

This is to certify that the
dissertation entitled

DOES STEREOTYPE THREAT DIFFERENTIALLY AFFECT
COGNITIVE ABILITY TEST PERFORMANCE OF MINORITIES AND
WOMEN? A META-ANALYTIC REVIEW OF EXPERIMENTAL
EVIDENCE

presented by

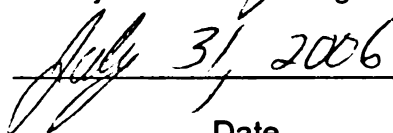
HANNAH-HANH DUNG NGUYEN

has been accepted towards fulfillment
of the requirements for the

PH.D. degree in INDUSTRIAL PSYCHOLOGY



Major Professor's Signature



Date

MSU is an Affirmative Action/Equal Opportunity Institution



PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

**DOES STEREOTYPE THREAT DIFFERENTIALLY AFFECT COGNITIVE ABILITY
TEST PERFORMANCE OF MINORITIES AND WOMEN? A META-ANALYTIC
REVIEW OF EXPERIMENTAL EVIDENCE**

By

Hannah-Hanh Dung Nguyen

A DISSERTATION

**Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of**

DOCTOR OF PHILOSOPHY

Department of Psychology

2006

ABSTRACT

DOES STEREOTYPE THREAT DIFFERENTIALLY AFFECT COGNITIVE ABILITY TEST PERFORMANCE OF MINORITIES AND WOMEN? A META-ANALYTIC REVIEW OF EXPERIMENTAL EVIDENCE

By

Hannah-Hanh Dung Nguyen

This dissertation is a qualitative and quantitative review of the cross-study effects of stereotype threat on cognitive ability test performance of target stereotyped groups (i.e., ethnic minorities; women). The theoretical framework of the “performance interference” hypothesis was reviewed. Methodological issues and empirical evidence for moderators and mediators were summarized and discussed. The qualitative review provided conceptual rationales for meta-analytic hypotheses, extending Walton and Cohen’s (2003) work on stereotype threat effects. Regarding within-group findings, an overall mean effect size of $-.26$ was found, meaning that targets underperformed compared with their true ability when under stereotype threat. However, study artifacts explained only a small proportion of the data variance and the credibility interval overlapped zero, indicating that this finding was inconclusive and there were true moderator effects. Each methodological moderator (stereotype threat-activating cues; threat-removing strategies) or conceptual moderator (domain identification; test difficulty) was hierarchically meta-analyzed across group-based stereotypes (i.e., race-based vs. gender-based). As far as mean effect sizes were concerned, for minorities, when a race-based stereotype was activated, moderately explicit stereotype threat-activating cues produced the largest mean d , followed by blatant and subtle cues ($-.64$, $-.41$, and $-.22$, respectively). For women, subtle cues produced the largest mean d , followed by blatant and moderately explicit cues ($-.24$, $-.18$, and $-.17$, respectively). In terms of stereotype threat-removing strategies, for minorities, explicit

strategies actually enhanced stereotype threat effects (mean $d = -.80$) compared with subtle strategies (mean $d = -.34$). For women, explicit strategies were more effective in reducing stereotype threat effects than subtle ones (mean $d = -.14$ and $-.33$, respectively). In terms of domain identification, stereotype threat affected moderately math-identified women more severely (mean $d = -.52$) than highly identified women (mean $d = -.29$), whereas low math-identified women suffered the least from stereotype threat (mean $d = -.11$). In terms of test difficulty, although more difficult tests produced greater mean effect sizes for both minorities and women, women suffered less than minorities when tests were difficult (mean $d = -.36$ and $-.43$, respectively), and stereotype threat effects increased women's math test performance when the math test was easy (mean $d = -.08$). Regarding between-group findings (i.e., women vs. men; minorities vs. majority), in terms of minorities-majority test score gaps, stereotype threat-activation produced a larger mean d ($-.69$) than control conditions (no stereotype threat manipulation; mean $d = -.56$). Where stereotype threat was removed, mean d was reduced to $-.38$. For women-men math score gaps, stereotype threat-activation conditions produced a larger mean d ($-.39$) than control conditions ($-.26$) but threat-removal conditions only yielded a comparable mean d to that in control conditions ($-.23$). Although these mean effect sizes may be suggestive of the existence of stereotype threat effects as well as differential patterns of effects between women and minorities, the 90% credibility intervals overlapped zero in most cases (i.e., true stereotype threat effects might be also zero or positive), and most of V% values were smaller than 75%, suggesting that one cannot be conclusive regarding the existence of stereotype threat effects or the condition(s) under which the effects may exist because of other potential moderators' effects. Theoretical and practical implications of these findings were discussed.

Copyright by

HANNAH-HANH DUNG NGUYEN

2006

This dissertation is dedicated to my parents, my sisters Hoa, Hop, Hien and their families,
and my adorable nieces Michelle Minh and Baby Mai. I would also like to dedicate my
work to all first-generation immigrant women who persistently pursue advanced degrees
in science and social science.

ACKNOWLEDGEMENTS

Many sincere thanks to my advisor, Dr. Ann Marie Ryan, for her guidance and continuous support in the Industrial-Organizational Psychology program at Michigan State University. She introduced me to the stereotype threat research area in the first place, funding my first stereotype threat project which subsequently resulted in a publication that we co-authored, and guiding me in establishing research programs that contribute to both theory development and practical applications in the domain of employment and educational cognitive ability testing.

Further, when I approached her about changing my dissertation topics last year so that I could pursue an emerging research question (of the existence of stereotype threat effects) that could only be addressed with the meta-analysis methodology, she trusted my judgment and competence enough to give me her blessings (which may not be very common in most advisor-graduate student relationships). However, she also challenged me to give her a first draft of my meta-analytic proposal within one week's time. (I managed to do that in two weeks though.)

Ann Marie has allowed me sufficient space to grow professionally over the years; her comments and suggestions for me were always straightforward and developmental. Further, I truly appreciate the fact that she generously shared with me not only her professional viewpoints but also her personal insights about an academic life, family, and relationships with students. I consider myself very fortunate to have known and worked with Ann Marie.

I would like to thank the faculty in my dissertation guidance and oral examination committees: I appreciated Dr. Linda Jackson's skeptical viewpoints that challenged me but also helped me to think more critically about several aspects of the stereotype threat theory and empirical literature. Dr. Kevin Ford cultivated in me a strong habit of building my stereotype threat research and hypothesis formulation (or any research questions) on the basis of conceptual premises, so that my findings could contribute meaningfully to the theory development. Dr. Remus Ilies is a true meta-analyst, prompting me to run full hierarchical meta-analyses, explore between-group findings, and accurately interpret the variance in the data. Dr. Neal Schmitt asked me straightforward technical and substantive questions that pushed me to articulate my rationales and results better and more thoroughly. (Sometimes he supplied the answers himself.) I could not wish for a fairer, more helpful and more intellectually stimulating dissertation and guidance committee.

I am indebted to Dr. Gregory Walton for his generous sharing of his stereotype threat data set and unpublished articles. Thanks to his help, I was able to compare and/or locate some missing pieces of statistical information whenever article authors could not be reached or could not help me. I hope he and his co-author (Dr. Cohen) would understand that my being critical of some aspects of their meta-analysis in the present dissertation was driven by a need for advancing the theory of stereotype threat and contributing to the scientific knowledge of interest. Personally, I deeply respect their work on which I based my dissertation.

I am grateful for many stereotype threat researchers and authors (including Dr. Claude Steele) who have promptly responded to my request for additional information, who kindly forwarded my request for papers to their colleagues (particularly Dr. David

Acevedo and other former APASSC colleagues of mine), and/or who sent me a nice word of encouragement that gave me a sense of gratification.

There are several individuals whose special support and assistance has made my dissertation journey smoother and more enjoyable. I would like to thank Dr. Huy Le, my colleague and old friend, for donating a copy of his meta-analytic programs, entertaining my technical questions, and warmly welcoming me in Virginia. Many thanks to Brian Kim sharing with me his tips and advice as well as his “home-made” spreadsheet-based meta-analytic programs, and to my research assistants, Irene Nga Lam Sze and Emily Harris, for diligently coding dozens of stereotype threat papers for months.

A special “thank you” to my lady friends, fellow graduate students as well as my dear kid sister, Sonia Ghumman (also my roommate), Lauren Ramsay, Jaclyn Nowakowski, and Kathy Hoa Nguyen, who all gave me a hand in navigating and maneuvering the last stage of my dissertation process when I was out of town. I relied a lot on their kindness in handling my moving logistics and university paperwork. Lauren and Sonia even fed me from time to time. I will terribly miss these ladies’ nurturing friendship.

Speaking about friendship, I am delighted to have known and been friends with Drs. Mike and Jennifer Gillespie, Dr. Patrick Converse and his wife Suzanne, Tony Boyce, Guihyun Park, Dr. Eva Deros from the Netherlands, and Dr. Adalgisa Battistelli from Italy, whom I have met at MSU and with whom I have bonded personally and/or professionally. I would also like to acknowledge the emotional and social support of old friends, former teachers and former colleagues whom I have known for decades and who are still part of my life: Co Mai, Co Thuy, Co Kinh, Truong Huy San, Phuc Tien, Thuy

Linh, Giang Dinh, and Doan Lan. I love my new group of young and energetic Vietnamese friends: Khang Hoang and Quyen, Thuan Do and his wife Thuy, and other members of the MSU Association of Vietnamese Scholars and Students. We had a lot of fun together and my last year at MSU became much more pleasant thanks to them.

Several persons have more or less helped to shape my academic identity over the years: Dr. Philip Taylor inspired me to pursue a doctoral degree fourteen years ago in Vietnam. Dr. Quang Le introduced me to the area of I/O psychology when we were still undergraduate students. Dr. Dave Whitney alerted me to a job prospect at my alma mater which became a strong motive for me to finish my dissertation. Dr. Virginia Binder and Dr. John Jung mentored me and/or my research at California State University Long Beach. Dr. Kellina Craig was my first research advisor at CSULB. Last but not least, Dr. Paul Sackett and Dr. Maria Rotundo at the University of Minnesota, Twin Cities who gave me the first taste of I/O research and the first chance at publication.

At the risk of forgetting someone else whose help and assistance I have received and should acknowledge here, I would also like to thank the staff of the department of psychology and I/O program, particularly Julie Detwiler and Marcy Schafer who were always nice to me and helpful whenever I asked them for a favor or two. They might only do their job, but these kind ladies definitely made my graduate student life a little bit easier.

I consider myself very fortunate, receiving several fellowships during my graduate school years, particularly the National Science Foundation Graduate Fellowship Program and the Michigan State University Enrichment Fellowship Program. Their

funding allowed me to explore various research areas of interest, which was a great blessing.

Last but not least, I thank my parents, my sisters and their families, as well my extended family of almost a hundred aunts, uncles, first cousins, second cousins and other relatives of mine around the world. I am sure they are very proud of me and my academic accomplishments are as much for them as for myself.

