar evidence for it. Additionally, the statistical analysis of the original study corrected for pre-existing SAT scores by including them as a covariate. This may have confounded their ANCOVA, considering it is vital for such a covariate analysis that the covariate does not systematically differ between the relevant groups (Miller & Chapman, 2001) – here: men and women. This assumption, however, clashes with stereotype threat theory, because said theory specifically predicts women to underperform in mathematical assessments such as the SATs. The consequence of this was demonstrated by the above referenced meta-analysis (Stoet & Geary, 2012): When only taking the non-adjusted studies into consideration, the overall success rate of the replications dropped from 55% to 30%. As such, performing these statistical adjustments may have led to an inflated discovery of stereotype threat effect in the original study, making it reasonable that the current study – which did not include SAT scores in its analysis – would not show the same effect. The RRR's analysis will, for American sub-samples with SAT scores, be performed both with and without SAT scores as a covariate, to further investigate the confounding effect of such an adjustment.

As a more general concern, this and other failed replications of stereotype threat are likely at least in part brought on by publication bias (Ferguson & Brannick, 2012). Johns et al. (2005) and other studies, which were able to demonstrate a significant stereotype threat effect (e.g., Spencer et al., 1999), may have been mere outliers in their field: It is very possible that many other studies were conducted without any evidence of stereotype threat being found, but that they were simply never published because the results were not deemed interesting enough. The consequence of this would have been an overestimation of the stereotype threat effect in the scientific community. As such, stereotype threat may have simply never been as prevalent or impactful as assumed, and it would come as no surprise that many, if not most, modern studies fail to replicate the effect.

Beyond the attempted replication of a stereotype threat effect, the present study also addressed the possibility of the individual testing groups' sex composition having an influence on the manifestation of stereotype threat. For the results to support such an influence, the model accounting for a three-way interaction between sex, condition, and sex composition would have needed to best fit the data. This would have not only meant that there were sex differences in accuracy depending on experimental condition – a conclusion which has already been ruled out by the results of the first confirmatory analysis –, but that these differences also varied depending on the sex composition of a participant's group. This was not the case, however: Out of the three models compared in this analysis, it was the one which only included the simple main effects of sex, condition, and sex composition that best explained the observed differences in accuracy scores. When more closely evaluating this model, however, it shows that none of its predictors – including the one for sex composition – were significant, and that all Beta coefficients were very weak in size (β = -0.07 – 0.02) (Acock, 2014). Furthermore, a direct comparison between this model and one only accounting for the main effects of sex and condition illustrates that the addition of sex composition as a predictor did not lead to a significantly better model fit. Out of all models across both confirmatory analyses, it is still the model with no predictors whatsoever which overall best explains the collected data. It can be concluded that a group's sex composition did not have any kind of notable influence on participants' performance.

Following these results, the H0 of the second hypothesis – that there is no significant three-way interaction between sex, condition, and sex composition – must be maintained. This is in line with research that has not been able to find any evidence of sex composition moderating stereotype threat (Doyle & Voyer, 2016). Other studies, however, found that the performance of women suffered proportionally to how many males were present, with more males leading to poorer performance (Inzlicht & Ben-Zeev, 2000). While this could be a stand-alone effect, independent of stereotype threat, the fact remains that the present study failed to find even a main effect of sex composition for the dependent variable accuracy. Once more, the question arises of why such an effect could not be demonstrated in the current study.

In terms of a strictly moderating effect of sex composition on stereotype threat, the absence of an observable stereotype threat effect in the present sample may have simply made it impossible to adequately investigate. As such, the present study delivers no direct evidence against such a moderating effect existing, and future studies that do find a notable stereotype threat effect should also take sex composition into account. But as mentioned above, the current study also failed to demonstrate an independent main effect of sex composition on accuracy. One explanation for this could be that there were simply not enough male participants in the individual groups for the women to feel intimidated by them. This seems unlikely, however, given that all groups bar one included at least one male student, and most groups featured more than a single male (see Table 2). According to the results of one study, just a single male being present already led to a significant decline in women's performance (Inzlicht & Ben-Zeev, 2000), meaning the groups of the present study would have had a sufficiently strong male presence to induce the proposed effect.

