Journal of School Psychology 53 (2015) 25–44

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis☆

Paulette C. Flore *, Jelte M. Wicherts

Tilburg University, The Netherlands

ARTICLE INFO

Article history:
Received 26 November 2013
Received in revised form 24 October 2014
Accepted 25 October 2014
Available online 13 November 2014

Keywords:
Stereotype threat
Math/science test performance
Gender gap
Test anxiety
Publication bias
Meta-analysis

ABSTRACT

Although the effect of stereotype threat concerning women and mathematics has been subject to various systematic reviews, none of them have been performed on the sub-population of children and adolescents. In this meta-analysis we estimated the effects of stereotype threat on performance of girls on math, science and spatial skills (MSSS) tests. Moreover, we studied publication bias and four moderators: test difficulty, presence of boys, gender equality within countries, and the type of control group that was used in the studies. We selected study samples when the study included girls, samples had a mean age below 18 years, the design was (quasi-)experimental, the stereotype threat manipulation was administered between-subjects, and the dependent variable was a MSSS test related to a gender stereotype favoring boys. To analyze the 47 effect sizes, we used random effects and mixed effects models. The estimated mean effect size equaled −0.22 and significantly differed from 0. None of the moderator variables was significant; however, there were several signs for the presence of publication bias. We conclude that publication bias might seriously distort the literature on the effects of stereotype threat among schoolgirls. We propose a large replication study to provide a less biased effect size estimate.

1. Introduction

Spencer, Steele, and Quinn (1999) first suggested that women's performance on mathematics tests could be disrupted by the presence of a *stereotype threat*. This initial paper inspired many researchers to replicate the stereotype threat effect and expand the theory by introducing numerous moderator variables and various dependent variables related to negative gender stereotypes, such as tests of Mathematics, Science, and Spatial Skills (MSSS). This practice resulted in approximately one hundred research papers and five meta-analyses (Nguyen & Ryan, 2008; Picho, Rodriguez, & Finnie, 2013; Stoet & Geary, 2012; Walton & Cohen, 2003; Walton & Spencer, 2009). Although four of these systematic reviews (Nguyen & Ryan, 2008; Picho et al., 2013; Walton & Cohen, 2003; Walton & Spencer, 2009) confirmed the existence of a robust mean stereotype threat effect, some ambiguities regarding this effect remain. For instance, it has been suggested (*Ganley et al., 2013; Stoet & Geary, 2012) that the stereotype threat literature is subject to an *excess of significant findings*, which might be caused by publication bias (Ioannidis, 2005; Rosenthal, 1979), *p-hacking* (i.e., using questionable research practices to obtain a statistically significant effect; Simonsohn, Nelson, & Simmons, 2013), or both (Bakker, van Dijk, & Wicherts, 2012). A less controversial but nevertheless interesting issue is the age at which stereotype threat begins to influence performance on MSSS tests: does stereotype threat already influence children's performance, or does this effect

☆ The preparation of this article was supported by grant numbers 016-125-385 and 406-12-137 from the Netherlands Organization for Scientific Research (NWO).
* Corresponding author at: Department of Methodology and Statistics, Tilburg School of Behavioral and Social Sciences, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands.
E-mail address: P.C.Flore@tilburguniversity.edu (P.C. Flore).
Action Editor: Craig Albers

http://dx.doi.org/10.1016/j.jsp.2014.10.002
0022-4405/

only emerge during early adulthood? Both of these issues are addressed in this article by means of a meta-analysis of the stereotype threat literature in the context of schoolgirls' MSSS test performance. We will introduce these topics by providing a general review of the literature on stereotype threat and the onset of gender differences in the domains of MSSS.

1.1. Stereotype threat

The effect of stereotype threat refers to the ramifications of an activated negative stereotype or an emphasized social identity (Steele, 1997). Individuals who are members of a stigmatized group tend to perform worse on stereotype relevant tasks when confronted with that negative stereotype (Steele & Aronson, 1995). In their seminal paper, Steele and Aronson (1995) focused on ethnic minorities as stereotyped group. Later experiments showed similar effects for other stigmatized groups, including women in the quantitative domain (e.g., Ambady, Paik, Steele, Owen-Smith, & Mitchell, 2004; Brown & Josephs, 1999; Oswald & Harvey, 2001; Schmader & Johns, 2003; Spencer et al., 1999). In these experiments, women were either assigned to a stereotype threat condition, where they were exposed to a gender-related stereotype threat (e.g., a written statement that men perform better on mathematics tests than women), or to a control condition, where they were not exposed to such a threat. When participants subsequently completed a MSSS test (e.g., a mathematical test), women who were assigned to the stereotype threat condition averaged lower scores than women who were assigned to the control condition (Ambady et al., 2004; Brown & Josephs, 1999; Oswald & Harvey, 2001; Schmader & Johns, 2003; Spencer et al., 1999). The results of these studies were deemed important, because researchers suspected that stereotype threat could be a driving force behind the decision of women to leave the science, technology, engineering, and mathematics (STEM) fields (Cheryan & Plaut, 2010; Schmader, Johns, & Barquissau, 2004). These developments led to an expansion of the stereotype threat literature, in which several moderator and mediator variables were studied.

Of all the studied moderator and mediator variables, we will summarize those variables that have been studied most frequently. Item difficulty appears to moderate the effects of stereotype threat, with difficult items leading to stronger effects (Campbell & Collaer, 2009; O'Brien & Crandall, 2003; Spencer et al., 1999; Wicherts, Dolan, & Hessen, 2005). Test-takers who are strongly identified with the relevant domain, in this case the domain of mathematics, science or spatial skills, appear to show stronger stereotype threat effects (Cadinu, Maass, Frigerio, Impagliazzo, & Latinotti, 2003; Lesko & Corpus, 2006; Pronin, Steele, & Ross, 2004; Steinberg, Okun, & Aiken, 2012). Another theoretical moderator is gender identification; the effects of stereotype threat are generally more severe for women who are highly gender-identified (Kiefer & Sekaquaptewa, 2007; Rydell, McConnell, & Beilock, 2009; Schmader, 2002; Wout, Danso, Jackson, & Spencer, 2008). However, the latter results were contradicted in a Swedish study (Eriksson & Lindholm, 2007). Moreover, the effects of stereotype threat appear stronger within a threatening environment (e.g., in the presence of men, or when negatively stereotyped test-takers hold a minority status) compared to a safe environment (e.g., in the presence of women only, or when holding a majority status; Gneezy, Niederle, & Rustichini, 2003; Inzlicht, Aronson, Good, & McKay, 2006; Inzlicht & Ben-Zeev, 2003; Sekaquaptewa & Thompson, 2003). The presence of role models also appears to moderate the effect of stereotype threat, in such a way that role models that contradict the stereotype (i.e., women who are good in mathematics or men who lack mathematical skills) appear to protect females from the debilitating effects of stereotype threat on MSSS test performance (Elizaga & Markman, 2008; Marx & Ko, 2012; Marx & Roman, 2002; McIntyre, Paulson, Taylor, Morin, & Lord, 2011; Taylor, Lord, McIntyre, & Paulson, 2011). Finally, several researchers suggested that the stereotype threat effect is (partly) mediated by arousal (Ben-zeev, Fein, & Inzlicht, 2005), anxiety and worries (Brodish & Devine, 2009; Ford, Ferguson, Brooks, & Hagadone, 2004; Gerstenberg, Imhoff, & Schmitt, 2012; Osborne, 2001, 2007), or the occupation of working memory (Beilock, Rydell, & McConnell, 2007; Bonnot & Croizet, 2007; Rydell, Rydell, & Boucher, 2010; Schmader & Johns, 2003).

The literature on the effects of stereotype threat has been summarized by five meta-analyses that covered heterogeneous subsets of studies (Nguyen & Ryan, 2008; Picho et al., 2013; Stoet & Geary, 2012; Walton & Cohen, 2003; Walton & Spencer, 2009). These broad-stroke meta-analyses estimated a small to medium significant effect before moderators were taken into account, with standardized mean differences ranging from 0.24 (Picho et al., 2013) to 0.48 (Walton & Spencer, 2009). These findings seemed to confirm that the effect is rather stable, although most of these meta-analyses reported heterogeneity in effect sizes (Picho et al., 2013; Stoet & Geary, 2012; Walton & Cohen, 2003). In fact, the previous meta-analyses included diverse tests, settings, and stereotyped groups, which makes it hard to pinpoint exactly why some studies show larger effects than others. Although these large scale meta-analyses are interesting to portray an overall picture, a more homogeneous subset of studies is preferred when dealing with specific questions, like the degree to which the stereotype threat related to gender also influences MSSS performance in schools. Thus, we addressed this issue by selecting a specific stereotyped group and stereotype (i.e., women and their supposed inferior capacity of solving mathematical or spatial tasks) and a specific age group (i.e., those younger than 18 years), which should result in a less heterogeneous set of effect sizes. These design elements enabled us to describe the influence of stereotype threat on MSSS test performance for females in critical periods of human development, namely childhood and adolescence.

1.2. Stereotype threat and children

Although the effects of stereotype threat on women was traditionally studied within adult populations (Spencer et al., 1999), multiple studies over the last 15 years have been carried out with children and adolescents as participants (e.g., Ambady, Shih, Kim, & Pittinsky, 2001; *Keller & Dauenheimer, 2003). Studies on children and adolescents in schools contribute to the literature for at least three reasons: (1) to find out at which age the stereotype threat effect actually emerges, (2) to study the stereotype threat effect in the natural setting of the classroom instead of the laboratory setting, and (3) to address the question whether variables that moderate the stereotype threat effect in adult samples similarly moderate the stereotype threat effect among children.

Table 1
Types of stereotype threat manipulations.

Manipulation condition	Manipulation	Example	Examples of papers
Explicit	Verbal or written statement that boys are superior to girls on the test	"It [the test] is comprised of a collection of questions which have been shown to produce gender differences in the past. Male participants outperformed female participants."	*Cherney and Campbell (2011), *Keller and Dauenheimer (2003)
Explicit	Verbal statement that boys are really good in the test	"Boys are really good at this game."	*Cimpian, Mu, and Erickson (2012)
Implicit	Participants filling out their gender	–	*Stricker and Ward (2004)
Implicit	Visual depiction of a stereotypical situation	Showed pictures of male scientists/mathematicians	*Good et al. (2010), *Muzzatti and Agnoli (2007)
Implicit	Priming female identity	The story described a girl using a number of traits that were stereotypically feminine in participants' cultural context (e.g., long blond hair, blue eyes, and colorful clothes).	*Tomasetto, Alparone, and Cadinu (2011)
Implicit	Framing the question as a geometric problem	–	*Huguet and Régner (2007), *Huguet and Régner (2009)

The primary research on stereotype threat with children as participants (i.e., studies that we included in our meta-analysis) roughly shared a similar design, although the details of the designs varied somewhat. Typically, the studies were conducted by means of an experiment or a quasi-experiment involving a stereotype threat condition and a control condition as predictor variable, sometimes in combination with a third or fourth condition (*Cherney & Campbell, 2011; *Picho & Stephens, 2012). These conditions were typically designed as a between-subjects factor. Some variations exist in the implementation of the stereotype threat and control conditions. The stereotype threat manipulation was administered either explicitly or implicitly. The explicit stereotype threat manipulation usually involved a written or verbal statement that informed participants that the MSSS test they were about to complete produced gender differences, whereas the implicit stereotype threat manipulations triggered the gender stereotype without explicitly mentioning the gender gap. Further examples of the two types of stereotype threat manipulations are illustrated in Table 1. The control condition was designed to either nullify or not nullify stereotype threat. In the nullified control condition the stereotype threat was actively removed, generally by a written or verbal statement which informed participants that the MSSS test they were about to complete did not produce gender differences, whereas in the non-nullified control condition no gender related information was provided. Further examples of the two types of control conditions are illustrated in Table 2.

The outcome measure in studies of stereotype threat among schoolgirls to date were MSSS tests; most studies involved a mathematical test properly adjusted to the age and ability level of the participants (e.g., *Keller & Dauenheimer, 2003; *Muzzatti & Agnoli, 2007). A few studies used the Mental Rotation Task (e.g., *Moè & Pazzaglia, 2006; *Neuburger, Jansen, Heil, & Quaiser-Pohl, 2012;

Table 2
Types of control conditions.

Control condition	Information	Example	Examples of papers
No Threat	No information given with regards to the relationship between gender and performance on the test	–	*Delgado and Prieto (2008), *Muzzatti and Agnoli (2007)
Nullified	Verbal or written statement that girls are superior to boys on the test	"It is comprised of a collection of questions which have been shown not to produce gender differences in the past. The average achievement of male participants was equal to the achievement of female participants."	*Cherney and Campbell (2011)
Nullified	Verbal or written statement that girls and boys perform equally well on the test	"In such tasks, boys and girls are equally skilled. Both have an equal ability to imagine how pictures and objects look when they are rotated. Therefore, such tasks are exactly equally difficult or easy for girls and boys."	*Neuburger et al. (2012)
Nullified	Education about the stereotype threat effect	"Research has shown that men perform better than women in this test and obtain higher scores. This superiority is caused by a gender stereotype, i.e., by a common belief in male superiority in spatial tasks, and has nothing to do with lack of ability."	*Moè (2009)
Nullified	Written description of a counter-stereotypical situation	"Marie was described as a successful student in math"	*Bagès and Martinot (2011)
Nullified	Visual depiction of a counter-stereotypical situation	"Participants were randomly assigned to one of three experimental conditions by inviting them to color a picture, in which a girl correctly resolves the calculation whereas a boy fails to respond"	*Galdi et al. (2014)

*Titze, Jansen, & Heil, 2010) which measured children's spatial abilities, a concept tightly linked to mathematics and gender stereotypes. Remaining dependent variables were the performance on a physics test (*Marchand & Taasoobshirazi, 2012), a chemistry comprehension test (*Good, Woodzicka, & Wingfield, 2010) or recall performance of a geometric figure (*Huguet & Régner, 2009). These tests generally consisted of 10 to 40 questions.