TABLE OF CONTENTS

LIST OF TABLES.....	xiv
LIST OF FIGURES	xvi
CHAPTER 1 - INTRODUCTION.....	1
Overview.....	1
The Present Study	3
Purpose.....	3
Qualitative review	3
Meta-analytic review	4
Structure.....	8
Stereotype Threat Effects	9
Definition	9
Stereotype Threat Research Paradigm	10
Explaining Subgroup Differences in Standardized Test Scores	11
The Performance Interference Conceptual Framework	14
Test Performance: A Key Behavioral Outcome	16
Hypothesis 1	16
Antecedent: Activated Group-Based Stereotypes.....	17
Group-based stereotype threat cues	18
Potential Moderator: Presentation Modes of Stereotype Threat-Activating Cues ...	18
Hypothesis 2.	25
Potential Moderator: Presentation Modes of Stereotype Threat-Removal Strategies	27
Hypothesis 3	29
Potential Moderator: Domain Identification.....	35
Hypothesis 4	37
Potential Moderator: Test Difficulty.....	37
Hypothesis 8	39
Mediators	40
Summary.....	45
CHAPTER 2 - METHOD.....	47
Literature Search.....	48
Inclusion Criteria	49
(1) Performance Interference Hypothesis.....	51
(2) Experimental Stereotype Threat Paradigm	51
(3) Measuring Cognitive Abilities	52
(4) Number of Correct Responses	55
(5) Available Within-Group Statistics	55
(6) Available Convertible Statistics.....	55
(7) Race- and/or Gender-Based Stereotypes.....	56

(8) Language.....	56
Cumulating Results within Studies.....	57
Treatment of Independent Data Points	57
Treatment of Non-Independent Data Points	57
Treatment of Studies with a Control Condition.....	63
Treatment of Studies with Stereotype Threat x Non-Target Moderator Design	63
Treatment of Studies with Stereotype Threat x Target Moderator Design.....	64
Treatment of Studies with Stereotype Threat x Multiple Target Moderators Design	65
Treatment of Studies Where Gender Is Nested in Race	65
Treatment of Studies with Large Sample Sizes	66
Coding Studies.....	67
Coding Form and Coding Manual	69
Coding Statistics and Continuous Variables.....	69
Coding Descriptive Information	70
Coding Study Characteristics.....	70
Coding Moderators	70
Methodological Moderators.....	70
Presentation modes of stereotype threat-activation cues	70
Presentation modes of stereotype threat-removing strategies.....	70
Group-based stereotypes.....	71
Conceptual Moderators.....	71
Domain identification	71
Test difficulty.....	72
Summary of the Meta-Analytic Data Set.....	72
Meta-Analytic Procedure.....	80
Correction for Measurement Unreliability	80
Computing Effect Size.....	81
Meta-Analytic Computations and Moderator Analyses	82
Software program	83
Testing for Publication Bias	83
Summary	85
CHAPTER 3 - RESULTS	86
Within-Group Meta-Analytic Findings	86
Overall Within-Group Stereotype Threat Effects.....	86
Moderator Analysis: Group-Based Stereotypes	88
Moderator Analysis: Stereotype Threat-Activating Cues.....	90
Moderator Analysis: Stereotype Threat-Removing Strategies	94
Moderator Analysis: Domain Identification	99
Moderator Analysis: Test Difficulty.....	102
Supplemental Bias Analysis.....	106
Between-Group Meta-Analytic Findings	112
Overall Between-Group Stereotype Threat Effects	112
Potential Moderator: Group-based Stereotypes.....	115
Supplemental Meta-Analyses	119

Cognitive Ability Tests = Stereotype Threat?	119
Reference Group Members' Test Performance	122
Summary.....	125
CHAPTER 4 - DISCUSSION	130
Theoretical Implications and Implications for Research	131
Within-Group Meta-Analytic Findings	131
Moderator Meta-Analytic Findings	131
Group-based stereotypes.....	132
Stereotype threat-activating cues	133
Stereotype threat-removing strategies	135
Domain identification	138
Test difficulty.....	140
Other Potential Conceptual Moderators	143
Defense mechanisms.....	144
Race identity	146
Gender identity	147
Between-Group Meta-Analytic Findings.....	149
Practical Implications	152
Existence of Stereotype Threat Effects.....	152
Magnitudes of Stereotype Threat Effects	156
A Threat Might Be in the Air, or Not?	157
Group-Based Stereotypes Matter	158
Limitations	160
APPENDICES	164
Appendix A - Stereotype Threat-Activating Cues.....	165
Appendix B - Stereotype Threat-Removing Strategies	170
Appendix C - Excluded Studies.....	173
Appendix D - Studies with a Stereotype Threat x Domain Identification Design ($k = 22$).....	176
Appendix E - Studies with a Stereotype Threat x Test Difficulty Design ($k = 81$)....	178
Appendix F - Coding Manual	183
Appendix G - Coding Form	192
Appendix H - Hypothesized Within-Group Meta-Analytic Findings from Sensitive Subsets	197
REFERENCES	202

LIST OF TABLES

Table 1 - Definitions and Examples of Presentation Modes of Stereotype Threat-Activating Cues.....	24
Table 2 - Examples of Stereotype Threat-Removing Strategies.....	28
Table 3a - Empirically Tested Moderators Associated with Test-Taker Characteristics .	31
Table 3b - Empirically Tested Moderators Associated with Test-Related Characteristics	33
Table 3c - Empirically Tested Moderators Associated with Testing Environment Characteristics.....	34
Table 4 - Tested Mediators of the Relation of Stereotype Threat to Cognitive Ability Performance	42
Table 5 - Hypotheses to Be Tested Meta-Analytically.....	46
Table 6 - Inclusion Criteria Chart	50
Table 7 - Cognitive Ability Tests Contributing Data to the Meta-Analyses	53
Table 8 - Studies with Multiple Measures of Cognitive Abilities	59
Table 9 - Studies with Multiple Stereotype Threat Experimental Conditions.....	61
Table 10 - Overview of the Meta-Analysis Database: Characteristics of Included Studies (K = 116).....	74
Table 11 - Hierarchical Meta-Analytical Findings (Within-Group).....	87
Table 12 - Hierarchical Moderator Analyses of Stereotype Threat Activating Cues	91
Table 13 - Hierarchical Meta-Analytic Findings: Stereotype Threat-Activating Cues by Group-Based Stereotypes	93
Table 14 - Hierarchical Moderator Analyses of Stereotype Threat Removal Strategies..	96
Table 15 - Hierarchical Meta-Analytic Findings: Stereotype Threat Removal Strategies by Group-Based Stereotypes	98
Table 16 - Hierarchical Moderator Analyses of Domain Identification.....	100
Table 17 - Hierarchical Moderator Analyses of Test Difficulty.....	103
Table 18 - Hierarchical Meta-Analytic Findings: Test Difficulty Mitigating Stereotype Threat Effects by Group-Based Stereotypes.....	105
Table 19 - Hierarchical Meta-Analytic Findings of Between-Group Mean Test Performance across Stereotype Threat Levels.....	113
Table 20 - Meta-Analytic Evidence for the Equivalency between Control and Subtle Stereotype Cues Conditions.....	121

Table 21 - Meta-Analytic Findings for Reference/Comparison Groups (Within-Group)
..... 123

Table 22 - Summary of Key Hypothesis Within-Group Findings..... 127

Table 23 - Summary of Between-Group Findings..... 129

LIST OF FIGURES

Figure 1 - A heuristic model of stereotype threat effects on test performance for members of a stereotyped group.....	15
Figure 2. The funnel graph of stereotype threat effects on target test takers' cognitive ability test performance.....	107
Figure 3. The funnel graph of stereotype threat effects on target test takers' cognitive ability test performance in the absence of large sample studies.....	109

Chapter 1

INTRODUCTION

Overview

Since Steele and Aronson's (1995) seminal experiments, the research literature on stereotype threat effects on cognitive ability test performance and performance in other ability domains has steadily grown. Researchers have generally tested two key hypotheses of stereotype threat theory: the "performance interference" or short-term effects of stereotype threat, and the "school disidentification" or long-term effects (see Steele, 1997; Steele, Spencer, & Aronson, 2002). The primary (and more popular) hypothesis of *performance interference* predicts that stereotyped individuals perform worse on a task (e.g., taking a cognitive ability test) in a stereotype threatening context than they do in a non-threatening condition. Stereotype threat is defined as a self-evaluative threat experienced by some members of a stereotyped social group, whereas stereotype threat effects are defined as the detrimental performance experienced by these group members where situational cues of a salient negative stereotype exist in the immediate environment. The present dissertation focused on reviewing research that investigated the performance interference hypothesis in the domain of cognitive ability test performance.

Stereotype threat effects are commonly thought of as a between-group phenomenon, explaining the gap in cognitive ability test performance between an ethnic minority group and a majority group, or between gender groups (e.g., Black-White standardized test score differences; gender differences on SAT-math tests; Hyde & Kling,

2001; Keller, 2002; Osborne, 2001a). Industrial-organizational psychologists are interested in finding whether stereotype threat effects may explain minority applicants' performance on personnel selection tests or other workplace performance indexes as compared with test or task performance of Whites (e.g., McFarland, Lev-Arey, & Ziegert, 2003; Nguyen, O'Neal, & Ryan, 2003; Kray, Galinsky, & Thompson, 2002; Roberson, Deitch, Brief, & Block, 2003; Stone, Lynch, Sjomeling, & Darley, 1999).

Because American society is increasingly aware of diversity issues, the social message that the theory of stereotype threat conveys is powerful: members of stigmatized social groups may be constantly at risk of underperformance in academic domains, and the risks are caused by situational factors. In other words, the theory of stereotype threat mainly attributes the suboptimal test performance of stigmatized group members to malleable situational characteristics of a test or a testing condition, and not to stable factors. For example, Spencer, Steele and Quinn (1999) argue that, since stereotype threat effects on difficult math tests were observed among a highly selected and identified group of women test takers, female deficiencies in math performance may not be caused by (the lack of) innate ability but by temporary situational factors, which can be alleviated.

A decade has passed since Steele and Aronson's first studies of stereotype threat effects on cognitive ability test performance were published; the body of literature on stereotype threat effects has grown substantially. This fact indicates that the theory has influenced research and furthered the understanding of effects of negative stereotypes on behaviors. According to Devos and Banaji (2003), the strong contribution of the stereotype threat theory is that it predicts (and empirically tests) the relationship between

negative in-group stereotypes and self-relevant *behavior changes* (e.g., domain-specific task performance), not only attitudinal or affective changes. The evidence for this type of relationship is relatively rare in the broad stereotype and self-identity literature. In sum, stereotype threat theory is a high impact framework both in the social science circle and in the public.

The Present Study

Purpose

I aimed at reviewing the theoretical framework of stereotype threat theory as posited by Steele and his colleagues (Steele, 1997; Steele & Aronson, 1995; Steele, Spencer, & Aronson, 2002) and identifying existing conceptual and methodological issues that may lead to a better understanding of operational and/or interpretational ambiguities. Based on this qualitative literature review, a series of hypotheses concerning the effects of stereotype threat on stereotyped individuals' cognitive ability test performance would be tested.