However, the cited study was conducted over two decades ago at the time of this writing, and there may have since been noteworthy changes in how women perceive men in an academic context. It is possible that the women of today have internalised that men and women can coexist as academic equals, and thus are no longer as intimidated by the former. Contradicting the above cited findings, it could also be that the effect only occurs when the sex ratio more heavily favours the men. This would make it difficult to investigate with the current sample, because all but five out of the 19 groups included more women than men. Lastly, there is always the possibility that a male presence never truly had a negative influence on female performance, but that previous studies only reported such an effect due to circumstances connected to publication bias (Ferguson & Brannick, 2012) or suboptimal statistical design. Both this interpretation and the one that suggests a genuine decline in the proposed effect over time are consistent with the findings of Doyle and Voyer (2016).

The evidence supporting stereotype threat has in the past been inconsistent at best – with studies such as the original by Johns et al. (2005) reporting strong indications of the described effect negatively influencing women's performance in mathematics, and many other studies and replications (Stoet & Geary, 2012) failing to find the same effect. This uncertainty is precisely why the scientific community stands to benefit from modern replication efforts such as the current study, as well as the overarching RRR proposed by Stoevenbelt et al. (2021). After all, a more definitive understanding of how sex-specific stereotypes influence women could have meaningful implications for the current state of education and academia. If stereotype threat were in fact proven detrimental to women's performance in mathematics, research would need to focus more on how its disruptive influence can be alleviated – be it through stereotype threat education or other means. Such methods could then eventually be applied in educational settings, to give all women their best chance and to further the equality of both sexes in modern academia. As such, understanding stereotype threat is not only a matter of scientific interest, but holds significant societal value, also.

All this being said, the current study provided no evidence of stereotype threat existing within its observed sample. This is in line both with the results of previous failed replications (Stoet & Geary, 2012), as well as with the expectations of the proposing authors of the RRR (Stoevenbelt et al., 2021). The latter specifically expected publication bias to be, at least partially, responsible for any non-significant results and smaller effect sizes of the to-be-replicated effects when compared to those of the original (Flore & Wicherts, 2015). However, this is not the only theory that should be considered as a possible cause for the reported difference in results: Stereotype threat may have declined over time for example, or it may simply not manifest as strongly in certain cultures. In any case, the current study was conducted with a sufficiently large sample and with an adequate statistical approach, which took the inherent data dependencies into account so that there would be no overestimation of any discovered effects. As such, it was able to provide substantial evidence towards stereotype threat currently not having any kind of negative influence on the mathematical performance of female psychology students under the age of 26 in Austria. There appears to be no need of incorporating stereotype threat education into local academic settings, because female students perform equally well without it. Beyond that even, no evidence was found of the participants' sex in any way predicting their performance, indicating that there may currently be no sex-specific performance gap at Austrian, or at least Viennese universities. Future studies – such as the RRR by Stoevenbelt et al. (2021) – should investigate whether these conclusions can be generalised to other countries and cultures, as well as to different age groups, disciplines, and contexts outside of tertiary education.

Whatever the reasons for it, stereotype threat appears to have an, at the very least, less notable influence on women's performance than was once thought – a conclusion which would not have been reached, had the scientific community not more openly embraced the value of replication studies in recent years (e.g., Open Science Collaboration, 2015). As such, one can only hope that other supposedly well-established psychological effects will be subject to similar replication efforts in the future.

References

Acock, A. C. (2014). *A Gentle Introduction to Stata* (4th ed.). Stata Press.

Aronson, J., Fried, C. B., & 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. https://doi.org/10.1006/jesp.2001.1491

Association for Psychological Science. (2021, August 12). *Registered Replication Reports*. https://www.psychologicalscience.org/publications/replication

Begley, C., & Ellis, L. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531–533. https://doi.org/10.1038/483531a

Blanz, M. (2015). *Forschungsmethoden und Statistik für die Soziale Arbeit: Grundlagen und Anwendungen*. Kohlhammer.