1.3. Developmental aspects of stereotype threat

The onset and development of the effects of stereotype threat on girls in mathematics throughout the life course is an interesting issue; however, few solid conclusions have been reached (Aronson & Good, 2003; Jordan & Lovett, 2007). To explore possible theories on how age might influence stereotype threat, we recollect the most important moderators that were identified in the research on young adults and subsequently consider whether these could influence stereotype threat differently throughout the development of children. The most important moderators among adults are gender identification, domain identification, stigma consciousness, and beliefs about intelligence (Aronson & Good, 2003). Thus, women who strongly identify with both the academic domain of mathematics (Cadinu et al., 2003; Lesko & Corpus, 2006; Pronin et al., 2004; Steinberg et al., 2012) and the female gender (Kiefer & Sekaquaptewa, 2007; Rydell et al., 2009; Schmader, 2002; Wout et al., 2008) are expected to experience stronger performance decrements compared to women who less strongly identify with those domains. Additionally, women who believe that the stereotypes regarding women and mathematics are true (Schmader et al., 2004) and that mathematical ability is a stable and fixed characteristic (Aronson & Good, 2003) are purported to show stronger stereotype threat effects. The current knowledge about the development of these four traits can be used as guidance for the expectations of the impact of stereotype threat throughout different age groups (Aronson & Good, 2003).

1.4. Gender identification

Gender identification is present at an early age. At the age of 3 years, a majority of children are able to correctly label themselves to their gender (Katz & Kofkin, 1997). A study on 3- to 5-year-olds (Martin & Little, 1990) showed that these children are not only able to correctly label their gender and distinguish men from women but also prefer sex-typed toys that correspond to their gender (i.e., boys preferring masculine sex-typed toys and girls preferring feminine sex-typed toys). When children reach the age of 6 to 7 years, they master the concept of gender constancy; and so understand that gender is stable over time and consistent (Bussey & Bandura, 1999). Based on these studies one could argue that because gender identity is already stable at a young age, even young children are potentially vulnerable to performance decrements caused by stereotype threat. However, Aronson and Good (2003) proclaimed that although children are already aware of their gender from an early age on, they do not form a coherent sense of the self until adolescence, which prevents younger children from vulnerability to stereotype threat.

1.5. Stigma consciousness

The studies on development of awareness of the stereotype (stigma consciousness) have showed mixed results. Various studies showed that children believe that boys are either better in mathematics or are identified more strongly with the field of mathematics compared to girls, for ages 6 to 11 (Cvencek, Meltzoff, & Greenwald, 2011; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Lummis & Stevenson, 1990) and ages 14 and 22 (Steffens & Jelenec, 2011). In Steffens and Jelenec (2011), older participants endorsed the stereotypes more strongly than the younger participants. A meta-analysis on affects and attitudes concerning mathematics showed that adolescents and young adults from different age groups (11 to 25 years old) all see mathematics more as a male domain (Hyde, Fennema, Ryan, Frost, & Hopp, 1990). These gender stereotypes are also present in the classroom; teachers tend to see boys as more competent in mathematics (Li, 1999), they expect mathematics to be more difficult for girls (Tiedemann, 2000), and they expect that failure in mathematics for girls more likely originates from a lack of ability, whereas failure for boys originates from a lack of effort (Fennema, Peterson, Carpenter, & Lubinski, 1990; Tiedemann, 2000). However, counterintuitive evidence regarding stigma consciousness has also been found more recently: some studies failed to find convincing evidence that children explicitly believe in the traditional stereotype (Ambady et al., 2001; Kurtz-Costes, Rowley, Harris-Britt, & Woods, 2008), other studies found that children believe in non-traditional stereotypes (Martinot, Bagès, & Désert, 2012; Martinot & Désert, 2007), and another study found that teachers do not hold stereotypical beliefs (Leedy, LaLonde, & Runk, 2003). Additionally a more recent study found that when it comes to overall academic competency 6- to 10-year-olds hold the stereotype that girls outperform boys (Hartley & Sutton, 2013), and these children actually believe that adults hold those stereotypes as well. A stereotype threat manipulation addressing this stereotype actually negatively influenced the performance of boys on a test that included different domains, including mathematics. Moreover, a longitudinal study showed that over different grades, teachers either rated the girls in their classes significantly higher in mathematical ability than boys, or rated girls and boys as roughly equivalent in mathematical ability, even when there was a significant gender gap in performance on a mathematics test favoring males (Robinson & Lubienski, 2011). Some argue that this evidence against the stereotype regarding mathematics and gender in recent studies might indicate that the gender stereotype as we know it is outdated (Martinot et al., 2012). Also, relatively little research has addressed whether gender stereotypes are comparable over time (e.g., during the 1980s vs. during the 2010s) or across different countries or smaller cultural units (as we addressed in the section Moderators).

1.6. Domain identification

Few studies have been conducted on the development of academic identification, or domain identification, in children (Aronson & Good, 2003). A study by *Keller (2007) on 15-year-olds indicated that domain identification moderated the effect of stereotype threat on math performance. Specifically, girls in a stereotype threat condition who considered themselves as low identifiers in the mathematical domain performed better on difficult math items, whereas girls who considered themselves as high identifiers in the mathematical domain performed worse on difficult math items. Although little attention has been given to domain identification in the context of stereotype threat and development, research on affect and attitude of girls towards mathematics over different age groups could provide information on how domain identification might fluctuate. For instance, the gender gap of positive attitudes towards and self-confidence in mathematics is virtually non-existent for children between the ages of 5 to 10 years but grows wider in older age groups, with boys being more positive and self-confident than girls (Hyde et al., 1990). Thus, it seems that, generally, adolescent girls have less confidence in and fewer positive attitudes towards mathematics compared to boys of their age, which might be an indication that older girls also identify themselves less with the mathematical domain. In the context of stereotype threat, this pattern of findings would lead us to expect that adolescent girls are actually less vulnerable to the effects of stereotype threat compared to pre-teenage girls.

1.7. Beliefs about intelligence

The literature on beliefs about intelligence and academic ability describes rather straightforwardly how those beliefs change throughout the development of children. Children younger than 7 years do not yet comprehend that intelligence and ability are personal traits that are stable over time and that the role of effort in academic performance is limited (Droege & Stipek, 1993; Stipek & Daniels, 1990). At this age, children confuse intelligence and ability with social–moral qualities: a good or nice person equals a smart person and vice versa (Droege & Stipek, 1993; Heyman, Dweck, & Cain, 1992). Because young children do not yet see academic abilities as fixed traits, they tend to be overly optimistic about their performances and overestimate their position on academic performances relative to their classmates (Nicholls, 1979). When children reach the age of 7 or 8, their theories seem to shift, in such a way that older children believe in more temporal constant abilities (Kinlaw & Kurtz-Costes, 2003). At this age, the children predict more stable levels of intelligence (Dweck, 2002; Wigfield et al., 1997), and they believe less in the role of effort (Stipek & Daniels, 1990). Additionally, they are better able to distinguish ability from social or moral abilities (Droege & Stipek, 1993; Heyman et al., 1992; Stipek & Daniels, 1990). As a consequence, beginning at approximately age 7 to 8 years, children are less optimistic and more realistic about their future academic performances and their position within the classroom compared to their peers (Eccles et al., 1989; Nicholls, 1979). These findings imply that stereotype threat would only have an effect on children who are at least 7 to 8 years old. If indeed these notions about abilities are crucial for stereotype threat, younger children most likely do not even see mathematical ability as a fixed trait; hence, there would be little reason for them to feel threatened by stereotypes regarding mathematical competency. In contrast, older children would have the capacity to understand that effort will not necessarily compensate for a lack of ability and hence be susceptible to stereotype threat.

Although studies on the development of gender identity, stigma consciousness, and beliefs about intelligence seem to imply that children below the age of 8 or 10 will probably not be influenced by stereotype threat, the line of evidence concerning these potential age-related moderating variables we discussed here is indirect. That is, it is unclear whether moderators that were found to be relevant for stereotype threat among young adults also are relevant among schoolgirls. In addition, the conclusion that children below the age of 8 or 10 will probably not be influenced by stereotype threat is in contrast with the theory on domain identification, which would actually predict the opposite. It is therefore important to collate all the evidence that speaks to the ages at which stereotype threat effects among schoolgirls actually emerge. In our meta-analysis, we therefore (a) explored whether age is a moderator of the stereotype threat effect among schoolgirls and (b) studied the moderators (at the level of studies) that are implicated in stereotype theory as being relevant for stereotype threat.

1.8. Moderators

1.8.1. Test difficulty

In our meta-analyses we considered, in addition to the exploratory moderator of age, four confirmatory moderators on the basis of theory and previous results (Nguyen & Ryan, 2008; Picho et al., 2013; Steele, 2010). The first moderator we hypothesized to have an influence on the effect of stereotype threat is *test difficulty*. Studies on the adult population showed that test difficulty is an important moderator (e.g., Nguyen & Ryan, 2008; Spencer et al., 1999). The moderation of test difficulty on the stereotype threat effect is often explained in terms of arousal (Ben-zeev et al., 2005), although psychometric reasons may also play a role (Wicherts et al., 2005). Studies showed that the stereotype threat effect appears to be mediated by arousal or anxiety (Ben-zeev et al., 2005; *Delgado & Prieto, 2008; Gerstenberg et al., 2012; Osborne, 2001); thus, the more anxious or aroused participants are, the worse they will perform on a mathematical test. Relatively difficult items are more threatening than easy items; therefore, they lead to a higher state of arousal, which in turn will result in a larger gender gap in mathematical test performance (*Delgado & Prieto, 2008; O'Brien & Crandall, 2003). These findings corresponded to traditional findings of social facilitation, which showed that arousal leads to diminished performance on a difficult task, whereas arousal leads to enhanced performance when the task is well learned (Markus, 1978; Zajonc, 1965). The moderating role of test anxiety might be explained by the fact that solving difficult questions requires a larger working memory capacity than solving easy questions (Beilock et al., 2007). When worrying thoughts provoked by stereotype threat occupy part of the

working memory, solving a difficult question becomes problematic, whereas easy questions are still solvable because they do not require a large working memory capacity (Eysenck & Calvo, 1992). This mechanism leads to score reduction for difficult tests but not for easy tests. With the former in mind, we expected that the effect of stereotype threat would be stronger in studies that use a relatively difficult test compared to studies that use a relatively easy test. We defined difficulty here as the degree to which those in the sample answer items in the test correctly. Psychometrically advanced analyses that formally model the item difficulties are beyond the scope of this meta-analysis because they require the raw data.

1.8.2. Presence of boys

The second variable that we predicted to moderate the stereotype threat effect among schoolgirls is the absence or *presence of boys* during test-taking. Several studies showed that female students tend to underperform on negatively stereotyped tasks in the presence of male students who are working on the same task (Gneezy et al., 2003; Inzlicht & Ben-zeev, 2000; Inzlicht & Ben-Zeev, 2003; Picho et al., 2013; Sekaquaptewa & Thompson, 2003). This effect might be explained by the salience of gender identity; gender becomes more salient for women who hold the minority in a group than for women who are in a same-sex group (Cota & Dion, 1986; Mcguire, Mcguire, & Winton, 1979). In turn, the heightened salience of gender identity might lead to stronger effects of stereotype threat. People who hold a minority or token status within a group tend to suffer from cognitive deficits (Lord & Saenz, 1985), a phenomenon that is even registered when women simply watch a gender unbalanced video of a conference in a mathematical domain (Murphy, Steele, & Gross, 2007). The combination of both the activation of gender identity and reduced cognitive performance due to social pressure caused by a minority status then leads to worse performance for women confronted with stereotype threat in a mixed-gender setting. Thus, we predicted the stereotype threat effect among schoolgirls to be stronger in studies in which boys were present during test administration, compared to studies in which no boys were present during test administration.