Qualitative review. Several qualitative review articles or book chapters have been written on stereotype threat (e.g., Aronson, 2002; Croizet, Desert, Dutrevis & Leyens, 2001; Steele & Aronson, 1998; Steele, et al., 2002). For example, Steele, et al. (2002) provided a summary of important boundary conditions for stereotype threat effects, as well as discussing practical applications and theoretical directions. Wheeler and Petty (2001) reviewed the antecedent aspect of stereotype threat theory: explicit, conscious mechanisms of stereotype threat activation as compared with implicit, automatic mechanisms of other-stereotyping. Wheeler and Petty concluded that both conscious and unconscious self-stereotyping mechanisms might be engaged in the activation of

stereotype threat although it was not clear which would be a dominant mechanism and under what circumstances. Further, Smith (2004) reviewed underlying psychological mechanisms mediating stereotype threat effects and concluded that it was still empirically unclear why the effects occurred.

The common thread in these reviews is a strong belief in the robustness and generalizability of the stereotype threat phenomenon. Lacking in these reviews, however, is an acknowledgment and/or discussion about some conceptual and methodological issues which may lead to an ambiguity in interpreting stereotype threat effects in the literature. Therefore, these issues are identified and evaluated in this dissertation; a meta-analytic review is also conducted on relevant research questions. Meta-analytic findings shed light on the extent to which the theory of stereotype threat explains potential changes in target test takers' cognitive ability test performance because these findings are a quantitative summary of the available experimental evidence, providing a proper estimation of the mean effect size and of the variability around the point estimate across studies after error variance is controlled (Hunter & Schmidt, 1990). A meta-analysis also allows researchers to judge whether substantial variability due to moderator variables exists or whether all observed effect sizes vary across studies because of sampling error and different reliability. If the former, researchers should investigate the effect of potential moderators of stereotype threat effects across studies.

Meta-analytic review. Walton and Cohen (2003) have conducted a meta-analysis on the effects of "stereotype lift," or to what degree stereotype threat manipulation can boost test performance of non-stereotyped, reference group members. Specifically, the researchers found some support for the hypothesis that stereotype lift occurred when non-

stereotyped individuals (e.g., men; Whites) were made aware of the intellectual ability-related negative stereotypes against another social group (e.g., women; ethnic minorities). The researchers attributed stereotype lift to a boost in non-stereotyped members' self-efficacy from downward social comparisons (i.e., comparing themselves with some "inferior" social groups).

Although the focus of their meta-analysis was on stereotype lift effects among non-target test takers, Walton and Cohen (2003) reviewed stereotype threat effects among target individuals for exploratory purposes. Overall, the researchers found that meta-analytic results supported the presence of stereotype threat effects across studies (mean $d = .29$; $k = 43$). However, other moderators (i.e., the perceived relevance of a negative stereotype; the explicit refutation of the link between a stereotype and a test, and target individuals' performance domain identification) were found to influence the manifestation of stereotype threat effects. Specifically, the researchers found that, in the stereotype-irrelevant condition (i.e., where no negative stereotype was present or a negative stereotype was refuted), there was a larger stereotype threat effect among studies that refuted the link between a test and the stereotype (mean $d = .45$) than among studies that did not (mean $d = .20$). In the stereotype-relevant condition (i.e., where a negative stereotype was introduced), the stereotype threat effect was also larger among studies that reinforced the link between the test and the stereotype (mean $d = .57$) than among studies that did not (mean $d = .29$). Walton and Cohen also found that individual performance domain identification affected stereotype threat effects in that stereotype threat was larger among studies that selected students who were identified with the performance domain (mean $d = .68$) than among studies that did not (mean $d = .22$).

Although Walton and Cohen's (2003) meta-analytic results on stereotype threat effects are informative, there is room for improvement both in the conceptual grounds and the methodology of their study. Specifically, there are five main limitations in their study. First, the meta-analytic portion on stereotype threat effects in Walton and Cohen's study was exploratory and additional to the researchers' main research goal (i.e., examining stereotype lift effects). That means that the researchers investigated stereotype threat effects as in a comparison with stereotype lift effects without directly proposing conceptual hypotheses for their meta-analytic tests of threat effects. As a consequence, their meta-analytic inclusion criteria excluded studies that did not have a non-target sample (e.g., Whites; men), resulting in an artificially smaller data set of stereotype threat effect sizes than what was available in the literature at that point.

Second, ability domain identification is an important conceptual moderator of stereotype threat effects according to the theory; this construct was tested meta-analytically in Walton and Cohen's study. However, it is not clear from the report how this moderator had been operationalized in the literature in the first place, and how the levels of domain identification were coded for meta-analyses by the researchers. Because domain identification is a controversial construct which has been defined inconsistently in the literature, this moderator deserves a more thorough examination than in Walton and Cohen's study.

Third, another important theoretical boundary condition of stereotype threat effects is the difficulty degree of ability performance tests. Walton and Cohen (2003) did not meta-analytically examine this conceptual moderator. The researchers opted instead to examine only studies that used a difficult ability performance test(s). Although the

researchers explained that “because stereotype lift is assumed to alleviate the doubt, anxiety, or fear of rejection that accompanies the threat of failure, it was also required that each included study use a difficult test rather than an easy one” (p. 457), it is unclear why this prescreen step would be also necessary for a stereotype threat data set. Omitting test difficulty as a potential moderator reduces the contribution of Walton and Cohen’s meta-analytic findings to the development and understanding of stereotype threat theory.

Fourth, in terms of methodology, Walton and Cohen’s (2003) approach in the treatment of non-independent data points could have been better clarified. For example, studies in the data set which yielded hundreds of non-independent data points each (i.e., identical or overlapping samples on multiple dependent measures; e.g., Stricker, 1998; Stricker & Ward, 1998) were given a weight of .5 in the effect size computation, but the reasons and/or implications of such a treatment in regard to the variance estimation of effect sizes were neither explained nor discussed (see Hunter & Schmidt, 1990).

Last but not least, Walton and Cohen’s (2003) interpretation of stereotype threat effects mainly focused on mean standardized effect sizes (i.e., concluding that stereotype threat effects were manifested at corresponding levels of tested moderators). However, all reported heterogeneity tests of these mean standardized effect sizes in their meta-analysis (of stereotype threat) were statistically significant ($p < .05$), suggesting that there would be other moderators that further explained the variance in the data set. The implication is that the researchers’ interpretation of these detected mean effect sizes as the conclusive evidence for stereotype threat effects across studies under certain circumstances may be too hasty.

In the present review, I build on Walton and Cohen's (2003) work on stereotype threat effects but extend the meta-analysis to address the conceptual and methodological limitations of their study.

Structure

In this introduction chapter, the theory of stereotype threat and related empirical evidence is reviewed qualitatively. Specifically, I review the definitions of key concepts and premises, hypothesized and/or tested psychological mechanisms and boundary conditions. While doing so, I also identify conceptual and methodological issues that have arisen in the literature but have not been reviewed in-depth elsewhere. Some of these issues provide the rationale for another meta-analytic study in addition to Walton and Cohen (2003), focusing on testing the effects of stereotype threat on stereotyped individuals' cognitive ability test performance.

In the method chapter, I describe the procedural steps of this meta-analysis: conducting a literature search; determining inclusion criteria for study eligibility; describing moderating variables; outlining meta-analytic procedures and explaining data treatment approaches.

In the result and discussion chapters, I present meta-analytic findings that either reject or support tenets of stereotype threat, as well as discussing theoretical and practical implications of these results for future research and applied practices.

Stereotype Threat Effects

Definition

Steele and Aronson (1995) define a stereotype threat as “a social-psychological predicament that can arise from widely-known negative stereotypes about one’s group. (...) Anything one does or any of one’s features that conform to it [the predicament] make the stereotype more plausible as a self-characterization in the eyes of others, and perhaps in one’s own eyes” (p. 797). About stereotype threat effects, Steele and Aronson write, “When the allegations of the stereotype are importantly negative, this predicament may be self-threatening enough to have disruptive effects of its own.”

Steele, et al. (2002) revised the theory of stereotype threat to encompass any threatening stereotypes against any individual’s social identity, not necessarily a minority group identity (e.g., ethnic minorities; women). The theorists compared stereotype threat to a “spot light”—beaming on individuals at a specific moment. In other words, the effects are conceptually situation-specific and context-bound.

Generally, the antecedent of stereotype threat effects mainly concerns a negative stereotype or social stigma associated with an ability domain against certain social groups (e.g., women; ethnic minorities). The activation of this stereotype may increase a sense of self-threat among target test takers, which in turn influences other psychological constructs such as apprehension, distraction, and reduced motivation. The stereotype threat predicament in turn leads to handicapped performance on the ability test.

Regarding moderators, Steele (1997) posited that the degree of stereotype threat effects may vary across settings depending on the relevance or salience of the negative stereotype to targets in the setting. The theorists proposed two important boundary

conditions: (a) a high individual level of domain identification with academic or intellectual abilities, because high domain identification facilitates a strong investment in the success in such a domain, and (b) a high level of test difficulty because target group members are threatened by stereotype threat cues (e.g., feeling pressured) only when a test is at the challenging, upper-bound level of their ability (Steele and colleagues, 1995; 2002). According to the theorists, getting intimidated by the difficulty level of a test makes the negative stereotype more salient to minority test takers because they are aware that under these circumstances, they are the most likely to be judged as having limited intellectual ability. The stereotype threat conceptual framework is revisited in a following section. Next, a couple of research design issues are described and discussed, such as characteristics of the stereotype threat research paradigm and issues related to the trend of between-subgroup comparative analyses in the literature.

Stereotype Threat Research Paradigm

To test the performance interference hypothesis, stereotype threat researchers typically adopt Steele and Aronson's (1995) experimental paradigm with some modification. Members of a target social group are randomly assigned to one condition or more in which stereotype threat is manipulated (e.g., a stereotype threat-activated group vs. a control group vs. a stereotype threat-removed group). Optionally, the same research design is replicated with members of a comparison or reference group (to whom the negative stereotype is not relevant).

Note that it is inconsistent how the control condition should be operationalized in stereotype threat research. Conventionally, a control group is defined as the group that does not receive any experimental treatment or manipulation (Fisher, 1925). For example,

Study 3 in Steele and Aronson (1995) includes three conditions: a stereotype threat-primed group (test directions emphasizing the diagnosticity of the cognitive ability test), a stereotype threat-removed group (test directions describing the test as a non-diagnostic task), and a control group (no special directions given). The within-group result patterns (for both Blacks and Whites in the study) showed that both the control and the threat-removed conditions were comparable (no significant within-group mean differences on test performance).

However, Steele and Davies (2003) argued that stereotype threat-removed conditions should be treated as “true control” conditions in a stereotype threat paradigm. The authors state, “the goal was to make the negative racial stereotype irrelevant to Black participants’ performance on this task —and thus, to reduce their felt stereotype threat” (p. 315) because cognitive ability tests in evaluative settings are inherently threatening to target test takers due to embedded group-based stigmas in intellectual domains.