Bolthausen, E., & Wüthrich, M. V. (2013). Bernoulli's law of large lumbers. *Astin Bulletin, 43*, 73-79. https://doi.org/10.1017/asb.2013.11

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*, 246–257. https://doi.org/10.1037//0022-3514.76.2.246

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic. https://doi.org/10.4324/9780203771587

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203774441

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*(12), 1615–1628. https://doi.org/10.1177/014616702237644

Doyle, R. A., & Voyer, D. (2016). Stereotype manipulation effects on math and spatial test performance: A meta-analysis. *Learning and Individual Differences, 47*, 103–116. https://doi.org/10.1016/j.lindif.2015.12.018

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist, 75*(3), 301–315. https://doi.org/10.1037/amp0000494

Fanelli, D. (2010). Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PLoS ONE, 5*(4), e10271. https://doi.org/10.1371/journal.pone.0010271

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. https://doi.org/10.1037/a0024445

Flore, P. C., Mulder, J., & Wicherts, J. M. (2019). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(5), 1-35. https://doi.org/10.1080/23743603.2018.1559647

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Geary, D. C. (2010). Male, female: The evolution of human sex differences (2nd ed.). American Psychological Association. https://doi.org/10.1037/12072-000

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbache, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, *21*(1), 33–46. https://doi.org/10.2307/749455

Huang, J., Gates, A. J., Sinatra, R., & Barabasi, A. L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(9), 4609-4616. https://doi.org/10.1073/pnas.1914221117

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155. https://doi.org/10.1037/0033-2909.107.2.139

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why women are susceptible to experience problem-solving deficits in the presence of men. *Psychological Science*, *11*, 365–371. https://doi.org/10.1111/1467-9280.00272

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*(3), 467–481. https://doi.org/10.1037/a0018951

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jamieson, J. P., & Harkins, S. G. (2009). The effect of stereotype threat on the solving of quantitative GRE problems: A mere effort interpretation. *Personality and Social Psychology Bulletin*, *35*(10), 1301–1314. https://doi.org/10.1177/0146167209335165

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science, 16*(3), 175-179. https://doi.org/10.1111/j.0956-7976.2005.00799.x

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and Social Psychology Bulletin*, *18*, 302-318. https://doi.org/10.1177/0146167292183006

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin*, *28*, 1183–1193. https://doi.org/10.1177/01461672022812004

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40-48. https://doi.org/10.1037//0021-843X.110.1.40

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, J. L., & White, P. H. (2001). Development of the Domain Identification Measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement, 61*, 1040-1057. https://doi.org/10.1177/00131640121971635

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. https://doi.org/10.1006/jesp.1998.1373

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

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. https://doi.org/10.1037/0022-3514.69.5.797

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*(1), 93–102. https://doi.org/10.1037/a0026617

Stoevenbelt, A. H., Flore, P. C., & Wicherts, J. M. (2021, August 17). *Registered Replication Report Stereotype Threat.* OSF. https://osf.io/69gxp/

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*, 1213–1227. https://doi.org/10.1037//0022-3514.77.6.1213

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin, 37*(8), 1055–1067. https://doi.org/10.1177/0146167211406506

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*(3), 252-260. https://doi.org/10.1056/NEJMsa065779

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127*(6), 797–826. https://doi.org/10.1037/0033-2909.127.6.797

Appendix

Abstract - English

As part of a large-scale Registered Replication Report, a well-known publication on stereotype threat from 2005 was replicated in the present study. For this purpose, 166 psychology students (112 female / 54 male) under the age of 26 were recruited in Austria for a 2x3 between-subjects experimental design with sex and experimental condition as its factors. Participants were randomly assigned at group level to one of three conditions, and would then complete a short mathematics test. The ST condition was meant to induce stereotype threat in the female students by characterising the test as diagnostic for arithmetic ability and by having them make note of their sex in advance. The STE condition was largely the same but attempted to mitigate the negative effects by educating the women on stereotype threat and how it might affect them. The PS condition served as a control group where participants were told the test was a neutral problem solving task and only had to write down their sex after completing the exam. A random-intercept multilevel analysis of various predictor models on participants' accuracy revealed that an “empty” model without any predictors best fit the data, and as such, that there was no significant interaction effect for sex and condition. As this interaction was considered the main to-be-replicated effect, the present study was not able to successfully replicate the results of the original. It is concluded that stereotype threat does not appear to have any notable influence on women’s mathematical performance.