1.8.3. Cross-cultural gender equality

The third moderator we studied was *cross-cultural gender equality*, or the degree in which women are deemed equal to men in the several nations where the selected stereotype threat studies took place. Recent studies showed marked cross-cultural differences in the gender gap in mathematical performance across countries (Else-Quest, Hyde, & Linn, 2010; Mullis, Martin, Foy, & Arora, 2012; Organisation for Economic Co-operation and Development (OECD), 2010). In the cross-cultural study on 15-year-old students carried out by OECD (i.e., the Programme for International Student Assessment or PISA) within 65 countries boys significantly outperformed girls on the mathematical test in 54% of the countries, whereas in 8% of the countries girls outperformed boys. In 38% of the countries, no significant difference between the two sex groups was found. Comparable are the Trends in International Mathematics and Science Study (TIMSS) studies (Mullis et al., 2012) on fourth graders within 50 countries, in which boys outperformed girls in 40% of the countries, girls outperformed boys in 8% of the countries, and no significant differences were found in 52% of the countries. However, the results of the TIMSS studies for eight graders in 42 countries were different: in 31% of the countries, girls outperformed boys, while in only 17% of the countries, boys outperformed girls, and in 52% of the countries no significant differences emerged. Overall, the sex differences for the majority of countries were quite small. The differences between countries concerning the gender gap in mathematics were proposed to be associated with the gender equality and amount of stereotyping within countries (Else-Quest et al., 2010; Guiso, Monte, & Sapienza, 2008; Nosek et al., 2009). Some studies showed that gender equality is associated with the gender gap in mathematics for school aged children (Else-Quest et al., 2010; Guiso et al., 2008). Gender equality also has as a negative relation with anxiety, and a positive relation with girls' self-concept and self-efficacy concerning the mathematical domain (Else-Quest et al., 2010). In addition, the gender gap in mathematical test performance could be predicted by cross-national differences in Implicit Association Test-scores on the gender–science relation (Nosek et al., 2009). Based on these results, we expected that the stereotype threat effect among schoolgirls would be stronger for studies conducted in countries with low levels of gender equality compared to countries with high levels of gender equality. To operationalize this variable, we used the Gender Gap Index (Hausmann, Tyson, & Zahidi, 2012), which is an index that incorporates economic participation, educational attainment, political empowerment, and health and survival of women relative to men. Higher scores on the GGI indicate a higher degree of gender equality. Geographical regions have been used before as moderator variable in the meta-analysis on stereotype threat and mathematical performance by Picho et al. (2013); however, they only studied regions within the United States of America.

1.8.4. Type of control condition

The last moderator we studied concerned the *type of control condition* participants were assigned to. Stereotype threat experiments involve the use of two or more conditions that differ in stereotype threat, such that conditions can be ranked by severity of stereotype threat. The condition that supposedly ranks lowest on stereotype threat severity is the control condition, which exists either of a situation where participants do not receive *any* gender related information (e.g., *Delgado & Prieto, 2008; *Muzzatti & Agnoli, 2007), or a so-called nullified control condition. This nullified control condition is designed to actively remove the stereotype threat, usually by informing test-takers that girls perform equally well as boys or even that girls outperform boys on the mathematical test (*Cherney & Campbell, 2011; *Neuburger et al., 2012). There are indications that test-takers who are assigned to a nullified control condition outperform those who are assigned to a condition in which no additional information has been given (Campbell & Collaer, 2009; Smith & White, 2002; Walton & Cohen, 2003; Walton & Spencer, 2009). This effect is explained by the fact that whenever women are confronted with a MSSS test their gender identity already becomes salient by the well-known stereotype (Smith & White, 2002); giving no additional information would thus entail a form of implicit threat activation. Therefore, we expected the effect of stereotype threat among schoolgirls to be stronger in studies that involved a nullified control condition compared to studies that involved a control condition without additional information.

1.9. Publication bias and p-hacking

Although the existence of the stereotype threat effect seems widely accepted, there are some reasons to doubt whether the effect is as solid as it is often claimed to be. Based on recent published and unpublished studies that fail to replicate the effects of stereotype threat, *Ganley et al. (2013) suggested that the literature on the stereotype threat effect in children might suffer from publication bias, a claim that had also been made for the wider stereotype threat literature involving females and mathematics (Stoet & Geary, 2012). Publication bias refers to the practice of primarily publishing articles in which significant results are shown, thus leaving the so-called null results in the file drawer (Ioannidis, 2005; Rosenthal, 1979; Sterling, 1959), a practice that can lead to serious inflations of estimated effect-sizes in meta-analyses (Bakker et al., 2012; Sutton, Duval, Tweedie, Abrams, & Jones, 2000).

According to Ioannidis (2005) a research field is particularly vulnerable to publication bias if the field (1) features studies with small sample sizes; (2) concerns small effect sizes; (3) focuses on a large number of relations; (4) involves studies with a large flexibility in design, definitions, and outcomes; (5) is popular and so features many studies, and (6) deals with topics relevant to financial or political interest. The field of stereotype threat is susceptible to publication bias, because all six characteristics are present to some extent in stereotype threat research. For instance, most studies (39 out of the 47 studies) have a total sample size smaller than 100; the averaged effect sizes found in the recent meta-analyses lie between 0.24 (Picho et al., 2013) and 0.45 (Walton & Spencer, 2009), which are classified as small to medium effect sizes[1] (Cohen, 1992); and the use of multiple dependent variables and covariates is common practice (Stoet & Geary, 2012), despite problems associated with covariate corrections (Wicherts, 2005). Furthermore, the design is often flexible with different kinds of manipulations, control conditions, and moderators. Moreover, the number of published studies attests to the popularity of the topic, and several stereotype threat researchers called for affirmative action based on their research (e.g., by means of a policy paper (Walton, Spencer, & Erman, 2013) or the Brief of Experimental Psychologists et al., 2012, for the case of *Fisher* vs. *the University*). With the former in mind, we expected to find indications of publication bias in our meta-analytic data set.

If we want to draw conclusions based on the outcomes of a meta-analysis, we assume that the outcomes of the included studies are reliable. Unfortunately the outcomes of some studies might be distorted due to questionable research practices (QRPs) in collection of data, reporting of results, and analysis of data. The term QRPs defines a broad set of decisions made by researchers that might positively influence the outcome of their studies. Four examples of frequently used QRPs are (1) failing to report all the dependent variables, (2) collecting extra data when the test statistic is not significant yet, (3) excluding data when it lowers the *p*-value of the test statistic, and (4) rounding down *p*-values (John, Loewenstein, & Prelec, 2012). The practice of using these QRPs with the purpose of obtaining a statistically significant effect is referred to as "*p*-hacking" (Simonsohn et al., 2013). *p*-Hacking can seriously distort the scientific literature because it enlarges the chance of a Type-I error (Simmons, Nelson, & Simonsohn, 2011), and it leads to inflated effect sizes in meta-analyses (Bakker et al., 2012). If many researchers who work within the same field invoke *p*-hacking, then an effect that does not exist at the population level might become established. Simonsohn et al. (2013) have developed the *p*-curve: a tool aimed to distinguish whether a field is infected by selective reporting, or whether results are truthfully reported. When most researchers within a field truthfully reported correct *p*-values, a distribution of statistical significant *p*-values should be right skewed (provided there is an actual effect in the population), whereas the distribution of *p*-values for a field in which researchers *p*-hack will be left skewed. With the *p*-curve, we can test whether it is likely that *p*-values within this field are *p*-hacked.

2. Method

2.1. Search strategies

A literature search was conducted using the databases ABI/INFORM, PsycINFO, ProQuest, Web of Science (searched in March 2013), and ERIC (searched in January 2014). Combined, these five databases cover the majority of the psychological and educational literature. The keywords that we used in the literature search (in conjunction with the phrase "stereotype threat", which needed to be present in the abstract) were "gender," "math," "performance," or "mental rotation," and "children," "girls," "women," or "high school." This search strategy resulted in several search strings that were connected by the search term "AND," such as "ab("stereotype threat") AND children AND gender." In addition two cited-reference searches on Web of Science were conducted; we targeted the oldest paper that we obtained from the first part of our literature search (Ambady et al., 2001) and the classical paper on stereotype threat and gender by Spencer et al. (1999). Additionally, we performed a more informal search on Google Scholar for which we used the same keywords as our other database searches. With this strategy we obtained two extra articles.

An important part of a meta-analysis is the search for unpublished studies or data (i.e., gray literature). We automatically searched parts of the gray literature by our search on Google Scholar and using databases PsycINFO, ERIC, and ProQuest; they do not only contain published papers but also dissertations and conference proceedings. Moreover, in order to find unpublished studies we used three additional strategies. First, we e-mailed the first authors of the included published papers with the question whether they possessed any unpublished data or were familiar with unpublished studies by other researchers. Second, we screened the abstracts of poster

[1] Although widely used, Cohen's rules of thumb for small, medium, and large effects may not be entirely appropriate here. Set against the typical effect sizes for gender differences in mathematics (e.g., $d = 0.16$, Hedges & Nowell, 1995), even a d of 0.1 for the stereotype threat effect among schoolgirls could be substantial in the sense that it may then explain a substantial part of the gender gap, all other things being equal. When considered in light of earlier meta-analyses of the stereotype threat effect the same effect size estimate of $d = 0.1$ could be seen as small. The core issue for understanding the potential effect of publication bias is that stereotype threat effects are small in relation to the sample sizes typical for psychological research (Bakker et al., 2012), leading to underpowered studies.

presentations held at the last 10 conferences of the Society for Personality and Social Psychology (SPSP), selected those abstracts that mentioned stereotype threat and children, and e-mailed the first author that worked on the project in question. Finally, we posted an open call for data on both the SPSP forum (www.spsp.org) and the Social Psychology Network forum (www.socialpsychology.org). We did not receive any papers through the second and third strategies; however, we obtained seven responses through the first strategy, which provided us with five additional studies. Five authors indicated that they had no unpublished works. Ultimately, we included five effect sizes (11%) in the meta-analysis that were a product of unpublished studies. In our literature search, we obtained one Italian study (*Tomasetto, Matteucci, & Pansu, 2010) that was translated by the first author.

2.2. Inclusion criteria

We included study samples based on five criteria. First, we selected only those studies in which schoolgirls were included in the sample and where the gender stereotype threat was manipulated. We excluded studies that focused on only boys or studies that concerned another negatively stereotyped group (e.g., ethnic minorities in other ability domains). Second, because we focused on studies with children and adolescents, we disregarded those studies for which the average age within the sample was above 18. Third, we used experiments in which students were randomly assigned[2] to the stereotype threat condition or control condition. This constraint meant that we included neither correlational studies nor studies that failed to administer a viable stereotype threat. A viable threat was either accomplished using explicit cues that address the ramifications of the gender stereotype (e.g., "Women perform worse on this mathematical test") or using implicit cues that are supposed to activate gender stereotypes (e.g., instructions to circle gender on a test form). Fourth, we included only studies for which the stereotype threat manipulation was treated as a between-subjects factor and thus excluded studies in which this variable was treated as a within-subjects factor. Fifth, the dependent variable had to be the score on a MSSS test. We coded the selected variables using the procedures described in the next section.

2.3. Coding procedures

The selection and coding of the independent and dependent variables was carried out following a number of rules. In some studies participants were assigned not only to a stereotype threat or control condition but also to an additional crossed factor. We treated the groups formed by the additional factor as different populations when this factor was a between-subjects factor.[3] Whenever the additional factor was a within-subjects factor, we took only the level of the factor that, based on the existing theories of stereotype threat, would be expected to have the strongest effect. For instance, we selected a difficult over an easy test in one study (*Neuville & Croizet, 2007). The control condition consisted of either a nullified control condition or a control condition in which no information had been given regarding gender and performance. For studies that involved multiple types of control groups, we selected the control group in the following order: (1) a nullified control condition which described that no differences in performance on the mathematical test have been found, (2) a nullified control condition which described that girls perform better on the mathematical test condition, (3) a nullified control condition in which test-takers were informed that the sex differences in performance on the mathematical test are due to stereotype threat, (4) a nullified control condition that entailed a description or visualization of a stereotype inconsistent situation, and (5) a control condition in which no additional information had been given. In selecting the dependent variable performance on a MSSS test we used the following rules: we first selected a test administered after the threat manipulation over a test administered before the threat manipulation, subsequently we selected published cognitive tests over self-constructed cognitive tests, and finally we selected math tests over other tests (i.e., spatial tests, physics tests, geometrical recall tests, or chemistry tests). We coded performance on a MSSS test via the official scorings rule for the test; if this rule was not reported, we used the reported percentage of correct answers or alternatively the average sum score (i.e., the raw mean number of correct answers per condition).

In addition to the independent and the dependent variable, six other variables were coded. Test difficulty was coded by 1 minus the proportion of correct answers within the control group of girls in the study sample; thus, a more difficult test resulted in a higher score on this moderator variable. We calculated test difficulty using the data from the control group of girls only instead of the entire sample because some (but not all) studies included boys in their samples and the test difficulty needed to be comparable across samples. Additionally, we did not use the data of girls in the experimental group because the effect of stereotype threat would probably distort the actual difficulty. Presence of boys was coded with *yes* when boys were present during test administration or alternatively with *no* when boys were not present. The type of control group was coded with *nullified* whenever the control condition consisted of an active threat removal, whereas a control condition without such an active threat removal was coded as *no information*. Cross-cultural gender equality in the country where the study took place was coded by the country's score on the Gender Gap Index

[2] To correct for random assignment on the cluster level instead of the individual level, we used cluster correction for equal cluster sizes (Hedges, 2007), which was applied to five studies. Both corrected and uncorrected effect sizes are reported in Table 3. We based the adjustment of the effect size on the following formula:

$$d_{T2} = \left(\frac{\overline{Y}^{T}_{\bullet\bullet} - \overline{Y}^{C}_{\bullet\bullet}}{S_T}\right)\sqrt{1-\frac{2(n-1)\rho}{N-2}}.$$

The decision to use an intra-class correlation of $\rho = .2$ was guided by the paper of Hedges and Hedberg (2007), in which calculations of the intra-class correlation for a large sample of schools showed an average of $\rho = .220$. This number was rather stable across grades (kindergarten through the 12th grade); thus, we felt confident to round this number down and use it in our analysis.

[3] In the experiment by *Keller (2007), the factor domain identification was obtained by a median split based on the continuous variable domain identification that we were unable to duplicate. Therefore, we chose to calculate the effect size over the entire sample pooled together, ignoring the variable domain identification.