For methodological clarity, in the present meta-analytic review, I follow Fisher’s (1925) recommendation in treating a non-treatment condition as a control condition and a stereotype threat-removed condition as one of the experimental conditions. This approach has implications for coding study design characteristics, which is discussed in the methodology section.

Explaining Subgroup Differences in Standardized Test Scores

In Steele and Aronson’s (1995) studies, Black-White cognitive ability test performance differences were directly compared across stereotype threat conditions. The practice gives grounds to a belief that stereotype threat theory explains between-group mean differences in cognitive ability test performance in academic settings (e.g., Black-

White standardized test score differences; gender differences on SAT-math tests; see Hyde & Kling, 2001; Keller, 2002). However, this belief is conceptually and empirically debatable.

First, the stereotype threat premises and assumptions are mainly constructed for within-subgroup mean comparisons (see Steele & Aronson, 1995, 1998; Steele 1997). Second, empirical evidence does not always support such a direct between-group comparison on cognitive ability test performance (e.g., Cullen, Hardison, & Sackett, 2004; Stricker & Bejar, 2004). Third, a direct comparison for between-group mean differences may not be appropriate due to a lack of measurement invariance in criterion measures under stereotype threat (see Wicherts, Dolan, & Hessen, 2005).

The theory of stereotype threat assumes that stereotype threat would not affect reference, non-stereotyped group members because the threat is not group-relevant. In other words, the theorists do not stipulate how stereotype threat activation can cause reaction changes among members of a comparison group (e.g., Whites; men). Further, some research evidence has shown that this assumption may be partially incorrect: although the activation of a stereotype threat relevant to one stereotyped subgroup does not negatively affect test performance of a comparison, non-stereotyped subgroup, such out-group negative stereotypes favoring the comparison subgroup (e.g., men are better at math than women; Whites are better on cognitive ability tests than Blacks) are actually beneficial for the comparison group members in test performance (e.g., Spencer, et al., 1999). Walton and Cohen (2003) coined the concept of *stereotype lift effects* to refer to the small boost in the performance level of test takers of a non-stereotyped, comparison group because of the activation of stereotype threat (mean $d = 0.10$, $k = 43$), particularly

when the stereotype threat activation explicitly conveys to members of the comparison group a message about stereotyped out-group members' inferiority (or in-group superiority) in an ability domain. The implication is that, in some stereotype threat studies, the observed between-group mean differences in evaluative testing conditions (e.g., men-women; Black-White) may be accounted for by both a debilitated performance of the target group members and a performance boost of the comparison group members. This fact has not been conceptualized in stereotype threat theory.

Further, Wicherts, et al. (2005) found that stereotype threat might be a source of measurement bias in test performance scores: stereotype threat priming differentially affected mean group test scores (e.g., those of Black-White; male-female) in the majority of data sets that the researchers had re-analyzed using multigroup confirmatory factor analysis. (The exception was Nguyen, et al., 2003's study where between-group measurement invariance was observed.) In other words, under stereotype threat, cognitive ability test scores may change as a function of group memberships: upward for comparison, non-stereotyped group members and downward for target group members. The implication is that the lack of measurement invariance renders meaningful interpretation of observed group mean differences unlikely (see, for example, Horn & McArdle, 1992; Vandenberg & Lance, 2000).

Nevertheless, to provide a broad picture of stereotype threat research in the domain of cognitive ability test performance, I proposed to meta-analyze within-group mean score differences (for stereotyped groups) and explored between-group mean score differences (for both stereotyped and comparison groups). This meta-analysis replicated

and extended Walton and Cohen's (2003) meta-analyses on stereotype threat effects within a hierarchical analysis framework.

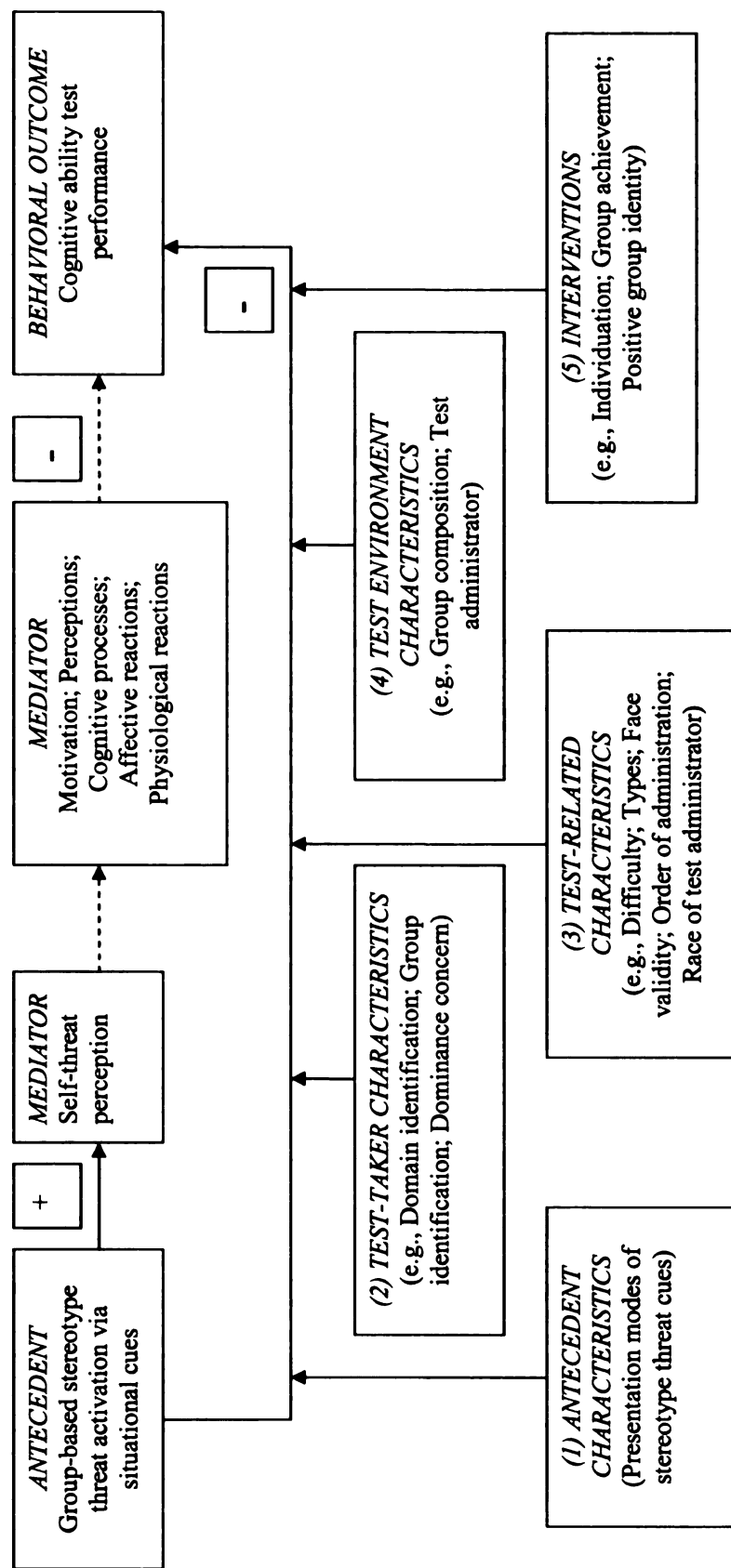
The Performance Interference Conceptual Framework

Figure 1 shows a graphic presentation of the performance interference research paradigm which pertains to a target stigmatized social group. The original stereotype threat framework has been modified to take into account recent research evidence: I redefined mediating paths and organized proposed moderators into five categories: (1) antecedent characteristics; (2) test-taker characteristics; (3) test-related characteristics; (4) test environment-related characteristics, and (5) researchers' interventions.

The components of this heuristic framework are described in the following sections. Sample empirical evidence for these components is reviewed next. Subsequently, I propose research questions and hypotheses to be tested meta-analytically.

Figure 1

A heuristic model of stereotype threat effects on test performance for members of a stereotyped group



Test Performance: A Key Behavioral Outcome

In Figure 1, the key outcome is cognitive ability test performance (e.g., on tests of quantitative, verbal, analytic, and spatial abilities). I focused on cognitive ability test performance in the present meta-analysis because researchers and society have paid the most attention to this dependent variable in the literature (e.g., most researched, discussed or debated). Practically, a majority of stereotype threat experimental studies assess cognitive ability test performance as a key outcome, thus facilitating a meta-analysis. Therefore, in the present review, it is predicted that:

Hypothesis 1. Situational stereotype threat negatively affects stereotyped test takers' cognitive ability performance across studies.

Note that the operational definition of cognitive ability test performance may be inconsistent across studies, which is problematic for evaluation and interpretation purposes. A majority of researchers have operationalized *test performance* as the number of test items correctly solved (e.g., a raw score or adjusted for guessing); this index of performance coincides with how test takers' performance is defined in the real world and yields the most meaningful interpretation. However, some researchers also defined test performance as a ratio of items correctly answered to the number of items attempted (*test accuracy*; e.g., Keller, 2002; Schmader & Johns, 2003; Shih, Pittinsky, & Ambady, 1999) because they considered this index more meaningful for the interpretation of their research questions. However, I believe that the constructs of test performance and test accuracy are distinguishable. Therefore, I consistently use correctly solved items as the definition of test performance in the present meta-analysis.

Steele and Davies (2003) asserted that about 100 studies had found support for stereotype threat effects across settings, ability domains and populations. Moreover, stereotype threat effects on cognitive ability test performance are believed to be robust across studies in laboratory settings. Walton and Cohen's (2003) meta-analysis showed that overall mean stereotype threat effect size across studies was not large (mean $d = .29$, $k = 43$). I expected to find a similar cross-study mean effect size to that in Walton and Cohen, or maybe even a smaller overall effect size because the study database of this review consists of several published stereotype threat studies which did not show significant stereotype threat effects on cognitive ability test performance (e.g., Nguyen, et al., 2003; Ployhart, Ziegert, & McFarland, 2003; Schneeberger & Williams, 2003), and these studies were published after Walton and Cohen's (2003) work had been conducted. The dataset in the present review also includes unpublished dissertations, conference papers, and working manuscripts, some of which did not find statistically significant supporting evidence for stereotype threat effects, or found weak or contrary evidence. For example, in a few studies, stereotype threatened test takers actually outperformed those who were not subjected to stereotype threat manipulation (e.g., McFarland, Kemp, Viera, & Odin, 2003).

In the following section, I review the antecedent of stereotype threat effects: group-based stereotype activation via situational cues. The presentation modes of stereotype cues are then proposed to serve as a methodological moderator of stereotype threat mean effect sizes.

Antecedent: Activated Group-Based Stereotypes

As seen in Figure 1, the activation of group-based stereotype threat via cues embedded in test directions or testing environment is the antecedent of stereotype threat effects. Specifically, when cognitive ability test performance is the outcome of interest, situational cues of group-based stereotypes are mostly related to a group-based stigma of intellectual inferiority (to other reference groups). Possibly because of the available participant pool (e.g., female college students), the most common stigma used to test stereotype threat effects is that of women's mathematics inferior capability (as compared with that of men; e.g., Inzlicht & Ben-Zeev, 2000; Sekaquaptewa & Thompson, 2002; Walsh, Hickey, & Duffy, 1999). Race/ethnicity-based stigmas in intellectual abilities have been investigated less frequently than the gender-based math stereotype but have drawn more public attention. Generally, ethnic minorities (except Asian Americans) are stereotyped as inferior in general intellectual abilities as compared with Whites (e.g., Blascovich, Spencer, Quinn, & Steele, 2001; McKay, Doverspike, Bowen-Hilton, & Martin, 2002; Steele & Aronson, 1995).