Abstract – German

Als Teil eines großflächigen Registered Replication Reports, wurde eine bekannte Publikation zu Stereotype Threat von 2005 für die diesige Studie repliziert. Zu diesem Zweck wurden 166 Psychologiestudenten (112 weiblich / 54 männlich) unter 26 Jahren in Österreich für ein 2x3 between-subjects experimentelles Design mit den Faktoren Geschlecht und experimentelle Bedingung rekrutiert. Probanden wurden in Gruppen zufällig einer von drei Bedingungen zugeteilt und bearbeiteten dann einen kurzen Mathematiktest. Die ST

Bedingung sollte Stereotype Threat in den weiblichen Studenten induzieren, indem der Test als diagnostisch für arithmetische Fähigkeit charakterisiert wurde und die Probanden ihr Geschlecht bereits vor dem Rechnen notieren mussten. Die STE Bedingung war ähnlich gestaltet, doch sollten die negativen Effekte hier dadurch verringert werden, dass die Frauen über Stereotype Threat und wie es sie beeinflussen könnte aufgeklärt wurden. Die PS Bedingung war eine Kontrollgruppe, in welcher den Probanden gesagt wurde, der Test sei eine neutrale Problemlöseaufgabe und in der sie ihr Geschlecht erst nach Absolvierung des Tests angeben mussten. Eine random-intercept multilevel Analyse für verschiedene Prädiktoren-Modelle auf Bearbeitungsgenauigkeit ergab, dass ein leeres Model mit keinerlei Prädiktoren die Daten am besten erklärte, und dass es somit keine signifikante Interaktion für Geschlecht und Bedingung gab. Da diese Interaktion der zentrale, zu replizierende Effekt war, konnte die diesige Studie die Ergebnisse des Originals nicht erfolgreich replizieren. Es wird geschlussfolgert, dass Stereotype Threat keinen bedeutenden Einfluss auf die mathematische Leistung von Frauen zu haben scheint.

Appendix A

Experiment

Kognitive Prozesse der Problemlösungskompetenz

Liebe*r Teilnehmer*in,

Unsere Forschung konzentriert sich auf die zugrundeliegenden kognitiven Prozesse Ihrer Problemlösungskompetenz. Um uns bei unserer Forschung zu helfen, möchten wir Sie bitten, einige Fragen zu beantworten, die mit Problemlösungskompetenzen zusammenhängen.

Die Untersuchung wird ca. 45 Minuten dauern, während der Sie 20 Minuten Zeit haben, unsere experimentelle Aufgabe zu lösen. Zunächst hören Sie eine kurze Audioaufnahme mit Anweisungen für die Aufgabe. Danach wird die Versuchsleitung die wichtigsten Teile der Anweisungen wiederholen. Die Versuchsaufgabe selbst wird 30 Fragen enthalten, die sich auf Ihre Problemlösungsfähigkeiten beziehen. Wir möchten Sie bitten, keine Fragen zu überspringen und Raten zu vermeiden. Nach der experimentellen Aufgabe erhalten Sie einige Fragen zu Ihrem Eindruck von der Aufgabe und zu einigen Ihrer demographischen Informationen.

Wir erwarten keine negativen Folgen dieser Studie.

Ihre Teilnahme ist vollkommen freiwillig und Sie können die Studie jederzeit ohne negative Folgen abbrechen. Ihre Zustimmung zur Teilnahme an dieser Studie ist bis zum Ende der Studie gültig. Ihre Daten werden anonym behandelt und sind nur mit einem Versuchspersonencode verknüpft. Die Daten sind in keiner Weise auf Sie zurückführbar. Die Daten dieser Forschung werden online, auf dem Open Science Framework, veröffentlicht und könnten daher in anderen Forschungsarbeiten verwendet werden. Darüber hinaus werden Ihre Daten für die kommenden 10 Jahre gespeichert.