(Hausmann et al., 2012). The exploratory variable *type of manipulation* was coded by either *explicit* or *implicit* as indicated in Table 1. *Age* was coded by using the mean age in the entire sample; however for papers that only reported an age range we took the midpoint of this range. Test difficulty, age, and cross-cultural gender equality were included as continuous moderators in the analysis, whereas presence of boys, type of control group, and type of manipulation were included as categorical moderators.

Whenever the papers provided insufficient information, we requested additional information from the authors via email. We sent the authors one reminder when they failed to respond. When we failed to obtain all information needed to calculate the effect size, we excluded the paper from that particular analysis. Missing pieces of information on moderator variables were treated as missing values, which were excluded pairwise from the analysis.

To ensure that the coding procedure would be as objective as possible, we developed a coding sheet.[4] The coding process was first carried out by the first author. To assess inter-rater agreement, five variables (type of control condition, presence of boys, cross-cultural gender equality, age, and type of manipulation) were rescored by two independent raters for all studies except for unpublished studies that were not reported in paper form ($k = 43$). The inter-rater agreement was assessed by calculating Fleiss' exact kappa (Conger, 1980; Fleiss, 1971) for categorical variables and the two-way, agreement, unit-measures intraclass correlation (Hallgren, 2012; Shrout & Fleiss, 1979) for continuous variables using the R-package irr (Gamer, Lemon, Fellows, & Singh, 2012). Those measures reached satisfactory levels of agreement for the nominal variables type of control condition (Fleiss' exact $\kappa = .76$) and presence of boys (Fleiss' exact $\kappa = .68$) as well as for continuous variables cross-cultural gender equality (ICC = 1.00) and age (ICC = .96). Only the agreement for the variable type of manipulation was lower (Fleiss' exact $\kappa = .10$), indicating only slight agreement among the three coders. However, as the type of manipulation was used as an exploratory variable in this study and was, therefore, not our main focus; low agreement on this variable is not overly problematic. Disagreements in scoring were solved by selecting the modal response. The dependent variable "performance on a MSSS test" and the moderator variable "test difficulty" were not retrieved by multiple coders because for these variables too much information was not reported in the original articles and needed to be retrieved by e-mailing the authors.

2.4. Statistical methods

We used Hedges's *g* (Hedges, 1981) as effect size estimator, which was calculated by means of the following formula:

$$\text{Hedges}'\text{s } g = \frac{\overline{Y}_{\bullet\bullet}^{experimental} - \overline{Y}_{\bullet\bullet}^{control}}{S_{pooled}} \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right).$$

Thus, study samples with negative effect sizes denote the expected performance decrement due to stereotype threat, whereas positive effect sizes contradict our expectations. The model fitted to the data was the random effects model (for the analyses without moderators) and the mixed effects model (for the analyses with moderators) because we wanted both to explain systematic variance by adding multiple moderators as well as to generalize to the entire population of studies (Viechtbauer, 2010). A characteristic of these two methods is that effect sizes are automatically weighted by the inverse of the study's sampling variance. We have not weighted the effect sizes with regards to other quality indicators. We estimated these models with the R-package metafor (Viechtbauer, 2010) in R version 3.0.2.

When fitting the random effects model, we automatically assume that the population level effect sizes values vary and are normally distributed. In this case, it is considered good practice (Hunter & Schmidt, 2004; Whitener, 1990) to calculate a credibility interval around the average effect size ($\overline{g}$) in addition to the more familiar confidence interval. We calculated the 95% credibility interval, which is an estimation of the boundaries in which 95% of values in the effect size distribution are expected to fall (Hunter & Schmidt, 2004). The boundaries of this interval are obtained using the standard deviation of the distribution of effect sizes (SD_{ES}), or more specifically adding and subtracting 1.96 times the SD_{ES} of $\overline{g}$. In contrast, for the 95% confidence interval the standard error is used to obtain the boundaries around a single value of $\overline{g}$. The confidence interval gives an indication of how the results can fluctuate due to sampling error, whereas the credibility interval gives an indication of the amount of heterogeneity in the distribution of effect sizes.

We estimated the amount of heterogeneity τ^2 with the restricted maximum likelihood estimator, which is the default in metafor (Viechtbauer, 2010) and an approximately unbiased estimator for the standardized mean difference (Viechtbauer, 2005). To address the issue of publication bias, we used several methods. First, we used three methods based on funnel plot asymmetry: the trim and fill method (Duval & Tweedie, 2000; Rothstein, 2007), the rank correlation test (Begg & Mazumdar, 1994), and Egger's test (Sterne & Egger, 2005). A combination of the three methods is desirable to obtain robust results because both the rank correlation test and Egger's test have low power when the amount of studies in the analysis is small (Kepes, Banks, & Oh, 2012). To take tests into account that are not based on the funnel plot, we conducted Ioannidis and Trikalinos's exploratory test (2007), which compares the observed amount of significant studies and the expected amount of significant studies based on power calculations (see also Francis, 2013, 2014). Finally, we created a *p*-curve to have an indication of the practice of *p*-hacking within the field (Simonsohn et al., 2013). A *p*-curve consists of only statistically significant *p*-values within a set of studies. So the *p*-curve analysis includes only the 15 studies for which the mean scores of the experimental group and the control group significantly differed from each other (based on a *t*-test and $\alpha = .05$). If the *p*-curve resembles a right skewed curve, this finding suggests that our set of findings has evidential value, whereas a left skewed curve suggests that some researchers have invoked *p*-hacking (Simonsohn et al., 2013).

[4] A list of excluded studies and the coding sheet are available upon request.

We pre-registered the hypotheses and inclusion criteria of our meta-analysis via the Open Science Framework (https://osf.io/bwupt/).

3. Results

Our literature search and the call for data yielded 972 papers that were further screened. Based on the inclusion criteria, 26 papers (i.e., studies) or unpublished reports were actually included in the meta-analysis, which resulted in 47 independent effect sizes (i.e., study samples). Additional information concerning the screening process is listed in Fig. 1. These 26 papers provided us with a wealth of new information because only 3 of these papers (12%) were also included in the most recent meta-analysis on this topic (Picho et al., 2013). The overlap with the four older meta-analyses is equal to or smaller than 12%. The total sample, obtained by simply adding all participants of the included studies, consisted of $N = 3760$ girls, of which $n_{ST} = 1926$ girls were assigned to the

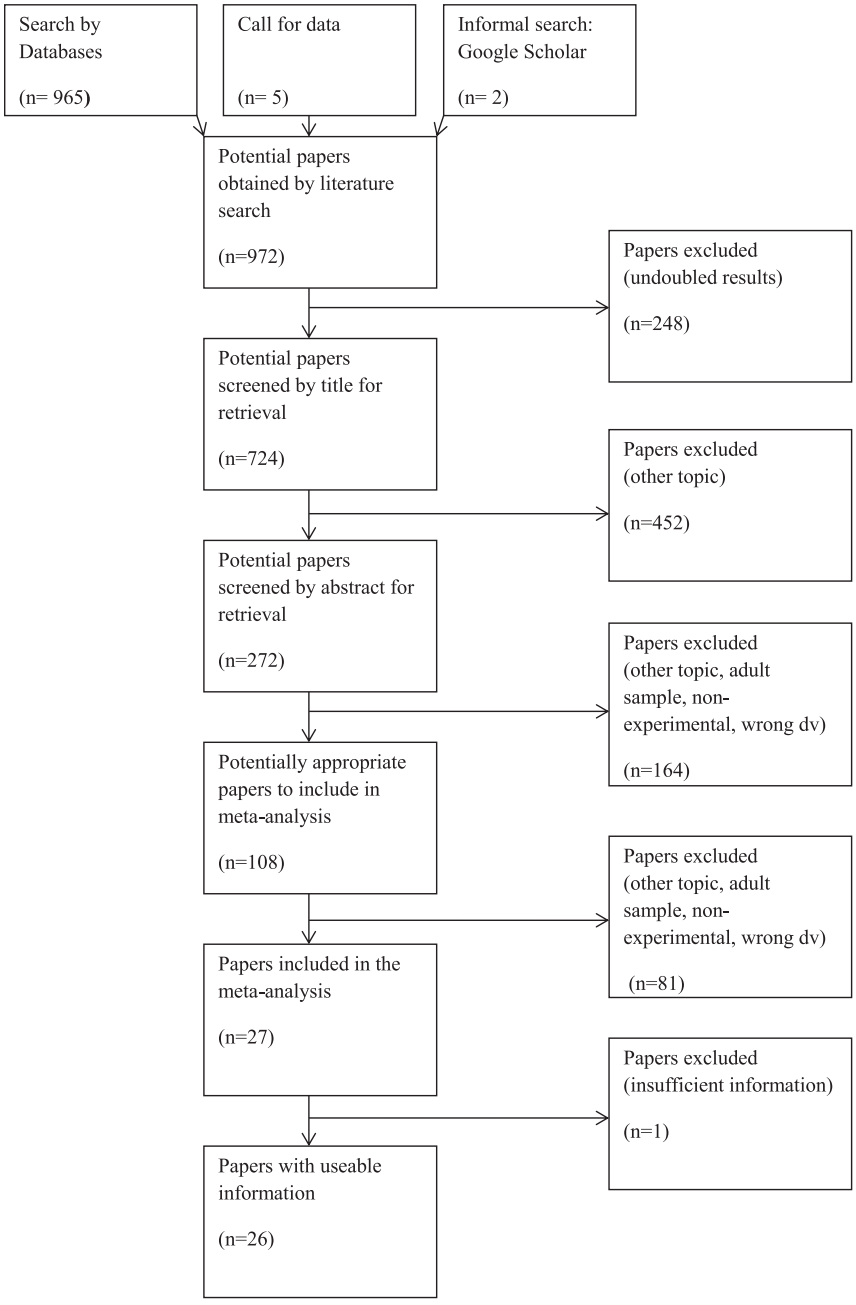

Fig. 1. Flow-chart of the literature search. n = number of papers.

Table 3
Characteristics and statistics of studies included in the meta-analysis.