Other group-based negative stereotypes include Whites' mathematics ability inferiority compared with the superiority of Asian Americans (e.g., Aronson, Lustina, Good, Keough, Steele, & Brown, 1999; Smith & White, 2002) or inferior intelligence of college students of lower social-economic status compared with those of higher status (e.g., Croizet & Claire, 1998; Croizet, Despres, Gauzins, Huguet, Leyens, & Meot, 2004).

Potential Moderator: Presentation Modes of Stereotype Threat-Activating Cues

Previous research and reviews on stereotype threat presentation modes have shown that stereotype threat activation involves both conscious and unconscious self-stereotyping mechanisms (Wheeler & Petty, 2001), and that the explicit presentation

mode of stereotype threat-activating cues is typically more successful than an implicit one in the inducement of stereotype threat effects on test performance (Walton & Cohen, 2003). These findings contributed to the understanding of stereotype threat activating mechanisms. In the present review, I investigate whether the presentation modes of stereotype threat cues have a non-linear relationship with test performance such that extremely blatant, explicit threat cues may unexpectedly produce a performance boost in stereotype threat-activated conditions.

Walton and Cohen (2003) found that the mean threat effect size of studies that explicitly activated threat cues in test directions (reinforcing a group-based stereotype; e.g., explicitly stating that members of one social group perform worse on a test than those of another group) was larger (mean $d = .57$) than the mean effect size of studies using a subtle stereotype threat cue (e.g., merely emphasizing that a test is diagnostic of cognitive ability; making group identity salience by asking about sub-group membership prior to tests; mean $d = .29$). However, the researchers did not explain why the explicitness level of stereotype threat cues moderated the magnitude of stereotype threat effects in such a fashion.

Yet, some theories in the broad stereotype literature may predict a different pattern of findings for stereotype threat effects than those in Walton and Cohen (2003): subtle cues may have a stronger effect on task performance than explicit ones because of the direct influence of ideomotor action or stereotype-based automaticity on behaviors of interest (see Bargh, 1997). This position has received some empirical support. For example, Levy (1996) examined older adults' performance on memory tests under the influence of negative aging stereotypes. The researcher serendipitously found that the

effects of subtle priming were stronger than those of explicit priming. The implication for stereotype threat theory is that the effects of threat cues may not be equal because it is contingent on presentation modes: In Levy's study, subtle priming of negative stereotypes had a *direct* effect on memory performance through the activation of associated behavioral tendencies, bypassing conscious mechanisms, whereas explicit priming indirectly affected memory performance through some psychological mediators, somehow weakening the effect of the aging stereotype. These results were also replicated in other studies (e.g., Hess, Hinson, & Statham, 2004; Stein, Blanchard-Fields, & Herzog, 2002).

One may wonder whether the findings on subtle negative stereotype cues (as compared with that of explicit ones) in terms of memory performance can be generalized to performance in other domains that involve more complex underlying cognitive-behavioral mechanisms. As mentioned above, Walton and Cohen's (2003) meta-analytic evidence does not support such a generalization (e.g., explicit threat cues producing stronger threat effects). Yet, a few studies in the stereotype threat literature reveal an intriguing pattern of findings: explicit threat cues might sometimes inadvertently reverse the direction of stereotype threat effects (i.e., stereotype threat activation being beneficial to stereotyped groups' test performance). For example, McFarland, et al. (2003) loaded their stereotype threat condition with multiple explicit, heavy-handed stereotype threat cues, each of which had been evidenced as producing stereotype threat effects in the literature. The researchers found that women in the stereotype threat condition solved more correct math test problems than women in the stereotype threat-removed condition.

In the domain of negotiation performance, Kray, et al. (2001) found that, by explicitly informing negotiators that women lacked stereotypically masculine traits that predicted high performance in bargaining, Kray, et al. caused female negotiators in the stereotype threat condition to outperform those in the non-threat condition and even male negotiators. Kray, et al. coined the term *stereotype reactance effects* for these unexpected findings. To explain their research findings the researchers drew from Brehm's (1966) reactance theory to explain that individuals might react against a threat to their freedom by exerting their freedom more forcefully than they would otherwise. Specifically, Kray, et al. speculated that, when a negative gender-based stereotype was blatantly and explicitly activated, it might be perceived by women as a limit to their freedom and ability to perform, thereby ironically invoking behaviors that were inconsistent with the stereotype (e.g., women overperforming on tests or tasks).

Given the above conceptual and empirical grounds, I expect that the explicit levels of presenting stereotype threat cues will influence how strongly stereotype threat effects are manifested in terms of stereotyped individuals' cognitive ability test performance. In other words, the presentation mode of stereotype threat cues is a potential methodological moderator of a mean stereotype threat effect size. In the present meta-analysis, I do not simply replicate the moderating tests of priming (stereotype threat-activating) cues in Walton and Cohen's (2003) study (i.e., consisting of two levels of subtle and blatant primes), but I further refine and elaborate the operational definitions of threat cue activation.

Provided that a negative stereotype conveys a negative social message about a subgroup's relatively inferior cognitive ability in comparison with other subgroups (e.g.,

women being worse than men in mathematics; a minority ethnic group being less intellectually capable than a majority ethnic group), the presentation modes of stereotype threat cues can be classified into three distinct categories (instead of two as in Walton & Cohen, 2003): (a) direct and blatantly explicit presentation of a negative stereotype; (b) direct and moderately explicit presentation, and (c) indirect and subtle presentation. This classification scheme is based on a question: How is a stereotypical message of group-based intellectual inferiority conveyed to test takers in a research design (i.e., made salient to subgroup members)?

First, when a negative stereotype about a subgroup's inferiority in the cognitive ability domain and/or cognitive ability performance is explicitly spelled out to target test takers prior to their taking a cognitive ability test, or at least indicated as such, the presentation mode of stereotype threat cues is categorized as *blatant*. In this condition, the stereotype is most likely to invoke a theorized reactance outcome as previously discussed.

Second, when the message of general subgroup differences in cognitive abilities is explicitly conveyed to test takers, but the *direction* of these group differences is not specified and left open for test takers' interpretation, the presentation mode is labeled as *moderately explicit*. In this condition, it is speculated that the stereotypic message is direct enough to draw targets' attention, ambiguous enough to cause targets to engage in off-task thinking (e.g., trying to figure out how the message should be interpreted), but not blatant enough to make some targets become more motivated to "prove it wrong."

Last, when no statement of a relevant stereotype is made explicitly, and the *context* of testing environment or test takers' experience is subtly manipulated so that the

negative stereotype becomes salient to test takers automatically and subconsciously, the presentation mode is labeled as *subtle*. In this condition, the message of stereotype threat might not be conveyed to some targets (i.e., not registering to them) and thus producing an overall weak effect or it might work on a subconscious level and automatically and directly affect targets' test performance, thus producing a strong negative reaction.

My classification scheme of stereotype threat-activating cues is built on previous researchers' works (Walton & Cohen, 2003; Wheeler & Petty, 2001) but it differs in that the classification is logically focused more on the explicitness of a stereotype message to test takers than on the explicitness of a cue delivery mechanism. For example, even when the stereotype of subgroup inferiority in cognitive ability domains is delivered to test takers via means other than an explicit statement in test directions (e.g., a handout with information favoring males' test performance, Bailey, 2004; a questionnaire of stereotype threat statements prior to tests, Seagal, 2001), stereotype threat cues are still categorized as "explicit" in the present review.

Table 1 summarizes the operational definitions of presentation modes of stereotype threat-activating cues and some examples in each presentation mode category. (See Appendix A for a comprehensive list of stereotype threat activation cues for studies reviewed in this paper.)

Table 1

Definitions and Examples of Presentation Modes of Stereotype Threat-Activating Cues

Presentation mode	Operational definition	Stereotype threat cue	Source
<i>Blatantly explicit cues (Blatant)</i>	The message involving a stereotype about a subgroup's inferiority in cognitive ability and/or ability performance is <i>explicitly</i> conveyed to test takers prior to their taking a cognitive ability test. The group-based negative stereotype becomes salient to test takers via a conscious mechanism.	<i>Emphasizing the target subgroup's inferiority on tests (or the comparison's subgroup's superiority).</i> For example, Whites performing better than Blacks/Hispanics; men scored higher than women on math tests, etc. <i>Priming targets' group-based inferiority.</i> For example, administering a stereotype threat questionnaire before tests; giving information favoring males.	e.g., Aronson, et al. (1999); Cadanu, Maass, Frigerio, Impagliazzo, & Latinotti (2003); Schneebberger & Williams (2003) e.g. Bailey (2004); Seagal (2001)
<i>Moderately explicit cues (Explicit)</i>	The message of subgroup differences in cognitive ability and/or ability performance is conveyed <i>directly</i> to test takers in test directions or via the test-taking context but the <i>direction</i> of these group differences is left open for test takers' interpretation. The group-based negative stereotype may become salient to test takers via a conscious mechanism.	<i>Race/Gender performance differences in general ability tests.</i> For example, generally men and women perform differently on standardized math tests <i>Race/Gender performance differences on the specific test.</i> For example, taking a specific math test producing gender differences; minorities' math ability being tested on a White-normed or bias test; certain groups of people performing better than others on math exams.	e.g., Edwards (2004); Brown & Pinel (2003); Rosenthal & Crisp (2006) e.g., Keller & Dauenheimer (2003); Pellegrini (2005); Tagler (2003)
<i>Indirect and subtle cues (Subtle)</i>	The message of subgroup differences in cognitive ability is <i>not</i> directly conveyed; instead, the context of tests, test-takers' subgroup membership, or test-taking experience is <i>manipulated</i> . The group-based negative stereotype may become salient to test takers via an automatic and/or subconscious mechanism.	<i>Race/Gender priming.</i> For example, making a race/gender inquiry prior to tests; race/gender priming by other means (e.g., a pre-test questionnaire; a pre-test task; a testing environment cue), etc. <i>Emphasizing test diagnosticity purpose.</i> For example, labeling the test as a "Diagnostic Test"; stressing tests of cognitive abilities (strengths; weaknesses); stressing the evaluative nature of performance, etc.	e.g., Anderson (2001); Dinella (2004); Oswald & Harvey (2000/2001); Schmader & Johns (2003); Spicer (1999); Steele & Aronson (1995) e.g., Marx & Stapel (in press); Ployhart, et al. (2003); Prather (2005); Martin (2004)

I hypothesize about the moderating effect of the presentation modes of stereotype threat-activating cues partially based on previous meta-analytic empirical evidence (see Walton & Cohen, 2003) and partially based on the stereotype reactance theory (see Kray, et al., 2001). Specifically, I predict a non-linear relationship in which moderately explicit threat-activating cues will produce the strongest stereotype threat effects, whereas subtle threat-activating cues may produce stronger effects than blatant, explicit threat-activating one.