Die Ethikkommission der Universität Wien hat die Genehmigung für diese Forschung erteilt. Bei Fragen oder Beschwerden können Sie sich an Ass.-Prof. Mag. Dr. Jakob Pietschnig (jakob.pietschnig@univie.ac.at) wenden.

Unter Berücksichtigung dieser Informationen möchten wir Sie bitten, an unserer Studie teilzunehmen und die Einverständniserklärung zu unterschreiben.

Vielen Dank!

Für weitere Informationen über diese Studie können Sie auch mit uns Kontakt aufnehmen:

Lorraine Phillips, lp@seneca-software.de

Appendix B

Einverständiserklärung

Ich stimme Folgendem zu:

Ich habe die Informationen in dem Informationsbrief sorgfältig gelesen und habe die Möglichkeit erhalten, Fragen über eventuelle Unsicherheiten, die ich bezüglich des Briefes oder des Experiments selbst hatte, zu stellen.

Ich weiß, dass meine Teilnahme freiwillig ist und dass ich die Studie jederzeit ohne negative Folgen und ohne Erklärung abbrechen kann. Außerdem gebe ich meine Zustimmung zur Teilnahme an dieser Studie.

Ich erlaube die Verwendung meiner Daten, die nicht auf mich zurückführbar sind, und die

Speicherung der Daten für einen Zeitraum von 10 Jahren. Außerdem gebe ich meine Zustimmung zur Veröffentlichung meiner anonymen Daten auf dem Open Science Framework.

Appendix C

Debriefing

Liebe*r Teilnehmer*in,

Vielen Dank für Ihre Teilnahme an unserem Experiment. Wie Sie vielleicht beim Ausfüllen der Aufgabe und der Fragebögen bemerkt haben, hat sich die aktuelle Studie nicht mit den kognitiven Prozessen befasst, die menschlichen Problemlösungsfähigkeiten zugrunde liegen.

Wir interessierten uns für die Auswirkungen des „Stereotype Threat" auf Frauen, die mit einem Mathematik-Test konfrontiert wurden. Vielen Studien zufolge schneiden Frauen in Mathematikaufgaben schlechter ab, wenn sie mit dem Stereotyp konfrontiert werden, dass Frauen weniger gut in Mathematik sind als Männer.

Wir wollten diesen Stereotyp in der Studie hervorrufen, indem wir die experimentelle Aufgabe entweder als Problemlösungs- oder als Mathematikaufgabe einführten. Außerdem fragten wir die Teilnehmer*innen in den mathematischen Bedingungen nach ihren demographischen Informationen, bevor sie mit der experimentellen Aufgabe begannen. Die Teilnehmenden in der Problemlösungsbedingung erhielten diese Fragen nach der experimentellen Aufgabe. Auf diese Weise hofften wir, die Leistung der Frauen bei der experimentellen Aufgabe zu beeinflussen.

Vielleicht haben Sie die Erfahrung gemacht, dass die Mathematik-Fragen schwierig waren und das hat vielleicht einige ängstliche Gefühle hervorgerufen. Wir wählten bewusst Fragen aus, die für die Studierenden, die an unserem Experiment teilnahmen, eine Herausforderung darstellten. Daher waren die Fragen vielleicht auch für Sie eine Herausforderung. Darüber hinaus war die vorgegebene Zeitgrenze auch bewusst kurz. Unsere Fragen stammen aus dem „American Graduate Record Examinations", einem Test, der zur Auswahl von Studierenden für „Graduate Schools" (Bildungseinrichtung mit Master- und Promotionsprogrammen) dient. Bei der echten GRE-Prüfung erhalten die Studierenden 35 Minuten Zeit, um 20 Fragen zu beantworten. Daher ist es nicht überraschend, dass Sie nicht alle Fragen beantworten konnten (wir baten Sie, 30 Fragen in 20 Minuten zu beantworten) und dass Sie möglicherweise einige Schwierigkeiten bei der Beantwortung der Fragen hatten.