	Study			Age	Country	Status	*N*	*g*[a]	CC	Boys	Difficulty	GGI	Manipulation
	Authors	Year	No.										
1	Agnoli, Altoè & Muzzatti	–	1A of 1	10.92	Italy	Unpub.	38	0.199	No information	Yes	.636	.673	Implicit
2	Agnoli, Altoè & Muzzatti	–	1B of 1	12.92	Italy	Unpub.	59	0.028	No information	Yes	.668	.673	Implicit
3	*Agnoli, Altoè & Pastro	–	1A of 1	14.01	Italy	Unpub.	41	−0.891	No information	Yes	.594	.673	Implicit
4	*Agnoli, Altoè & Pastro	–	1B of 1	16.03	Italy	Unpub.	49	0.557	No information	Yes	.500	.673	Implicit
5	Bagès & Martinot	2011	1A of 1	10.58	France	Pub.	63	−0.705	Nullified	Yes	.508	.698	Implicit
6	Bagès & Martinot	2011	1B of 1	10.58	France	Pub.	59	−0.864	Nullified	Yes	.552	.698	Implicit
7	Cherney & Campbell	2011	1A of 1	16.02	USA	Pub.	124	0.293	Nullified	No	.500	.737	Explicit
8	Cherney & Campbell	2011	1B of 1	16.02	USA	Pub.	135	0.507	Nullified	Yes	.370	.737	Explicit
9	Cimpian, Mu, & Erickson	2012	2 of 2	5.98	USA	Pub.	48	−0.656	No information	No	.458	.737	Explicit
10	Delgado & Prieto	2008	1 of 1	15.5	Spain	Pub.	168	−0.270 (−0.277)	No information	Yes	.365	.727	Explicit
11	Galdi, Cadinu, & Tomasetto	2013	1	6.47	Italy	Pub.	80	−0.620	Nullified	No	NA	.673	Implicit
12	*Galdi et al.	2014	1 of 3	13.5	USA	Pub.	110	0.137	Nullified	Yes	.620	.737	Explicit
13	*Galdi et al.	2014	2A of 3	12.5	USA	Pub.	115	0.276	No information	Yes	.230	.737	Explicit
14	*Galdi et al.	2014	2B of 3	13.5	USA	Pub.	99	−0.158	No information	Yes	.360	.737	Explicit
15	*Galdi et al.	2014	3A of 3	9.5	USA	Pub.	29	0.165	No information	Yes	.560	.737	Explicit
16	*Galdi et al.	2014	3B of 3	13.5	USA	Pub.	65	0.141	No information	Yes	.550	.737	Explicit
17	*Galdi et al.	2014	3C of 3	17.5	USA	Pub.	76	−0.268	No information	Yes	.480	.737	Explicit
18	Good et al.	2010	1 of 1	14.81	USA	Pub.	34	−0.693	No information	Yes	.782	.737	Implicit
19	Huguet & Régner	2009	1	12	France	Pub.	92	−0.867	No information	Yes	.589	.698	Implicit
20	Huguet & Régner	2007	1 of 2	12	France	Pub.	20	−0.742	No information	No	.538	.698	Implicit
21	Huguet & Régner	2007	2A of 2	12	France	Pub.	136	0.010 (0.010)	No information	No	.598	.698	Implicit
22	Huguet & Régner	2007	2B of 2	12	France	Pub.	87	−0.808 (−0.815)	No information	Yes	.578	.698	Implicit
23	Keller & Dauenheimer	2003	1 of 1	15.7	Germany	Pub.	35	−0.457	Nullified	Yes	.531	.763	Explicit
24	Keller	2007	1 of 1	15.9	Germany	Pub.	55	0.040	Nullified	Yes	.705	.763	Explicit
25	Marchand & Taasoobshirazi	2012	1 of 1	16	USA	Pub.	90	−0.576 (−0.581)	Nullified	Yes	.310	.737	Explicit
26	*Moè	2012	1 of 1	15.5	Italy	Pub.	49	−0.541	Nullified	Yes	.572	.673	Explicit
27	Moè	2009	1A of 1	17.97	Italy	Pub.	24	−0.497	Nullified	Yes	.643	.673	Explicit
28	Moè	2009	1B of 1	17.97	Italy	Pub.	23	−0.620	Nullified	Yes	.554	.673	Explicit
29	Moè & Pazzaglia	2006	1 of 2	17	Italy	Pub.	71	−0.266	Nullified	No	.582	.673	Explicit
30	Muzzatti & Agnoli	2007	1A of 2	7.2	Italy	Pub.	35	0.047	No information	Yes	.509	.673	Implicit
31	Muzzatti & Agnoli	2007	1B of 2	8.4	Italy	Pub.	68	0.230	No information	Yes	.663	.673	Implicit
32	Muzzatti & Agnoli	2007	1C of 2	9.4	Italy	Pub.	64	0.132	No information	Yes	.610	.673	Implicit
33	Muzzatti & Agnoli	2007	1D of 2	10.4	Italy	Pub.	42	−0.424	No information	Yes	.663	.673	Implicit
34	Muzzatti & Agnoli	2007	2A of 2	8.2	Italy	Pub.	42	0.028	No information	Yes	.364	.673	Implicit
35	Muzzatti & Agnoli	2007	2B of 2	10.2	Italy	Pub.	48	0.148	No information	Yes	.305	.673	Implicit
36	Muzzatti & Agnoli	2007	2C of 2	13	Italy	Pub.	30	−1.197	No information	Yes	.325	.673	Implicit
37	Neuburger et al.	2012	1 of 1	10.18	Germany	Pub.	72	−0.143	Nullified	Yes	.741	.763	Explicit
38	Neuville & Croizet	2007	1 of 1	7.3	France	Pub.	45	−0.639	No information	Yes	.200	.698	Implicit
39	Picho & Stephens	2012	1A of 1	15.5	Uganda	Pub.	38	−0.744	No information	Yes	.330	.723	Explicit
40	Picho & Stephens	2012	1B of 1	15.5	Uganda	Pub.	51	−0.135	No information	No	.390	.723	Explicit
41	Stricker & Ward	2004	1 of 2	17.5	USA	Pub.	730	−0.160 (−0.160)	No information	Yes	.522	.737	Implicit
42	Titze et al.	2010	1 of 1	10.47	Germany	Pub.	84	0.273	Nullified	Yes	.272	.763	Explicit
43	Tomasetto et al.	2010	1 of 1	15.59	Italy	Pub.	118	−0.125	Nullified	Yes	.338	.673	Implicit
44	Tomasetto et al.	2011	1A of 1	5.43	Italy	Pub.	33	−0.652	No information	No	NA	.673	Implicit
45	Tomasetto et al.	2011	1B of 1	6.05	Italy	Pub.	64	−0.339	No information	No	NA	.673	Implicit
46	Tomasetto et al.	2011	1C of 1	7.47	Italy	Pub.	27	−0.322	No information	No	NA	.673	Implicit
47	*Twamley	2009	1 of 1	11	USA	Unpub.	74	−0.252	No information	No	.730	.737	Implicit

Note. Status = published versus unpublished papers. $N = N_{\text{threat condition}} + N_{\text{control condition}}$. CC = control condition. Boys = presence of boys (yes) or not (no). GGI = Gender Gap Index. NA indicates a cell with missing data.
[a] The primary number is the corrected effect size; the number in parentheses is the uncorrected effect size.

experimental condition and $n_C = 1834$ girls were assigned to the control condition. The most important characteristics of the included study samples are summarized in Table 3.

3.1. Overall effect

To estimate the overall effect size, we used a random effects model. In accordance with our hypothesis as well as the former literature, we found a small average standardized mean difference, $\overline{g} = -0.22$, $z = -3.63$, $p < .001$, $CI_{95} = -0.34; -0.10$, indicating that girls who have been exposed to a stereotype threat on average score lower on the MSSS tests compared to girls who have not been exposed to such a threat. Furthermore, we found a significant amount of heterogeneity using the restricted maximum likelihood estimator, $\hat{\tau}^2 = 0.10$, $Q(46) = 117.19$, $p < .001$, $CI_{95} = 0.04; 0.19$, which indicates there is variability among the underlying population effect sizes. This estimated heterogeneity accounts for a large share of the total variability, $I^2 = 61.75\%$. The 95% credibility interval, an estimation of the boundaries in which 95% of the true effect sizes are expected to fall, lies between −0.85 and 0.41 (Viechtbauer, 2010). This range constitutes a wide interval. The forest plot (Fig. 2) depicts the effect sizes against the precision with which each effect was estimated.

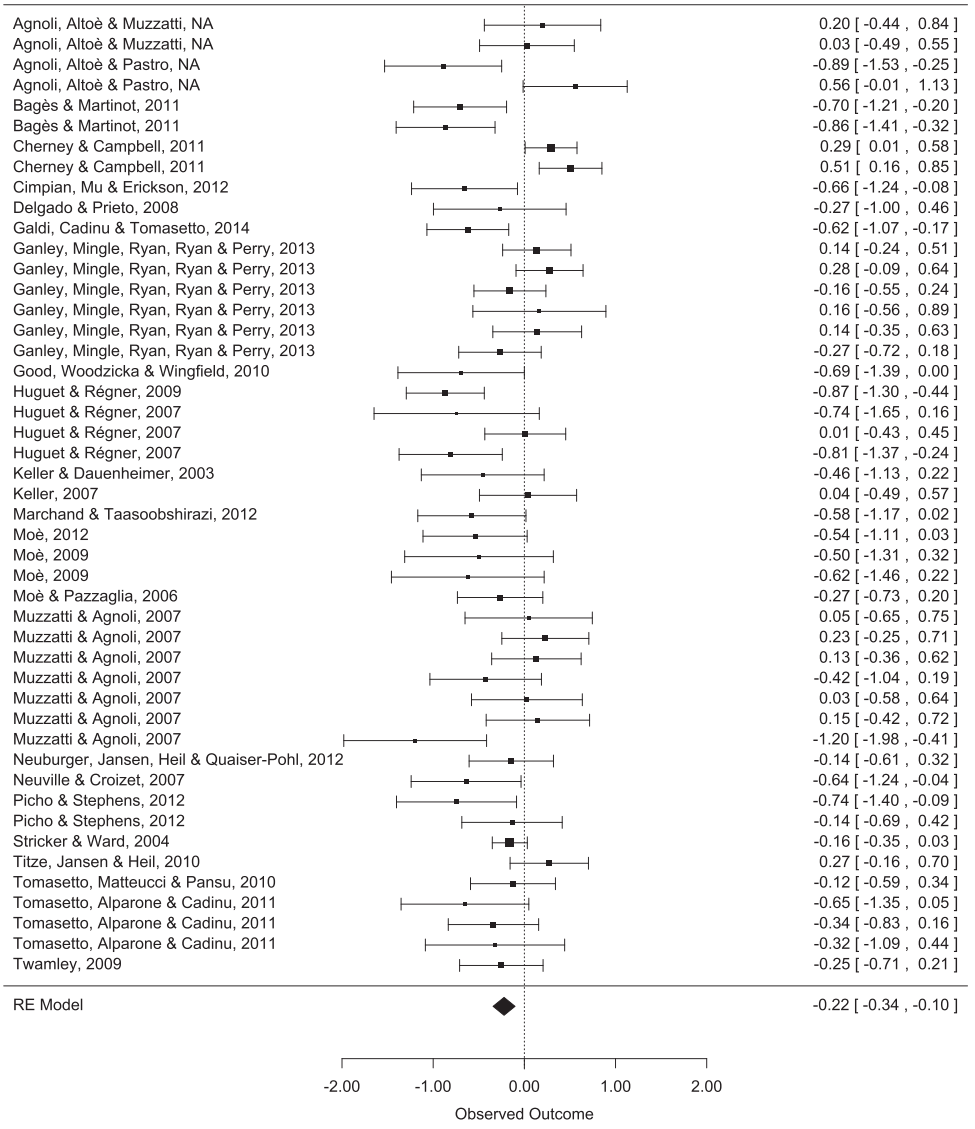

Fig. 2. The forest plot of included effect sizes. NA = missing value. RE model = Random Effects model. The observed outcome is the standardized mean difference Hedges's *g*.

Table 4
Results of the univariate mixed effects meta-regression per moderator.

Variable	*k*	*N*	Intercept	Slope coefficient	*SE*	*z*	*p*	95% CI		Q_E	τ^2	Q_M	I^2	R^2
GGI	47	3760	−2.23	2.83	1.85	1.53	.13	−0.80	6.46	107.33*	0.09	2.34	60%	.07
Boys (factor)	47	3760	−0.28	0.08	0.15	0.54	.59	−0.21	0.36	117.08*	0.10	0.29	62%	0
Difficulty	43	3556	−0.43	0.45	0.42	1.09	.28	−0.37	1.28	105.28*	0.10	1.18	63%	.02
Control (factor)	47	3760	−0.23	0.03	0.13	0.25	.80	−0.22	0.29	115.17*	0.10	0.06	62%	0

* $p < .001$.

Table 5
Results of the multivariate mixed effects meta-regression with four moderators included.

Variable	Slope coefficient	*SE*	*z*	*p*	95% CI	
Intercept	−2.07	1.52	−1.36	.17	−5.06	0.91
GGI	2.30	2.10	1.09	.27	−1.82	6.41
Boys (factor)	−0.05	0.18	−0.27	.79	−0.39	0.30
Difficulty	0.52	0.43	1.20	.23	−0.33	1.37
Control (factor)	−0.03	0.14	0.22	.83	−0.24	0.31

Table 6
Sensitivity analysis: estimating the effect using different amounts of heterogeneity.

$\hat{\tau}^2$	*g*	*SE*	*z*	*p*
0.0447	−0.20	0.05	−4.06	<.001
0.1001	−0.22	0.06	−3.63	<.001
0.1940	−0.24	0.08	−3.10	.002

3.2. Moderator analyses

We submitted the data to separate mixed effects meta-regressions for each of the four moderators and used the REML estimator to obtain the residual $\hat{\tau}^2$ (i.e., unexplained variance in underlying effect sizes). The results of the simple meta-regression analyses for each moderator variable separately are presented in Table 4, where the variables presence of boys and control condition were treated as categorical variables, and the remaining variables were treated as continuous variables. None of the moderators were statistically significant. Additionally, the results for the multiple meta-regression as given in Table 5, showed no statistically significant moderation, $Q_M(4) = 2.68$, $p = .61$, $\hat{\tau}^2 = .11$, $Q_E(38) = 95.59$, $p < .001$. Additional exploratory analyses did not yield any statistically significant explanation for differences between the effect sizes. The moderation of the exploratory variable age, $Q_M(1) = 0.65$, $p = .42$, $\hat{\tau}^2 = .10$, $Q_E(45) = 112.80$, $p < .001$, did not turn out be statistically significant, indicating that we found no evidence for systematic variety in the magnitude of the effect sizes due to differences in age. Additionally the exploratory variable type of manipulation, $Q_M(1) = 3.16$, $p = .08$, $\hat{\tau}^2 = .09$, $Q_E(45) = 103.87$, $p < .001$, did not result in a statistically significant moderation either.

3.3. Sensitivity analyses

To verify the robustness of our results (notably the estimated effect size), we ran several sensitivity analyses, as is recommended for meta-analyses (Greenhouse & Iyengar, 2009). Specifically, we verified the robustness of our results with respect to the use of a different statistical meta-analytic model, an alternative heterogeneity estimator, re-analyses of the random effects model using different estimates of τ^2, diagnostic tests, and different subsets of effect sizes. First, in a fixed effects model, we also found a statistically significant mean effect size of $\bar{g} = -0.16$, $z = -4.35$, $p < .001$.[5] Using the DerSimonian–Laird estimator yielded a similar effect size estimate as the restricted maximum likelihood estimator, $\bar{g} = -0.22$, $z = -3.66$, $p < .001$, $CI_{95} = -0.34; -0.10$, with roughly the same amount of estimated heterogeneity, $\hat{\tau}^2 = 0.10$, $Q(46) = 117.19$, $p < .001$, $CI_{95} = 0.04; 0.19$. We also reran the original analysis with three different amounts for $\hat{\tau}^2$: the originally estimated $\hat{\tau}^2$, the upper bound around $\hat{\tau}^2$, and the lower bound of the confidence interval around $\hat{\tau}^2$. The results of these analyses are summarized in Table 6. Although the estimated effect sizes varied slightly, they all were negative and differed significant from zero.

We also considered potential outliers, by inspecting the studentized residuals, and found that the second study of *Cherney and Campbell (2011) displayed a studentized residual larger than 2. Running the analysis without this study gave an estimated effect size of $\bar{g} = -0.24$, $z = -4.05$, $p < .001$, which indicates that the estimated mean effect size is only slightly influenced by this

[5] Although we report this analysis for the sake of robustness of the estimated effect size, we would not advocate interpreting this result due to the heterogeneity we found among effect sizes.