Hypothesis 2. *The presentation mode of stereotype threat cues moderates stereotype threat mean effect size in that studies using moderately explicit cues will produce the largest mean effect size, followed by studies using subtle cues, and then by studies using blatant cues.*

Note that a theoretical question that one may ask is how stereotype threat effects induced by explicit cues are distinguishable from the effects of a self-efficacy or self-fulfilling prophecy manipulation. In other words, an explicit message of stereotype threat may not trigger a reaction based on a fear of being stereotyped among target test takers as the theory of stereotype threat predicts, but instead such a message may reduce test takers' task self-efficacy, or increasing targets' internalized self-doubt, causing them to fulfill a prophecy of low achievement. According to Steele and colleagues, stereotype threat effects are external and situational, and thus conceptually distinctive from the internal constructs of self-efficacy and self-fulfilling prophecy (Steele, 1999; Steele & Aronson, 1995). Empirically, test self-efficacy was found unrelated to a stereotype threat manipulation and to test performance in some studies (e.g., Nguyen, et al., 2003). When a significant relationship between stereotype threat manipulation and self-efficacy was

found, women in the stereotype threat condition actually reported *higher* self-efficacy than women in the control condition (Bailey, 2004). Further, when White males, who should have had no internalized self-doubt about their group inferiority in mathematic performance, were told that Asians generally did better than Whites on math tests, they underperformed compared with those who did not learn of the Asian superiority comment (Aronson, et al., 1999). Steele (1999) concluded that the findings of this particular study showed that situational stereotype threat alone was responsible for White men's disrupted test performance. In sum, situational stereotype threat effects are conceptually distinguishable from effects of self-efficacy and self-fulfilling prophecy.

Also, one may ask whether cognitive ability tests themselves are capable of activating stereotype threat. Some researchers defined cognitive ability tests administered without any special instructions as control conditions in stereotype threat research (e.g., Ployhart, et al., 2003; Wicherts, et al., 2005). Other researchers defined cognitive ability test-only conditions as stereotype threat groups (e.g., Keller & Bless, unpublished; Oswald & Harvey, 2000/2001; Quinn & Spencer, 2001). Stereotype threat theorists believe that merely presenting relevant cognitive ability tests to stereotyped test takers may be sufficient to subtly activate stereotype threat behavior changes (c.f., Steele & Davis, 2003). This assumption was supported in some studies (e.g., Spencer, et al., 1999). Therefore, Walton and Cohen (2003) categorized studies that did not manipulate a test into the implicit stereotype threat group. However, in the present review, I choose to operationally define test-only conditions as control conditions. There are two reasons: (a) to maintain the methodological consistency in the present review, and (b) to uphold the operational concept of stereotype threat cues as *manipulated, situational cues*.

Further, I meta-analytically explore a research question of whether non-manipulated cognitive ability tests actually equate to a situational stereotype threat cue. For example, I test whether the magnitudes of stereotype threat effects differ between subsets of control and subtle stereotype threat-activated conditions. The subtlety of these stereotype threat cues (e.g., “test diagnosticity” in Steele and Aronson, 1995) may ensure that an experimental condition of stereotyped threat activation is as equivalent as possible to a non-manipulated cognitive ability test. If standardized mean test score differences between these conditions approach zero, Steele and Davis’ (2003) position is supported.

Potential Moderator: Presentation Modes of Stereotype Threat-Removal Strategies

As mentioned above, the stereotype threat research paradigm typically consists of a stereotype threat-activated condition, a stereotype threat-removed condition, and/or a control condition. Similar to stereotype threat cues, stereotype threat-removed strategies were presented to test takers subtly or explicitly in the literature. Table 2 shows some examples of stereotype threat-removing strategies. (See Appendix B for a comprehensive list of the stereotype threat removing strategies used in the meta-analytic dataset.)

Table 2

Examples of Stereotype Threat-Removing Strategies

Mode	Threat Removal	Source
<i>Subtle</i>	<i>Task purpose:</i> A problem solving task (no race inquiry before task)	e.g., Steele & Aronson (1995)
	<i>Test purpose:</i> A test development project; test performance would not be assessed.	e.g., Wout, Shih, Jackson, & Sellers (unpublished)
	<i>Indirect intervention:</i> TV commercials show women in astereotypical role (e.g., engineering and healthcare)	e.g., Davies, Spencer, Quinn, & Gerhardstein (2002)
<i>Explicit</i>	<i>Explicit intervention:</i> A handout with information favoring females.	e.g., Bailey (2004)
	<i>Explicit intervention:</i> Math test is free of gender bias (men = women)	e.g., Brown & Pinel (2003)
	<i>Explicit intervention:</i> Blacks perform better than Whites.	e.g., Cadinu, et al. (2003) e.g., Guajardo (2005)
	<i>Explicit intervention:</i> Educating subjects about stereotype threat phenomenon.	

Walton and Cohen (2003) found that studies that explicitly refuted the link between a negative stereotype and a test (or stereotype threat-removing conditions) yielded a larger mean effect size (mean $d = .45$) than studies that did not (or control conditions; mean $d = .20$). In other words, targets performed worse on a cognitive ability test in a condition where researchers tried to remove threat effects (e.g., by framing it as a non-diagnostic test or disputing group difference in performance) than in a baseline condition (e.g., the test being presented as diagnostic of ability). The researchers interpreted these findings as supporting the notion that targets might link evaluative tests to negative stereotypes automatically.

The research question to be examined here is whether studies with explicit threat removals are effective or ineffective in reducing stereotype threat effects. Walton and Cohen's (2003) findings indicate that stereotype threat effects were worsened when overt efforts were made to rectify the situation. However, explicit threat-removing strategies may serve as a catalyst to motivate individuals to avoid being prejudiced or stereotyped; this motivation can in turn inhibit negative stereotypes by shaping activated thoughts into actions toward their goals (see Spencer, Fein, Strahan, & Zanna, 2005). In other words, the explicit presentation of stereotype threat removals might play the role of a source of test-taking motivation instead of an inhibitor of performance.

Hypothesis 3. Among studies with a stereotype threat-removed condition(s), the presentation mode of stereotype threat-removing strategies moderates stereotype threat mean effect size in that studies using explicit strategies will produce a smaller mean effect size than that produced in studies using subtle strategies.

In sum, I conceptually and empirically reviewed the antecedent of stereotype threat effects: group-based stereotype threat activation. The presentation mode of stereotype-activating cues was hypothesized as a potential moderator of stereotype threat effects. Further, I hypothesized that stereotype threat-removing strategies were another potential moderator of stereotype threat effects. The research question of whether stereotype threat is inherent in any cognitive ability tests (as compared with subtly threat activated conditions) was explored.

In the next section, I review other substantive moderators of stereotype threat effects.

Other Moderators of Stereotype Threat Effects

As seen in Figure 1 (above), there are five types of methodological and conceptual moderators of stereotype threat effects on cognitive ability test performance. The methodological moderator of cue presentation modes has been discussed above. Three other categories of moderators are comparable to those in test-taking research literature (c.f., Ryan, 2001) and include (a) test-taker characteristics (e.g., domain identification; group identification); (b) test-related characteristics (e.g., types of tests; test difficulty), and (c) testing environment characteristics (e.g., test group composition; experimenters' demographic characteristics). Tables 3a-3c summarize these moderators.

Table 3a

Empirically Tested Moderators^a Associated with Test-Taker Characteristics

Moderator	Source	Stereotyped Group ^b	Relevant Research Design ^c	Significant Interaction?	Outcome ^d
<i>Domain identification</i>	Aronson, et al. (1999, Study 2)	<i>Whites</i>	ST ^e x Math identification	<i>Yes</i>	High: ST < non-ST; Moderate: ST > non-ST
	Cadinu, et al., (2003, Study 1)	<i>Women</i>	ST x Math identification	<i>Yes</i>	High: ST < non-ST; Low: no difference
	Cullen, et al. (2004)	<i>Blacks</i>	[Correlational study]	<i>n/a</i>	[Domain identification did not moderate ST effects]
	McFarland, et al. (2003)	<i>Blacks</i>	ST x Intelligence identification	<i>No</i>	[No correlation between domain identification and performance]
	Philipp & Harton (2004)	<i>Women</i>	ST x Math identification (x other IVs)	<i>No</i>	
<i>Group identity/identification (e.g., racial; gender; culture)</i>	Prather (2005)	<i>Women</i>	ST x Domain identity (x other IVs)	<i>No</i>	
	Spicer (1999)	<i>Blacks</i>	ST x Academic identity (x other IVs)	<i>(not reported)</i>	
	McFarland, et al. (2003) ^f	<i>Blacks</i>	ST x Racial identity	<i>No</i>	
	Schmader (2002)	<i>Women</i>	ST x Gender identification	<i>Yes</i>	High: ST < non-ST; Low: no difference
	Prather (2005)	<i>Women</i>	ST x Gender identity (x other IVs)	<i>No</i>	
	Nguyen, et al. (2004)	<i>Women</i>	ST x Gender identity (x other IVs)	<i>No</i>	
	Pellegrini (2005)	<i>Women (Hispanic)</i>	ST x Acculturation (x other IVs)	<i>n/a</i>	
	Rivadeneyra (2001)	<i>Latinos</i>	ST x Acculturation/cultural identity (x other IVs)	<i>No</i>	
	Smith & Hopkins (2004)	<i>Blacks</i>	ST x Acculturation (x other IVs)	<i>No</i>	

<i>Gender identity threat</i>	Seagal (2001)	<i>Blacks & Latinos</i>	ST x Social identity (x other IVs)	No	
	Schmader & Johns (2003)	<i>Women</i>	ST x gender identity threat	No	
<i>Stigma consciousness</i>	Brown & Pinel (2003)	<i>Women</i>	ST x Stigma consciousness (hi v. low)	Yes	Low: no ST effects
<i>Dominance concern tendency (testosterone)</i>	Josephs, Newman, Brown, & Beer (2003)	<i>Women</i>	ST x Dominance concern	Yes	High: ST < non-ST; Low: no difference
<i>Locus of control</i>	Cadinu, et al. (2003, Study 1)	<i>Italian women</i>	ST x Locus of control (internal v. external)	No	
<i>Coping sense of humor</i>	Smith & Hopkins (2004)	<i>Blacks</i>	ST x Locus of control (x other IVs)	No	
	Ford, Ferguson, Brooks, & Hagadone (2004)	<i>Women</i>	ST x Humor sense	Yes	High: no ST effects. Low: ST effects

^a Only variables that have been empirically examined as a moderator of stereotype threat effects are reviewed here. ^b Stereotyped Group: In some studies, the design was replicated with a comparison (non-stereotyped) group (e.g., males; Whites) but the results are not reported here because they are not relevant to the issues of interest. ^c Relevant Research Design: Study designs may be more complex (e.g., testing more than one moderator). For clarity, I reviewed only the interaction between the stereotype threat antecedent and the moderator of interest in this table. ^d Outcomes mean test performance of target group members in either stereotype threat condition or non-threat condition; ">" means "performing better than;" "<" means "performing worse than." ^e ST: stereotype threat. ^f McFarland, et al. (2003) did not find main effects for stereotype threat.