Dieses Experiment ist Teil eines groß angelegten, registrierten Replikationsberichts. In der Studie wollen wir die Wirkung von Stereotypen in möglichst vielen Laboren und Ländern untersuchen. Alle Informationen zu diesem Projekt werden online auf dem Open Science Framework (https://osf.io/63gs9/) veröffentlicht. Sollten Sie noch Fragen haben, können Sie Lorraine Phillips unter lp@seneca-software.de kontaktieren.

Da wir noch nicht alle Teilnehmenden getestet haben und das Wissen um den Effekt des „Stereotype Threat“ unsere Ergebnisse beeinflussen kann, möchten wir Sie bitten, den wahren Zweck dieser Studie nicht mit anderen Studierenden zu diskutieren.

Vielen Dank für Ihre Zeit!

Appendix D

Stereotype Threat (ST) Bedingung:

„Guten Tag und vielen Dank, dass Sie heute hier sind. Mein Name ist Dr. Hamelang. Lassen Sie mich Ihnen von der Studie erzählen, an der Sie gleich teilnehmen werden. Mein primäres Forschungsinteresse gilt der Erfassung mathematischer Fähigkeiten. In den letzten Jahren habe ich Geschlechtsunterschiede in Bezug auf mathematische Fähigkeiten untersucht.

In der heutigen Sitzung wollen wir Ihre mathematischen Fähigkeiten messen, indem wir Sie einen standardisierten Mathematiktest bearbeiten lassen. Der Test, den Sie bearbeiten werden, ist ein gut etabliertes Instrument zur Messung mathematischer Begabung und Ihre Leistung wird uns eine klare Einschätzung Ihrer natürlichen mathematischen Fähigkeiten vermitteln.

Einige der Fragen mögen Ihnen vielleicht schwierig erscheinen, jedoch liegen sie alle im Rahmen der Fähigkeiten der meisten Studierenden. Wir bitten Sie, diesen Test ernst zu nehmen und sich wirklich zu bemühen, damit wir akkurate Daten erheben können. Ihre Leistung in diesem Test wird uns helfen, Leistungsnormen für Männer und Frauen zu bestimmen. Nach dem Test werden wir Ihnen Feedback zu Ihrer Leistung geben und Ihnen einige Fragen über ihre Erfahrungen bei der Testbearbeitung stellen.“

Appendix E

Stereotype Threat Education (STE) Bedingung:

„Guten Tag und vielen Dank, dass Sie heute hier sind. Mein Name ist Dr. Hamelang. Lassen Sie mich Ihnen von der Studie erzählen, an der Sie gleich teilnehmen werden. Mein primäres

Forschungsinteresse gilt der Erfassung mathematischer Fähigkeiten. In den letzten Jahren habe ich Geschlechtsunterschiede in Bezug auf mathematische Fähigkeiten untersucht.

Aktuelle Forschungsergebnisse legen nahe, dass Frauen bei mathematischen Aufgaben manchmal schlechter als Männer abschneiden, NICHT, weil sie weniger Fähigkeiten aufweisen, sondern weil es negative Stereotype gibt, die nahelegen, dass Frauen schlechter abschneiden *sollten*, da sie für Mathe nicht so „verdrahtet“ seien wie Männer. Mit anderen Worten: das bloße Bewusstsein, dass negative Stereotype über die mathematischen Fähigkeiten von Frauen existieren, kann tatsächlich eine Rolle dabei spielen, die Leistungen von Frauen bei standardisierten Mathematiktests zu beeinträchtigen. Bedenken Sie, dass ein niedriges Ergebnis von Frauen in einem Mathematiktest von jedweder Person, die das Ergebnis sieht, verwendet werden könnte, um Generalisierungen über die mathematischen Fähigkeiten von Frauen im Allgemeinen vorzunehmen. Wenn Sie also erkennen, dass ein niedriges Ergebnis in einem Mathematiktest als Evidenz verwendet werden könnte, dass Frauen nicht gut in Mathematik sind, könnte dies zusätzlichen Druck erzeugen, sich zu bemühen und gut abzuschneiden. Die Forschung hat gezeigt, dass dieser zusätzliche durch das Stereotyp ausgelöste Druck, tatsächlich die Konzentration in einem Ausmaß beeinträchtigen kann, dass Frauen bei einem Mathematiktest schlechter abschneiden, als sie es sonst würden. Das kann sogar Frauen passieren, die sehr begabt in Mathematik sind.