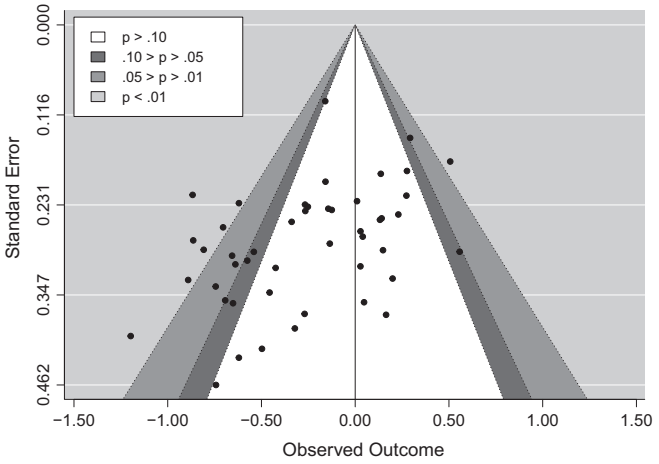

Fig. 3. The contour-enhanced funnel plot of included effect sizes. The observed outcome is the standardized mean difference Hedges's *g*.

study. Finally, we created different subsets to see whether the effect is stable over different categories. We found a few differences between some subsets: the estimated effect size was larger for samples with an implicit stereotype threat manipulation $\bar{g} = -0.32$, $z = -3.76$, $p < .001$, $k = 26$, compared to samples with an explicit stereotype threat manipulation, $\bar{g} = -0.10$, $z = -1.20$, $p = .23$, $k = 21$, and samples gathered outside of the United States of America showed a stronger stereotype threat effect, $\bar{g} = -0.30$, $z = -4.15$, $p < .001$, $k = 34$, than samples gathered in the United States of America, $\bar{g} = -0.05$, $z = -0.48$, $p = .63$, $k = 13$. Additionally we created subsets of young (younger than 13 years) and older (13 years or older) participants; the estimated effect size was larger in samples with younger students, $\bar{g} = -0.25$, $z = -2.92$, $p = .004$, $k = 25$, than in samples with older students, $\bar{g} = -0.20$, $z = -2.19$, $p = .03$, $k = 22$. Using an alternative cut-off at the age of 10 yielded similar results (for younger students, $\bar{g} = -0.24$, $z = -2.06$, $p = .04$, $k = 11$, and for older students, $\bar{g} = -0.22$, $z = -3.07$, $p = .002$, $k = 36$). These subset analyses are exploratory analyses and should be interpreted as such; however, they might be an inspiration for future research.

3.4. Excess of significance results

We used several methods to test for the presence of publication bias. First, we ran several tests on the funnel plot (see Fig. 3) to assess funnel plot asymmetry. According to the estimations of the trim and fill method (Duval & Tweedie, 2000), the funnel plot would be symmetric if 11 effect sizes would have been imputed on the right side of the funnel plot. Actual imputation of those missing effect sizes (Duval & Tweedie, 2000) reduced the estimated effect size to $\bar{g} = -0.07$, $z = -1.10$, $p = .27$, $CI_{95} = -0.21; 0.06$. Because this altered effect size did not differ significantly from zero whereas our original effect size estimation of $\bar{g} = -0.22$ did, this pattern is a first indication that our results might be distorted by publication bias. Both Egger's test (Sterne & Egger, 2005; $z = -3.25$, $p = .001$) and Begg and Mazumdar's (1994) rank correlation test, Kendall's $\tau = -.27$, $p = .01$, indicated funnel plot asymmetry. This finding indicates that imprecise study samples (i.e., study samples with a larger standard error) on average contribute to a more negative effect than precise study samples. The relation between imprecise samples and the effect sizes is illustrated in Fig. 4 using a cumulative meta-analysis sorted by the sampling variance of the samples (Borenstein, Hedges, Higgins, & Rothstein, 2009). This cumulative process first carries out a "meta-analysis" on the sample with the smallest sampling variance and proceeds adding the study with smallest remaining sampling variance and re-analyzing until all samples are included in the meta-analysis. The drifting trend of the estimated effect sizes visualizes the effect that small imprecise study samples have on the estimations of the mean effect. We created subsets to estimate the effects of large study samples ($N \geq 60$) and small study samples ($N < 60$). We found a stronger effect in the subset of smaller study samples, $\bar{g} = -0.34$, $z = -3.76$, $p < .001$, $CI_{95} = -0.52; -0.16$, $CrI_{95} = -0.96; 0.27$, $k = 24$, and a small and nonsignificant effect for the subset of larger study samples, $\bar{g} = -0.13$, $z = -1.63$, $p = .10$, $CI_{95} = -0.29; 0.03$, $CrI_{95} = -0.75; 0.49$, $k = 23$.

Finally, Ioannidis and Trikalinos's exploratory test (Ioannidis & Trikalinos, 2007) showed that this meta-analysis contains more statistically significant effects than would be expected based on the cumulative power of all study samples, $\chi^2(1) = 8.50$, $p = .004$.[6] The excess of statistically significant findings is another indicator of publication bias (Bakker et al., 2012; Francis, 2012). To check the alternative explanation that the excess of statistically significant findings is due to the practice of *p*-hacking we created a *p*-curve (Fig. 5) using the online app from Simonsohn et al. (2013). The *p*-curve depicts the theoretical distribution of *p*-values when there is no effect present (solid line), the theoretical distribution of *p*-values when an effect is present and the tests have 33% power (dotted line), and the observed distribution of the significant *p*-values in our meta-analysis (dashed line). The observed distribution was right-skewed,

[6] To calculate the cumulative power we used the estimated effect size obtained by the random effects model, $|g| = 0.2226$. Although we detect a significant difference between the observed and expected significant study samples based on this effect size, the test is rather sensitive. For an effect size of 0.27, the test is no longer statistically significant.

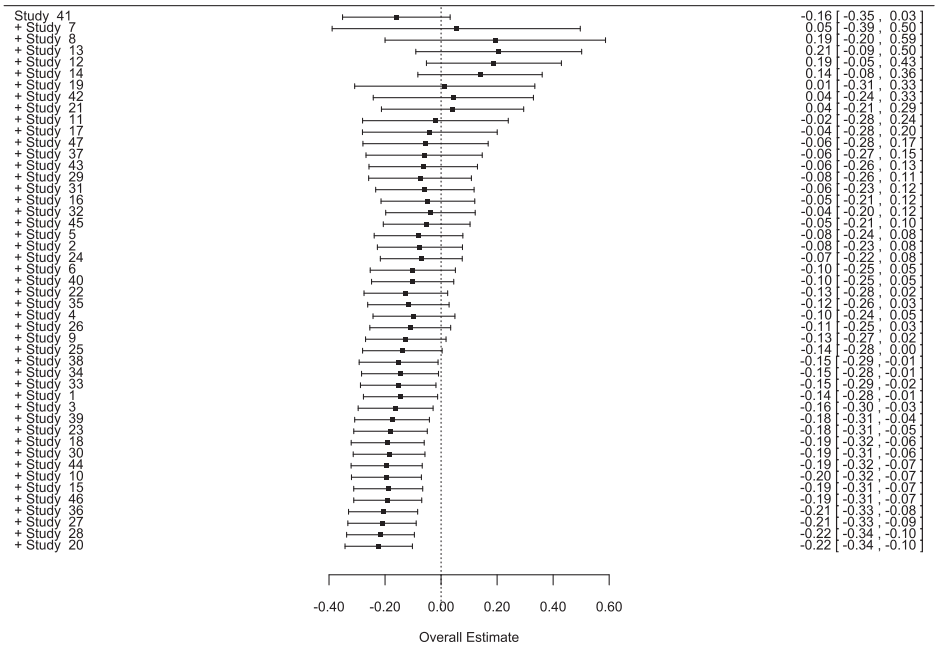

Fig. 4. Cumulative meta-analysis sorted by the sampling variance of the studies. The overall estimate is the estimated average effect size.

$\chi^2(30) = 62.87, p < .001$, which indicated that there is an effect present that is not simply the result of practices like *p*-hacking.[7] Overall, most publication bias tests indicate that the estimated effect size is likely to be inflated.

4. Discussion

Analyzing 15 years of stereotype threat literature with children or adolescents as test-takers, we found indications that girls underperform on MSSS tests due to stereotype threat. Consistent with findings by Nguyen and Ryan (2008), Picho et al. (2013), Walton and Cohen (2003), and Walton and Spencer (2009), we estimated a small effect of −0.22. The estimations of heterogeneity indicated that there was a large share of heterogeneity among population effect sizes. We ran multiple sensitivity analyses, and most of these tests indicated that the mean effect size is rather robust against fluctuations due to alternative decisions regarding the analyses or the removal of influential studies. Yet our results failed to corroborate predictions drawn from stereotype threat theory with regards to the moderating variables. None of the four variables (difficulty, presence of boys, type of control group, and cross-cultural gender equality) significantly moderated the effect of stereotype threat. Exploratory analyses with moderators as age or type of manipulation did not yield significant moderation either. However, we did find some strong indications that publication bias is present in the field of stereotype threat.

In future research, the exploratory variables age and type of manipulation deserve more attention. With regards to the variable age, the effect of stereotype threat overall appears to be rather stable over different ages. However, surprisingly, the subset analyses indicated that the estimated effect size for samples with children younger than 13 was slightly larger than the effect size for samples with older children. An additional subset analysis on our data using only samples with early grade school children (i.e., younger than 8 years old) shows a relatively large estimated mean effect size, $\bar{g} = -0.48, z = -4.30, p < .001, k = 7$. This outcome is rather counterintuitive, because three theories on stereotype threat predict that very young children would not yet be sensitive to detrimental effects of stereotypes: preadolescent children have not obtained a coherent sense of the self yet (Aronson & Good, 2003), young children fail to understand that effort will not necessarily compensate for a lack of mathematical abilities (e.g., Droege & Stipek, 1993; Stipek & Daniels, 1990), and older children endorse gender stereotypes more strongly than younger children (Steffens & Jelenec, 2011). The variable type of manipulation also deserves extra attention. Although type of manipulation did not have a statistically significant effect on stereotype threat ($p = .08$), the intercoder agreement for this variable was suboptimal, and most likely the power for the test of this variable is low. In other words, the circumstances under which we measured this variable were not ideal, and future inspection of it might be valuable. Due to these issues, we conclude that the type of manipulation and age are variables that require more attention in the stereotype threat literature.

Unfortunately the robustness of the stereotype threat effect can be questioned by the presence of publication bias. All three tests based on funnel plot asymmetry—trim and fill (Duval & Tweedie, 2000), Egger's test (Sterne & Egger, 2005), and Begg and Mazumdar's rank correlation test (Begg & Mazumdar, 1994)—indicated that publication bias was present. Additionally Ioannidis and Trikalinos's (2007) exploratory test highlighted an excess of significant findings, which can be due to publication bias. These

[7] The test for the left-skewed distribution is not statistically significant, $\chi^2(30) = 18.24, p = .95$.

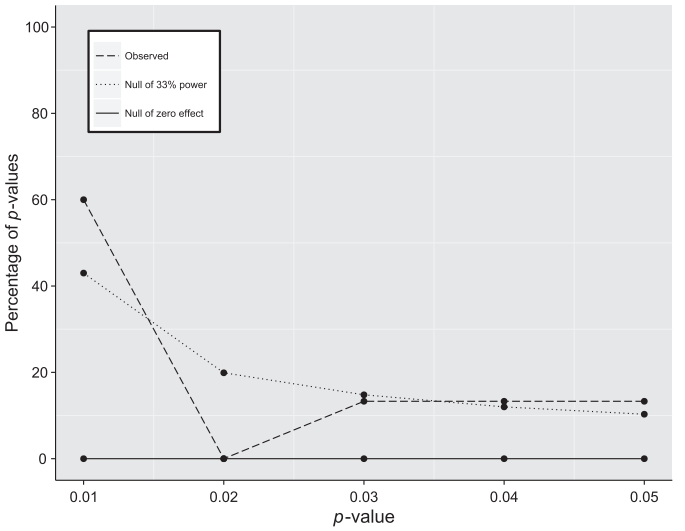

Fig. 5. The *p*-curve of the included studies.

findings might not be entirely reliable when heterogeneity between effect sizes is present (Ioannidis & Trikalinos, 2007). However, this test is deemed appropriate (Francis, 2013) if the included experiments, used methods, and selected populations are similar. Because the methods of the selected studies only vary in details and the population is restricted to schoolgirls, we see no reason to disregard the results of Ioannidis and Trikalinos's exploratory test. Moreover, when we compared the subsets of large study samples and small study samples, only the latter obtained a significant mean effect. That result is striking because smaller studies are associated with lower power to detect an effect and more sampling variability. Taking all aforementioned tests into account, we conclude that the stereotype threat literature among children and adolescents is subject to publication bias.

Currently, we have no good explanation for the large amount of heterogeneity between study samples. None of the four confirmatory moderator variables did explain a significant amount of variance; of those, the variable cross-cultural gender equality closest approached significance, with less equality being predictive of stronger stereotype threat effects, $B = 2.83, z = 1.53, p = .13$. The difference in estimated effects between subsets of study samples conducted inside and outside the United States also indicates that there are cross-cultural differences of the estimated effects. This corresponds to the large cross-cultural gender gap differences in mathematical performance (Else-Quest et al., 2010; Mullis et al., 2012; OECD, 2010). A post hoc power analysis (Hedges & Pigott, 2004) for the omnibus test of this simple meta-regression denotes that with a power of .44 the test for the moderator has quite low power.[8] A lack of power thus might be an alternative explanation for the nonsignificant effects of this moderator. Unfortunately, power is difficult to enforce when performing a meta-analysis, but it might be interesting for researchers who are planning future stereotype threat meta-analyses using adult samples to consider cross-cultural gender equality as moderator variable because the studies in the adult population are more numerous and will lead to more powerful meta-analyses.