Table 3b

Empirically Tested Moderators^a Associated with Test-Related Characteristics

Moderator	Source	Target group ^b	Design ^c	Significant interaction?	Outcome ^d
<i>Test difficulty</i>	O'Brien & Crandall (2003)	<i>Women</i>	ST ^e x Test difficulty (within-subject)	<i>Yes</i>	Difficult: ST < non-ST; Easy: ST > non-ST
	Quinn & Spencer (2001)	<i>Women</i>	[ST ^f x] Math difficulty (word problems v. numerical problems)	<i>n/a</i>	Word problems < numerical problems
	Spencer, et al. (1999)	<i>Women</i>	[ST ^f x] Math difficulty	<i>n/a</i>	Difficult < Easy
	Stricker & Bejar (2004)	<i>Women; Blacks</i>	ST x Test difficulty (computer adaptive)	<i>No</i>	
	Prather (2005) Spicer (1999)	<i>Women Blacks</i>	ST x Test difficulty (x other IVs) ST x Test difficulty (x other IVs)	<i>No Yes (partially)</i>	Difficult: ST > non-ST (contrary to prediction) Easy: ST > non-ST
<i>Test face validity</i>	Ployhart, et al. (2003) ^g	<i>Blacks</i>	ST x Face validity (valid v. generic)	<i>n/a</i>	
<i>Cognitive ability domain tested</i>	Sawyer & Hollis-Sawyer (2005)	<i>Blacks; Hispanics</i>	ST x Face validity (x other IVs)	<i>Yes</i>	(contrary to prediction)
	Davies, et al. (2002; Study 2); Inzlicht & Ben-Zeev (2000)	<i>Women</i>	ST x Cognitive abilities (math v. verbal tests)	<i>Yes</i>	Math test: ST < non-ST; Verbal test: no difference
	McFarland, et al. (2003)	<i>Blacks</i>	ST x Order of test (prior to a personality measure v. after)	<i>No</i>	
<i>Test administration order</i>	Sternberg, Sternberg, Jarvin, Leighton, Newman, Moon, et al. (unpublished) ^f	<i>Women</i>	ST x Order of tests (verbal v. math)	<i>No</i>	

Note. ^a, ^b, ^c, ^d, ^e See notes from Table 3a. ^f Quinn and Spencer (2001) and Spencer, et al. (1999) did not manipulate stereotype threat primes but assumed that a math test was an inherent stereotype threat predicament for women. ^g Ployhart, et al. (2003) did not find main effects for stereotype threat and did not report significant tests for the effects of test face validity within Blacks. However, mean test scores for Blacks across conditions are compared here.

Table 3c
Empirically Tested Moderators^a Associated with Testing Environment Characteristics

Moderator	Source	Target group ^b	Relevant Design ^c	Significant interaction?	Outcome ^d
<i>Test group composition</i>	Inzlicht & Ben-Zeev (2003)	<i>Women</i>	ST ^e x Solo status (male-dominant group v. female-dominant)	<i>No</i>	
<i>Test environment (e.g., hostile)</i>	Oswald & Harvey (2000/2001)	<i>Women</i>	ST x Hostile test environment (sexist cartoon v. no cartoon)	<i>Yes</i>	ST: hostile > non-hostile Non-ST: hostile < non-hostile
<i>Test administrator (ethnicity)</i>	Wout, et al. (unpublished)	<i>Blacks</i>	ST x Test administrator ethnicity (Black v. White)	<i>Yes</i>	ST: White administrator > Black administrator Non-ST: no differences
	Walters (2000)	<i>Blacks</i>	ST x Test administrator ethnicity (Black v. White) (x other IV)	<i>No</i>	

Note. a, b, c, d, e See notes from Table 3a.

The last category of moderators contains “intervention” or “protective” strategies that some stereotype threat researchers may employ to buffer stereotype threat effects for stereotyped group members. These strategies include invoking individuation by asking one to disclose his or her personal information so that his or her individuality became more identifiable (Ambady, Paik, Steele, Owen-Smith, & Mitchell, 2004), emphasizing a group’s achievements in society (McIntyre, Paulson, & Lord, 2002), reminding test takers of a group identity linked to a positive stereotype (Shih, et al., 1999), or forewarning test takers about potential stereotype threat in cognitive ability tests (Williams, 2004).

Among factors listed in Tables 3a-3c, I focus on two potential moderators for mean stereotype threat effect size across studies in the present meta-analysis: (a) domain identification, and (b) test difficulty. There are conceptual and empirical rationales for my selection (to be discussed in following sections); there is also a practical reason. Compared with other potential moderators of stereotype threat effects, these variables have been investigated more frequently and/or more conceptually emphasized than other variables, thus providing sufficiently large sub-samples to test relevant moderating hypotheses.

Potential Moderator: Domain Identification.

According to Steele et al. (2002), the strength of stereotype threat effects is contingent on “how much the person identifies with the domain of activity to which the stereotype applies,” or “the degree to which one’s self-regard, or some component of it, depends on the outcomes one experiences in the domain” (p. 390). Specifically, only those who strongly identify themselves with a domain in which a negative stigma against

their social group is embedded are susceptible to the possibility of confirming group-based stigma about their own ability. The theorists further equate high domain identification with elite, academically high-achieving African American college students, as opposed to urban Black students or “those who identify less with school—often weaker, less confident students—because they do not care so much about academic success” (Steele & Aronson, 1998, p. 402). This notion is believed to hold true for other stereotyped groups, such as women in a mathematic ability domain.

Domain identification as a moderator of stereotype threat effects has been studied only sparingly (Aronson, et al, 1999; Cadinu, et al., 2003; Leyens, Desert, Croizet, & Darcis, 2000; McFarland, et al., 2003). Although Steele and Aronson (1995) measured academic identification of Blacks and Whites (operationally defined as perceptions of the importance of verbal and math skills to education and intended career), Aronson, et al. (1999, Study 2) provided the first direct empirical evidence for the moderating effect of domain identification on stereotype threat. Using a highly math-able sample, the researchers found that math domain identification significantly interacted with stereotype threat. Specifically, among high math-identifiers, the threat-removed group outperformed the threat-primed group. However, among the moderately high math-identifiers, the reverse results were obtained. Some studies replicated Aronson, et al.’s findings (e.g., Cadinu, et al., 2003; Leyens, et al., 2000); other studies did not (e.g., McFarland, et al., 2003).

Based on Steele and colleagues’ theoretical premise (1995; 2002) and Aronson, et al.’s (1999) empirical evidence, domain identification has become a pre-screening criterion in some stereotype threat studies (e.g., Brown & Pinel, 2003; Davies, et al.,

2002; Quinn & Spencer, 2001) based on the assumption that stereotype threat effects are most likely to be observed among highly domain identified individuals. The role of domain identification as a boundary condition of stereotype threat effects was supported by Walton and Cohen's (2003) meta-analytic findings: studies using a selective sample of students of high domain identification yielded a larger mean threat effect size (0.68) than the mean effect size in studies with non-selective samples (0.22).

In the present dissertation, Walton and Cohen's (2003) study was partially replicated and extended by testing domain identification as a substantive moderator with three levels (instead of two; high; moderate and low). The question to address was whether these levels of domain identification mitigate stereotype threat effects across studies. Based on the theory of stereotype threat, it is predicted that the magnitude of threat effects across studies increases as the level of domain identification increases.

Hypothesis 4. Studies with a sample of test takers who are highly domain identified will produce the largest mean effect size, followed by studies with a sample moderate in domain identified, and then by studies with a sample low in domain identification.

Potential Moderator: Test Difficulty

Stereotype threat theorists consider test difficulty an important moderator of stereotype threat effects on cognitive ability test performance. The rationale is that target group members are most likely to be threatened by stereotype threat cues (e.g., feeling pressured) only when a test is at the challenging, upper-bound level of their ability (Steele and colleagues, 1995; 2002). In other words, only when facing the intimidating difficulty of a test that stereotyped individuals become aware of the fact that they are very

likely to confirm the negative stereotype of ability inferiority about their social group. Difficult-test conditions may enhance the effects of stereotype threat because facing a challenging test in stereotype threat conditions, target individuals may become dejected, distracted, and de-motivated, or experience decreased mental workload, which in turn may negatively influence their test performance (see Figure 1 for and Table 4 above for a review of mediators of stereotype threat effects).

Empirical evidence for the moderating effects of test difficulty is mixed. In a sample of highly math-able women and men (prescreened for college math grades of “B” or better and for scoring at the 85th percentile or above on SAT math tests), Spencer, et al. (1999, Study 1) found no gender differences in performance on a test of easier math problems (i.e., algebra, trigonometry & geometry), but there were gender differences in performance on a difficult math test (i.e., advanced calculus). Note that Spencer, et al. did not manipulate stereotype threat explicitly but assumed that stereotype threat was inherently embedded in a difficult math test for women. Therefore, the researchers interpreted the findings as supporting evidence for the link between test difficulty and stereotype threat effects for women.

Explicitly manipulating stereotype threat (e.g., telling subjects in threat condition that the math test “shows gender differences” vs. “no gender differences” in the control group), O’Brien and Crandall (2003) found that a difficult SAT math test produced stereotype threat effects for women under stereotype threat as compared with women in stereotype threat-removed conditions. However, on an easier math test of 3-digit multiplication problems, women in the stereotype threat-primed condition outperformed women in the stereotype threat-removed condition.

Stricker and Bejar (2004) administered a standardized cognitive ability test (GRE quantitative, verbal, and analytic tests) to participants across gender groups and racial groups (women vs. men; Blacks vs. Whites). Stereotype threat was manipulated in this study. Because the administration of the GRE tests was computerized, the researchers were able to manipulate the difficulty levels of the whole tests directly (i.e., either administering an actual, more difficult battery of GRE tests or a modified, easier battery of GRE tests). They found no significant within-group stereotype threat effects regardless of test difficulty levels.

The construct of test difficulty as a conceptual moderator was chosen in the present meta-analysis for similar reasons as those for domain identification. Theory-wise, test difficulty is a boundary condition of stereotype threat effects. Design-wise, several researchers purposefully employed a highly difficult cognitive ability test to investigate stereotype threat effects (e.g., Croizet, et al., 2004; Gonzales, Blanton, & Williams, 2002; Inzlicht & Ben-Zeev, 2003; McIntyre, et al., 2002; Schmader, 2002; Steele & Aronson, 1995). Some other researchers used moderately difficult tests (e.g., Dodge, Williams, & Blanton, 2001; McKay, et al., 2002; Smith & White, 2002; Stricker & Ward, 2004), suggesting potential variance in stereotype threat effects across these studies. In accordance with stereotype threat theory, it is hypothesized that test difficulty levels moderate threat effect magnitudes across studies.

Hypothesis 5. Studies using highly difficult cognitive ability tests will yield the largest mean effect size, followed by studies using moderately difficult tests, and then by studies using easy tests.