In der heutigen Sitzung werden Sie einen Test bearbeiten, der gut validiert wurde als Maß für Ihre natürlichen mathematischen Fähigkeiten. Aber wenn Sie eine Frau sind, ist es wichtig, nicht zu vergessen, falls Sie sich bei der Bearbeitung dieses Tests ängstlich fühlen, dass diese Angst ein Ergebnis dieser negativen Stereotype sein könnte, die in der Gesellschaft weithin bekannt sind und nichts mit Ihrer tatsächlichen Fähigkeit zu tun haben, den Test gut zu absolvieren. Sollten Sie sich ein wenig ängstlich fühlen, ist das komplett normal und Sie sollten nicht annehmen, dass es bedeutet, dass Sie in dem Test schlecht abschneiden. Nach dem Test werden wir Ihnen Feedback zu Ihrer Leistung geben und Ihnen einige Fragen über ihre Erfahrungen bei der Testbearbeitung stellen.“

Appendix F

Problem Solving (PS) Bedingung:

„Guten Tag und vielen Dank, dass Sie heute hier sind. Mein Name ist Dr. Hamelang. Lassen Sie mich Ihnen von der Studie erzählen, an der Sie gleich teilnehmen werden. Mein primäres Forschungsinteresse besteht darin, die kognitiven Prozesse zu verstehen, die der Problemlösungen zu Grunde liegen. In den letzten Jahren habe ich individuelle Unterschiede zwischen Menschen bei der Lösung von Problemen untersucht.

In der heutigen Sitzung hätten wir gerne, dass Sie eine Problemlöseaufgabe bearbeiten. Diese Aufgabe ist nicht diagnostisch für irgendeine Fähigkeit – die Aufgabe ist lediglich eine einfache Laborübung, die es uns erlaubt zu untersuchen, wie Menschen an Problemlösungen arbeiten.

Einige der Fragen mögen Ihnen vielleicht schwierig erscheinen, jedoch liegen sie alle im Rahmen der Fähigkeiten der meisten Studierenden. Wir bitten Sie, diese Aufgabe ernst zu nehmen und sich wirklich um die Lösung der Probleme zu bemühen, damit wir korrekte

Informationen erheben können. Ihre Leistung wird uns helfen die verschiedenen Faktoren zu verstehen, die mit Problemlöseprozessen zusammenhängen. Im Anschluss werden wir Ihnen Feedback zu Ihrer Leistung geben und Ihnen einige Fragen zu den Problemlöseaufgaben stellen.“

Appendix G

Table 15

*Item difficulties, item-rest correlations and alpha-if-deleted for the GRE items (*N *= 159)*

Item	Difficulty / Solution frequency	Item-rest-correlations (bis)	Alpha-if-deleted
1	.799	.289	.721
2	.748	.184	.727
3	.748	.214	.725
4	.176	.034	.733
5	.478	-.008	.742
6	.170	.095	.730
7	.642	.273	.721
8	.742	.248	.723
9	.509	.447	.708
10	.440	.579	.698
11	.428	.525	.703
12	.346	.699	.691
13	.308	.717	.691
14	.233	.513	.708
15	.264	.661	.697
16	.094	.675	.710
17	.189	.536	.708
18	.088	.777	.707
19	.082	.761	.709
20	.025	.880	.717
21	.031	.505	.721
22	.013	.445	.723
23	.013	.559	.723

Item	Difficulty / Solution frequency	Item-rest-correlations (bis)	Alpha-if-deleted
24	.013	.388	.724
25	.000	*NA*	.725
26	.000	*NA*	.725
27	.000	*NA*	.725
28	.000	*NA*	.725
29	.000	*NA*	.725
30	.000	*NA*	.725