A different explanation for the heterogeneity in our analysis is the presence of moderator variables that we did not take into account. Domain identification for instance appears to be an important moderator for the stereotype threat effect, which has been found in adult samples (Cadinu et al., 2003; Lesko & Corpus, 2006; Pronin et al., 2004; Steinberg et al., 2012) as well as in samples of children (*Keller, 2007). The difficulty with this moderator variable is that few studies report the degree to which students identify themselves with mathematics or the like, which makes it problematic to take the variable into account.[9] In addition, publication bias could have played a role in our failure to find moderation of the stereotype threat effect. Specifically, because the effects of publication bias are directly proportional to the size of the underlying effect (cf. Bakker et al., 2012), publication bias may obscure actual differences between these underlying effect sizes.

4.1. Limitations

The amount of unexplained heterogeneity is the first of a few limitations concerning our meta-analysis. Due to this heterogeneity it is difficult to substantively interpret the stereotype threat effect and the degree to which publication bias is a serious issue. Also, publication bias itself can have an effect on heterogeneity of effects (Jackson, 2006). However, with the multiple sensitivity analyses and different signs of publication bias, we are rather confident that publication bias is at play in this literature. Another limitation of this study is the low power for the tests of the moderators. This limitation is mainly due to the small sample sizes within the studies (only seven studies had an $N > 100$ required to detect with sufficient power a d of .50) and the limited amount of studies included in the

[8] We calculated the power using the method of Hedges and Pigott (2004) for the mixed-effects omnibus test, with $\beta = 3$ and $\hat{\tau}^2 = 0.6$.

[9] Studies seldom indicated whether participants were specifically selected on this moderator variable. Moreover, identification with the domain and gender roles are both variables that consist of individual differences that can best be modeled at the individual level for which the raw data are needed.

meta-analysis. Although it is unfortunate that the tests for the moderators are underpowered, it does not affect the conclusion that publication bias is a serious issue within this line of research. Finally, it would have been informative if the dataset contained more unpublished studies, especially because a subset analysis with a fair amount of unpublished studies could have been a good estimator for the effect of stereotype threat that is not influenced by publication bias.[10] Unfortunately an extensive gray literature search did not yield more than five effect sizes, which corresponds to a percentage of 11% unpublished effect sizes in our meta-analysis. However, such a low percentage of gray literature papers within psychological meta-analyses seems rather common even in top journals (Ferguson & Brannick, 2012). The most important difficulties we encountered with the gray literature search is that the amount of details in documents like conference abstracts or even doctoral dissertations was insufficient to successfully include the study in the meta-analysis and authors were often unreachable. We want to stress that pre-registration of studies including contact information of the first author is of vital importance for more reliable future meta-analyses.

4.2. Conclusion

To conclude, we estimated a small average effect of stereotype threat on the MSSS test-performance of school-aged girls; however, the studies show large variation in outcomes, and it is likely that the effect is inflated due to publication bias. This finding leads us to conclude that we should be cautious when interpreting the effects of stereotype threat on children and adolescents in the STEM realm. To be more explicit, based on the small average effect size in our meta-analysis, which is most likely inflated due to publication bias, we would not feel confident to proclaim that stereotype threat manipulations will harm mathematical performance of girls in a systematic way or lead women to stay clear from occupations in the STEM domain. Of course, we do not challenge the fact that stereotypes might strongly influence a person's life under unfortunate circumstances; however, we want to avoid the unjustifiable generalization that stereotype threat, based on the evidence at hand (i.e., the average small effect that stereotype threat manipulations have on instant test performance within this meta-analysis), generally leads to lower math grades and women leaving the STEM field. Due to the scientific and societal importance of the topic, we urge that future research is needed to disentangle the effects of stereotype threat from publication bias. As directions for future research we propose simple, large replication studies, preferably administered cross-culturally. In our opinion, only studies with large sample sizes will contribute to acquiring an accurate picture of the actual effect of stereotype threat among schoolgirls. A power calculation for a one-tailed *t*-test indicated that, with an effect size of 0.223, roughly 250 participants are needed per condition to achieve a power of .80. In addition, these studies should be appropriately registered (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) via the Open Science Framework or the What Works Clearinghouse to avoid publication bias and related biases introduced during the analyses of the data.

References[11]

Agnoli, F., Altoè, G., & Muzzatti, B. (n.d.). Unpublished data: Università degli Studi di Padova.

* Agnoli, F., Altoè, G., & Pastro, M. (n.d.). Unpublished data: Università degli Studi di Padova.

Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., & Mitchell, J. P. (2004). Deflecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology*, *40*, 401–408. http://dx.doi.org/10.1016/j.jesp.2003.08.003.

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, *12*, 385–390. http://dx.doi.org/10.1111/1467-9280.00371.

Aronson, J., & Good, C. (2003). The development and consequences of stereotype vulnerability in adolescents. In F. Pajares, & T. Urdan (Eds.), *Adolescence and education. Academic motivation of adolescents*, *2*. (pp. 299–330). Greenwich, CT: Information Age Publishing.

*Bagès, C., & Martinot, D. (2011). What is the best model for girls and boys faced with a standardized mathematics evaluation situation: A hardworking role model or a gifted role model? *British Journal of Social Psychology*, *50*, 536–543. http://dx.doi.org/10.1111/j.2044-8309.2010.02017.x.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. http://dx.doi.org/10.1177/1745691612459060.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlationtest for publication bias. *Biometrics*, *50*, 1088–1101.

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General*, *136*, 256–276. http://dx.doi.org/10.1037/0096-3445.136.2.256.

Ben-zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, *41*, 174–181. http://dx.doi.org/10.1016/j.jesp.2003.11.007.

Bonnot, V., & Croizet, J. C. (2007). Stereotype internalization and women's math performance: The role of interference in working memory. *Journal of Experimental Social Psychology*, *43*, 857–866. http://dx.doi.org/10.1016/j.jesp.2006.10.006.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Cumulative meta-analysis. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to meta-analysis* (pp. 371–376). Chichester: John Wiley & Sons Ltd.

Brief of Experimental Psychologists et al. as Amici Curiae Supporting Respondents (2012, August 13). *Fisher v. University of Texas.* (No. 01-1015).

Brodish, A. B., & Devine, P. G. (2009). The role of performance–avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal of Experimental Social Psychology*, *45*, 180–185. http://dx.doi.org/10.1016/j.jesp.2008.08.005.

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.

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*, 676–713 (Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10560326).

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology*, *33*, 267–285. http://dx.doi.org/10.1002/ejsp.145.

Campbell, S. M., & Collaer, M. L. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychology of Women Quarterly*, *33*, 437–444. http://dx.doi.org/10.1111/j.1471-6402.2009.01521.x.

*Cherney, I. D., & Campbell, K. L. (2011). A league of their own: Do single-sex schools increase girls' participation in the physical sciences? *Sex Roles*, *65*, 712–724. http://dx.doi.org/10.1007/s11199-011-0013-6.

[10] The estimated effect of the unpublished subset was $\bar{g} = -0.07$ ($z = -0.29$, $p = .77$), however this effect is based only on $k = 5$ effect sizes.

[11] References with asterisk were included in meta-analysis.

Cheryan, S., & Plaut, V. C. (2010). Explaining underrepresentation: A theory of precluded interest. *Sex Roles*, *63*(7–8), 475–488. http://dx.doi.org/10.1007/s11199-010-9835-x.
*Cimpian, A., Mu, Y., & Erickson, L. C. (2012). Who is good at this game? Linking an activity to a social category undermines children's achievement. *Psychological Science*, *23*, 533–541. http://dx.doi.org/10.1177/0956797611429803.
Cohen, J. (1992). A power primer. *Quantitative Methods In Psychology*, *112*, 155–159.
Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*, 322–328.
Cota, A. A., & Dion, K. L. (1986). Salience of gender and sex composition of ad hoc groups: An experimental test of distinctiveness theory. *Journal of Personality and Social Psychology*, *50*, 770–776. http://dx.doi.org/10.1037//0022-3514.50.4.770.
Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*, 766–779. http://dx.doi.org/10.1111/j.1467-8624.2010.01529.x.
*Delgado, A. R., & Prieto, G. (2008). Stereotype threat as validity threat: The anxiety–sex–threat interaction. *Intelligence*, *36*, 635–640. http://dx.doi.org/10.1016/j.intell.2008.01.008.
Droege, K. L., & Stipek, D. J. (1993). Children's use of dispositions to predict classmates' behavior. *Developmental Psychology*, *29*, 646–654. http://dx.doi.org/10.1037//0012-1649.29.4.646.
Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method. *Biometrics*, *56*, 455–463.
Dweck, C. S. (2002). The development of ability conceptions. In A. Wigfield, & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 57–88). San Diego, CA: Academic Press.
Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, *64*, 830–847. http://dx.doi.org/10.2307/1131221.
Eccles, J. S., Wigfield, A., Constance, A., Miller, C., Reuman, D. A., & Yee, D. (1989). Self-esteem: Relations and changes at early adolescence. *Journal of Personality*, *57*, 283–310.
Elizaga, R. A., & Markman, K. D. (2008). Peers and performance: How in-group and out-group comparisons moderate stereotype threat effects. *Current Psychology*, *27*, 290–300. http://dx.doi.org/10.1007/s12144-008-9041-y.
Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103–127. http://dx.doi.org/10.1037/a0018053.
Eriksson, K., & Lindholm, T. (2007). Making gender matter: The role of gender-based expectancies and gender identification on women's and men's math performance in Sweden. *Scandinavian Journal of Psychology*, *48*, 329–338. http://dx.doi.org/10.1111/j.1467-9450.2007.00588.x.
Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition & Emotion*, *6*, 409–434. http://dx.doi.org/10.1080/02699939208409696.
Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics*, *21*, 55–69. http://dx.doi.org/10.1007/BF00311015.
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*, 120–128. http://dx.doi.org/10.1037/a0024445.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382. http://dx.doi.org/10.1037/h0031619.
Ford, T. E., Ferguson, M. A., Brooks, J. L., & Hagadone, K. M. (2004). Coping sense of humor reduces effects of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin*, *30*, 643–653. http://dx.doi.org/10.1177/0146167203262851.
Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*, 585–594. http://dx.doi.org/10.1177/1745691612459520.
Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153–169. http://dx.doi.org/10.1016/j.jmp.2013.02.003.
Francis, G. (2014). The frequency of excess success for articles in psychological science. *Psychonomic Bulletin & Review*, *1–8*.
*Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child Development*, *85*, 250–263.
Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement. http://cran.r-project.org/package=irr (Retrieved from)
*Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology*, *49*, 1886–1897. http://dx.doi.org/10.1037/a0031412.
Gerstenberg, F. X. R., Imhoff, R., & Schmitt, M. (2012). "Women are bad at math, but I'm not, am I?" Fragile mathematical self-concept predicts vulnerability to a stereotype threat effect on mathematical performance. *European Journal of Personality*, *26*, 588–599. http://dx.doi.org/10.1002/per.
Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economic*, *118*, 1049–1074.
*Good, J. J., Woodzicka, J. A., & Wingfield, L. C. (2010). The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology*, *150*, 132–147.
Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnositics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 417–434). New York, NY: Russell Sage Foundation.
Guiso, L., Monte, F., & Sapienza, P. (2008). Culture, gender and math. *Science*, *320*, 1164–1165.
Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorial in Quantitative Methods for Psychology*, *8*, 23–34.
Hartley, B. L., & Sutton, R. M. (2013). A stereotype threat account of boys' academic underachievement. *Child Development*, *84*, 1716–1733. http://dx.doi.org/10.1111/cdev.12079.
Hausmann, R., Tyson, L. D., & Zahidi, S. (2012). *The global gender gap report.* , 1–371 (Retrieved from http://www3.weforum.org/docs/GGGR12/MainChapter_GGGR12.pdf.).
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimations. *Journal of Educational Statistics*, *6*, 107–128.
Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370. http://dx.doi.org/10.3102/1076998606298043.
Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. http://dx.doi.org/10.3102/0162373707299706.
Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability and numbers of high-scoring individuals. *Science*, *269*, 41–45.
Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*, 426–445. http://dx.doi.org/10.1037/1082-989X.9.4.426.
Heyman, G. D., Dweck, C. S., & Cain, K. M. (1992). Young children's vulnerability to self-blame and helplessness: Relationship to beliefs about goodness. *Child Development*, *63*, 401–415.
*Huguet, P., & Régner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology*, *99*, 545–560. http://dx.doi.org/10.1037/0022-0663.99.3.545.
*Huguet, P., & Régner, I. (2009). Counter-stereotypic beliefs in math do not protect school girls from stereotype threat. *Journal of Experimental Social Psychology*, *45*, 1024–1027. http://dx.doi.org/10.1016/j.jesp.2009.04.029.
Hunter, J. E., & Schmidt, F. L. (2004). Technical questions in meta-analysis of correlations. *Methods of meta-analysis. Correcting error and bias in research findings* (pp. 189–240). Thousand Oaks, CA: Sage Publications.
Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. A meta-analysis. *Psychology of Women Quarterly*, *14*, 299–324. http://dx.doi.org/10.1111/j.1471-6402.1990.tb00022.x.
Inzlicht, M., Aronson, J., Good, C., & McKay, L. (2006). A particular resiliency to threatening environments. *Journal of Experimental Social Psychology*, *42*, 323–336. http://dx.doi.org/10.1016/j.jesp.2005.05.005.
Inzlicht, M., & Ben-zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*, 365–371.

Inzlicht, M., & Ben-Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology*, *95*, 796–805. http://dx.doi.org/10.1037/0022-0663.95.4.796.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701. http://dx.doi.org/10.1371/journal.pmed.0020124.
Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. http://dx.doi.org/10.1177/1740774507079441.
Jackson, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine*, *25*, 2911–2921. http://dx.doi.org/10.1002/sim.2293.
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. http://dx.doi.org/10.1177/0956797611430953.
Jordan, A. H., & Lovett, B. J. (2007). Stereotype threat and test performance: A primer for school psychologists. *Journal of School Psychology*, *45*, 45–59. http://dx.doi.org/10.1016/j.jsp.2006.09.003.
Katz, P. A., & Kofkin, J. A. (1997). Race, gender, and young children. In S. S. Luthar, J. A. Burack, D. Cicchetti, & J. R. Weisz (Eds.), *Developmental psychopathology: Perspectives on adjustment, risk, and disorder* (pp. 51–74). Cambridge, UK: Cambridge University Press.
*Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology*, *77*, 323–338. http://dx.doi.org/10.1348/000709906XI.
*Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*, *29*, 371–381. http://dx.doi.org/10.1177/0146167202250218.
Kepes, S., Banks, G. C., & Oh, I. S. (2012). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology*, *1–21*. http://dx.doi.org/10.1007/s10869-012-9279-0.
Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: A prospective study of female college students. *Psychological Science*, *18*, 13–18. http://dx.doi.org/10.1111/j.1467-9280.2007.01841.x.
Kinlaw, R. C., & Kurtz-Costes, B. (2003). The development of children's beliefs about intelligence. *Developmental Review*, *23*, 125–161. http://dx.doi.org/10.1016/S0273-2297(03)00010-8.
Kurtz-Costes, B., Rowley, S. J., Harris-Britt, A., & Woods, T. A. (2008). Gender stereotypes about mathematics and science and self-perceptions of ability in late childhood and early adolescence. *Merrill-Palmer Quarterly*, *54*, 386–409. http://dx.doi.org/10.1353/mpq.0.0001.
Leedy, M. G., LaLonde, D., & Runk, K. (2003). Gender equity in mathematics: Beliefs of students, parents, and teachers. *School Science and Mathematics*, *103*, 285–292.
Lesko, A. C., & Corpus, J. H. (2006). Discounting the difficult: How high math-identified women respond to stereotype threat. *Sex Roles*, *54*, 113–125. http://dx.doi.org/10.1007/s11199-005-8873-2.
Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research*, *41*, 63–76. http://dx.doi.org/10.1080/0013188990410106.
Lord, C. G., & Saenz, D. S. (1985). Memory deficits and memory surfeits: Differential cognitive consequences of tokenism for tokens and observers. *Journal of Personality and Social Psychology*, *49*, 918–926.
Lummis, M., & Stevenson, H. W. (1990). Gender differences in beliefs and achievement: A cross-cultural study. *Developmental Psychology*, *26*, 254–263. http://dx.doi.org/10.1037//0012-1649.26.2.254.
*Marchand, G. C., & Taasoobshirazi, G. (2012). Stereotype threat and women's performance in physics. *International Journal of Science Education*, 1–12. http://dx.doi.org/10.1080/09500693.2012.683461.
Markus, H. (1978). The effect of mere presence on social facilitation: An unobtrusive test. *Journal of Experimental Social Psychology*, *14*, 389–397.
Martin, C. L., & Little, J. K. (1990). The relation of gender understanding to children's sex-typed preferences and gender stereotypes. *Child Development*, *61*, 1427–1439.
Martinot, D., Bagès, C., & Désert, M. (2012). French children's awareness of gender stereotypes about mathematics and reading: When girls improve their reputation in math. *Sex Roles*, *66*, 210–219. http://dx.doi.org/10.1007/s11199-011-0032-3.
Martinot, D., & Désert, M. (2007). Awareness of a gender stereotype, personal beliefs and self-perceptions regarding math ability: When boys do not surpass girls. *Social Psychology of Education*, *10*, 455–471. http://dx.doi.org/10.1007/s11218-007-9028-9.
Marx, D. M., & Ko, S. J. (2012). Superstars "like" me: The effect of role model similarity on performance under threat. *European Journal of Social Psychology*, *42*, 807–812. http://dx.doi.org/10.1002/ejsp.1907.
Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin*, *28*, 1183–1193. http://dx.doi.org/10.1177/01461672022812004.
Mcguire, W. J., Mcguire, C. V., & Winton, W. (1979). Effects of household sex composition on the salience of one's gender in the spontaneous self-concept. *Journal of Experimental Social Psychology*, *15*, 77–90.
McIntyre, R. B., Paulson, R. M., Taylor, C. A., Morin, A. L., & Lord, C. G. (2011). Effects of role model deservingness on overcoming performance deficits induced by stereotype threat. *European Journal of Social Psychology*, *41*, 301–311. http://dx.doi.org/10.1002/ejsp.774.
*Moè, A. (2009). Are males always better than females in mental rotation? Exploring a gender belief explanation. *Learning and Individual Differences*, *19*, 21–27. http://dx.doi.org/10.1016/j.lindif.2008.02.002.
*Moè, A. (2012). Gender difference does not mean genetic difference: Externalizing improves performance in mental rotation. *Learning and Individual Differences*, *22*, 20–24. http://dx.doi.org/10.1016/j.lindif.2011.11.001.
*Moè, A., & Pazzaglia, F. (2006). Following the instructions! Effects of gender beliefs in mental rotation. *Learning and Individual Differences*, *16*, 369–377. http://dx.doi.org/10.1016/j.lindif.2007.01.002.
Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*, 879–885. http://dx.doi.org/10.1111/j.1467-9280.2007.01995.x.
*Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children. *Developmental Psychology*, *43*, 747–759. http://dx.doi.org/10.1037/0012-1649.43.3.747.
*Neuburger, S., Jansen, P., Heil, M., & Quaiser-Pohl, C. (2012). A threat in the classroom. Gender stereotype activation and mental-rotation performance in elementary-school children. *Zeitschrift für Psychologie*, *220*, 61–69. http://dx.doi.org/10.1027/2151-2604/a000097.
*Neuville, E., & Croizet, J. C. (2007). Can salience of gender identity impair math performance among 7–8 years old girls? The moderating role of task difficulty. *European Journal of Psychology of Education*, *12*, 307–316.
Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *The Journal of Applied Psychology*, *93*, 1314–1334. http://dx.doi.org/10.1037/a0012702.
Nicholls, J. G. (1979). Development of perception of own attainment and causal attributions for success and failure in reading. *Journal of Educational Psychology*, *71*, 94–99 (Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/438417).
Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., et al. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 10593–10597. http://dx.doi.org/10.1073/pnas.0809921106.
O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, *29*, 782–789. http://dx.doi.org/10.1177/0146167203252810.
OECD (2010). *PISA 2009 results: What students know and can do — Student performance in reading, mathematics and science Vol. I*. http://dx.doi.org/10.1787/9789264091450-en.
Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology*, *26*, 291–310. http://dx.doi.org/10.1006/ceps.2000.1052.
Osborne, J. W. (2007). Linking stereotype threat and anxiety. *Educational Psychology*, *27*, 135–154. http://dx.doi.org/10.1080/01443410601069929.
Oswald, D. L., & Harvey, R. D. (2001). *Hostile environments, stereotype threat, and math performance among undergraduate women19. (pp. 338–356), 338–356.*
Picho, K., Rodriguez, A., & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: A meta-analysis. *The Journal of Social Psychology*, *153*, 299–333.

*Picho, K., & Stephens, J. M. (2012). Culture, context and stereotype threat: A comparative analysis of young Ugandan women in coed and single-sex schools. *The Journal of Educational Research*, *105*, 52–63. http://dx.doi.org/10.1080/00220671.2010.517576.
Pronin, E., Steele, C. M., & Ross, L. (2004). Identity bifurcation in response to stereotype threat: Women and mathematics. *Journal of Experimental Social Psychology*, *40*, 152–168. http://dx.doi.org/10.1016/S0022-1031(03)00088-X.
Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*, 268–302. http://dx.doi.org/10.3102/0002831210372249.
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. http://dx.doi.org/10.1037//0033-2909.86.3.638.
Rothstein, H. R. (2007). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology*, *4*, 61–81. http://dx.doi.org/10.1007/s11292-007-9046-9.
Rydell, R. J., McConnell, A. R., & Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology*, *96*, 949–966. http://dx.doi.org/10.1037/a0014846.
Rydell, R. J., Rydell, M. T., & Boucher, K. L. (2010). The effect of negative performance stereotypes on learning. *Journal of Personality and Social Psychology*, *99*, 883–896. http://dx.doi.org/10.1037/a0021139.
Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology*, *38*, 194–201.
Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*, 440–452. http://dx.doi.org/10.1037/0022-3514.85.3.440.
Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles*, *50*, 835–850.
Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*, *39*, 68–74.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
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. http://dx.doi.org/10.1177/0956797611417632.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, 1–40.
Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles*, *47*, 179–191.
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. http://dx.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*, 613–629.
Steele, C. M. (2010). *Whistling Vivaldi and other clues to how stereotypes affect us*. New York, NY: W.W. Norton & Company.
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 (Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7473032).
Steffens, M. C., & Jelenec, P. (2011). Separating implicit gender stereotypes regarding math and language: Implicit ability stereotypes are self-serving for boys and men, but not for girls and women. *Sex Roles*, *64*, 324–335. http://dx.doi.org/10.1007/s11199-010-9924-x.
Steinberg, J. R., Okun, M. A., & Aiken, L. S. (2012). Calculus GPA and math identification as moderators of stereotype threat in highly persistent women. *Basic and Applied Social Psychology*, *34*, 534–543. http://dx.doi.org/10.1080/01973533.2012.727319.
Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance — Or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.
Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). New York, NY: Wiley.
Stipek, D. J., & Daniels, D. H. (1990). Children's use of dispositional attributions in predicting the performance and behavior of classmates. *Journal of Applied Developmental Psychology*, *11*, 13–28. http://dx.doi.org/10.1016/0193-3973(90)90029-J.
Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*, 93–102. http://dx.doi.org/10.1037/a0026617.
*Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, *34*, 665–693.
Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *BMJ*, *320*, 1574–1577.
Taylor, C. A., Lord, C. G., McIntyre, R. B., & Paulson, R. M. (2011). The Hillary Clinton effect: When the same role model inspires or fails to inspire improved performance under stereotype threat. *Group Processes & Intergroup Relations*, *14*, 447–459. http://dx.doi.org/10.1177/1368430210382680.
Tiedemann, J. (2000). Gender-related beliefd of teachers in elementary school mathematics. *Educational Studies in Mathematics*, *41*, 191–207.
*Titze, C., Jansen, P., & Heil, M. (2010). Mental rotation performance in fourth graders: No effects of gender beliefs (yet?). *Learning and Individual Differences*, *20*, 459–463. http://dx.doi.org/10.1016/j.lindif.2010.04.003.
*Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls' math performance under stereotype threat: the moderating role of mothers' gender stereotypes. *Developmental Psychology*, *47*, 943–949. http://dx.doi.org/10.1037/a0024047.
*Tomasetto, C., Matteucci, M. C., & Pansu, P. (2010). Genere e matematica: si puo ridurre la minaccia dello stereotipo in classe? In R. Ghigi (Ed.), *Adolescenti in genere. Stili di vita e atteggiamenti dei giovani in Emilia Romagna* (pp. 99–104). Roma: Carocci.
*Twamley, E. E. (2009). *The role of gender identity on the effects of stereotype threat: An examination of girls' math performance in a single-sex classroom.*
Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*, 261–293. http://dx.doi.org/10.3102/10769986030003261.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.
Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. http://dx.doi.org/10.1177/1745691612463078.
Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology*, *39*, 456–467. http://dx.doi.org/10.1016/S0022-1031(03)00019-2.
Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*, 1132–1139. http://dx.doi.org/10.1111/j.1467-9280.2009.02417.x.
Walton, G. M., Spencer, S. J., & Erman, S. (2013). Affirmative meritocracy. *Social Issues and Policy Review*, *7*, 1–35. http://dx.doi.org/10.1111/j.1751-2409.2012.01041.x.
Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, *75*, 315–321. http://dx.doi.org/10.1037//0021-9010.75.3.315.
Wicherts, J. M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *The American Psychologist*, *60*, 267–269. http://dx.doi.org/10.1037/0003-066X.60.3.267.
Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*, 696–716. http://dx.doi.org/10.1037/0022-3514.89.5.696.
Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., et al. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology*, *89*, 451–469. http://dx.doi.org/10.1037/0022-0663.89.3.451.
Wout, D., Danso, H., Jackson, J., & Spencer, S. (2008). The many faces of stereotype threat: Group- and self-threat. *Journal of Experimental Social Psychology*, *44*, 792–799. http://dx.doi.org/10.1016/j.jesp.2007.07.005.
Zajonc, R. B. (1965). Social facilitation. *Science*, *149*, 269–274.