Note that, among studies that include sub-samples at various levels of test difficulty (difficult, medium, easy), each sub-sample would yield independent data to be meta-analyzed. More common are studies in which only one level of test difficulty is used also contributing independent data points for the dataset. The procedure is further described in the Method chapter.

Potential Mediators

Figure 1 (above) shows the hypothesized mediating path in stereotype threat theory. (The plus and minus signs show a hypothesized direction of relationships). First, stereotyped individuals might feel a heightened perception of self-threat in a stereotype threatening situation. This perception might in turn lead to a host of other psychological mechanisms that can mediate the relationship between a situational threat and one's test performance.

Dashed lines are used for part of the mediating paths to reflect the fact that researchers have been mostly unsuccessful in testing these paths statistically. I next briefly review some previous (and mostly unsuccessful) efforts in testing for mediation in the literature. However, I do not propose meta-analytic tests of mediators because of the lack of empirical support for these mediators.

There are approximately two dozen variables that have been investigated and/or tested statistically as potential mediators of stereotype threat effects (in published papers only). Table 4 lists these mediators in several categories, such as perceptual (self-threat perceptions and those of situations), emotional, motivational, and cognitive constructs. It shows that researchers have focused on an array of theoretically sound psychological mechanisms that might explain why stereotype threat effects occur. Note that in some

earlier studies, researchers failed to subject a hypothesized mediating path to a rigorous statistical test of mediation. For example, Steele and Aronson (1995, Study 3) found that stereotype threat activation significantly transformed participants' perceptual, cognitive and emotional reactions. Stangor, Carr, and Kiang (1998) demonstrated that stereotype threat activation lowered target participants' performance expectancies. However, the researchers only inferred the existence of mediating effects of these perceptual and emotional changes to African Americans' test performance, but did not statistically test them. Most mediation tests in other studies were either unfruitful or researchers did not find stereotype threat effects in the first place to test mediation (e.g., Mayer & Hanges, 2003).

Table 4

Tested Mediators of the Relation of Stereotype Threat to Cognitive Ability Performance

Potential mediator	Study	Significantly statistical test of mediation?
<u>Self-threat Perceptions</u>		
<i>Perception of stereotype threat</i>	Steele & Aronson (1995)	No
	Gonzales, et al. (2002)	No
	Leyens, et al. (2000)	No
	McKay, et al. (2002)	No
		No
	Mayer & Hanges (2003) [threat-general vs. threat- specific]	No
<i>Activated self-relevant stereotypes</i>	Davies, et al. (2002)	Yes (full mediation)
<i>Gender identity threat</i>	Schmader & Johns (2003)	No
<i>Self-perceived task competence</i>	Steele & Aronson (1995)	No
	Gonzales, et al. (2002)	No
<u>Perceptions of the situation</u>		
<i>Perceived test difficulty</i>	Schmader & Johns (2003)	No
<i>Perceived test face validity</i>	Dodge, et al. (2001)	No
<u>Other individual characteristics</u>		
<i>Intelligence domain identification (general)</i>	McFarland, et al. (2003)	No
	Leyens, et al. (2002)	No
<i>Stigma consciousness</i>	Brown & Pinel (2003)	No
<i>Self-perceived math ability</i>	Brown & Pinel (2003)	No
<u>Emotions</u>		
<i>Test anxiety (situational, state)</i>	Steele & Aronson (1995)	No
	Dodge, et al. (2001)	No
	Keller & Dauenheimer (2003)	No
	Smith & White (2002)	No
	Oswald & Harvey (2000/2001)	No
	Osborne (2001b)	[Yes] ^a
<i>Evaluation apprehension</i>	Mayer & Hanges (2003)	No
	Spencer, et al. (1999)	No
<i>Expected affective reactions</i>	Stangor, et al. (1999)	No
	Oswald & Harvey (2000/2001)	No

<i>Mood</i>	Smith & White (2002)	No
<i>Dejection</i>	Keller & Dauenheimer (2003)	Yes (full mediation)
<u>Motivational constructs</u>		
<i>Test-taking effort/motivation (reduced)</i>	Steele & Aronson (1995)	No
	Gonzales, et al. (2002)	No
	Brown & Pinel (2003)	No
	Keller (2002); Keller & Dauenheimer (2003)	No
	Schneeberger & Williams (2003)	No
	Dodge, et al. (2001)	No
<i>Test-taking self-efficacy</i>	Dodge, et al. (2001)	No
	Spencer et al. (1999)	No
	Mayer & Hanges (2003)	No
<i>Performance expectancies</i>	Cadinu et al. (2003, Study 1)	Yes (partial mediation)
	Sekaquaptewa & Thompson (2002)	No
<i>Heightened arousal</i>	O'Brien & Crandall (2003)	No
<i>Regulatory foci</i>	Seibt & Forster (2004)	Yes
<u>Cognitive processes</u>		
<i>Cognitive interference (off-task thinking; distractibility)</i>	Steele & Aronson (1995)	No
	Gonzales, et al. (2002)	No
	Dodge, et al. (2001)	No
	Mayer & Hanges (2003)	No
	Oswald & Harvey (2000/2001)	No
	Prime (2000)	No
	Croizet, et al. (2004)	Yes (full mediation)
<i>Self-handicapping</i>	Croizet & Claire (1998)	No
	Keller & Dauenheimer (2003)	No
	Keller (2002)	[Yes] ^b
<i>Inhibited mental processes</i>	Quinn & Spencer (2001)	No
<i>Working memory capacity</i>	Schmader & Johns (2003; Study 3)	[Yes] ^b
	Schneeberger & Williams (2003)	No
<i>Testing speed</i>	Prime (2000)	No

Note. (*) Some researchers have tested mediating paths statistically; some others did not attempt to do so; yet other researchers did not find significant stereotype threat effects to test mediation. These studies are all listed as not providing significant statistical evidence for a mediating path. ^a Osborne (2001b) actually conducted tests of test anxiety as a mediator for the relationship between group memberships (e.g.,

ethnicity or gender) and standardized cognitive ability test scores, not stereotype threat effects on test scores per se. Therefore, I would hesitate to include this study in the small group of studies that have successfully established a mediating path for stereotype threat effects.^b Keller (2002) and Schmader & Johns (2003) used the ratio of numbers of correct items over numbers of items attempted for mediation analyses, not the numbers of correct items.

In sum, the theory of stereotype threat is found wanting in explaining why performance interference happens (or not) for target individuals under activated stereotype threat. Findings from the present meta-analysis may help to clarify when stereotype threat effects occur, so that tests of why they occur can be more systematic and fruitful in the future.

Summary

A heuristic model of stereotype threat effects was presented, which reflects the theory and takes into account recent research evidence. The key components in this model (an antecedent, a behavior outcome, moderators, and mediators) were reviewed. The qualitative review provided rationales for subsequent meta-analytic hypotheses of general cross-study mean effect size and conceptual and methodological moderators of stereotype threat effects. Some of these hypotheses were built on and extended Walton and Cohen's (2003) work on the effects of stereotype threat. Table 5 summarizes the hypotheses.

Table 5

Hypotheses to Be Tested Meta-Analytically

No.	Hypothesis
H1	<i>Situational stereotype threat negatively affects stereotyped test takers' cognitive ability performance across studies.</i>
H2	<i>The presentation mode of stereotype threat cues moderates stereotype threat mean effect size in that studies using moderately explicit cues will produce the largest mean effect size, followed by studies using subtle cues, and then by studies using blatant cues.</i>
H3	<i>Among studies with a stereotype threat-removed condition(s), the presentation mode of stereotype threat-removing strategies moderates stereotype threat mean effect size in that studies using explicit strategies will produce a smaller mean effect size than that produced in studies using subtle strategies.</i>
H4	<i>Studies with a sample of test takers who are highly domain identified will produce the largest mean effect size, followed by studies with a sample moderate in domain identified, and then by studies with a sample low in domain identification.</i>
H5	<i>Studies using highly difficult cognitive ability tests will yield the largest mean effect size, followed by studies using moderately difficult tests, and then by studies using easy tests.</i>

Chapter 2

METHOD

The main goal of this dissertation manuscript is to analyze the cross-study effect of stereotype threat on cognitive ability test performance of members of stereotyped groups, in comparison with test performance of themselves when in non-stereotyped situations, and other social groups. The findings may have implications for both theory development and practical applications. I used the meta-analytic method to achieve this goal. Meta-analysis was first proposed by Glass (1976) to integrate and summarize the findings from individual studies in a body of research. Glass writes, “Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature” (Glass, 1976, p.3). In other words, a meta-analysis can help a reviewer to code studies into a data set and then manipulate, measure, and display the data in a common, comparable metric system. The outline of the meta-analytic steps in this dissertation is as follows:

1. Conducting a literature search;
2. Setting inclusion criteria;
3. Making treatment decisions of data to cumulate within studies;
4. Creating a coding scheme for coding studies and coding moderators;
5. Entering the data for each study, and
6. Exploring and displaying the data with meta-analytic statistical techniques.

Literature Search

First, a computerized bibliographic search of electronic databases such as PsycINFO (including PsycARTICLE) and PROQUEST was conducted, using the combined keywords of “stereotype” and “threat” as search parameters. This literature search yielded peer-reviewed journal articles and dissertation abstracts dated between 1995 (the publication year of the original article by Steele and Aronson) and April 2006. All available stereotype threat articles and loaned circulating dissertations were obtained. If a dissertation copy was not circulated, online search for the dissertation author’s contact information was conducted and a direct request for a copy of dissertation was sent. (A few dissertations were downloadable from the internet.)

Second, I conducted a manual search by reviewing the reference list of each identified article to find additional citations that may not be revealed by a computerized search (e.g., conference papers; unpublished manuscripts). I also used the internet search engines of google.com and scholar.google.com to search for unpublished empirical papers of interest and self-identified stereotype threat researchers (using the key terms “stereotype threat” and “test performance”).

The scholar.google.com search returned 620 hits some of which contained original information about published articles on stereotype threat. Because the google.com search returned 17,200 hits, I browsed through the first 1,720 hits (10%) for original entries on downloadable unpublished empirical studies and/or self-identified stereotype threat researchers (e.g., from their own description of research interests). (Many of the subsequent entries became repetitive and did not offer original information; they were thus disregarded.) Once stereotype threat researchers were found and if their

email addresses were available online, I sent these researchers a “cold” email, requesting their manuscripts and/or working papers if any. I also posted the same request on various psychological list-serves. Furthermore, several prominent researchers in the stereotype threat area were contacted for unpublished manuscripts, in-press papers, as well as other additional sources of research data on stereotype threat effects on cognitive ability test performance. A portion of these requests was successful.

The preliminary database of stereotype threat papers gathered were then subject to a set of inclusion criteria to determine which studies should be retained and coded for this meta-analysis.

Inclusion Criteria

Bobko, Roth, and Potosky (1999) emphasize the importance of applying consistent decision rules in selecting studies to include in a meta-analysis. Therefore, several criteria had to be consistently met for a study to be included in the present meta-analysis. Table 6 summarizes the inclusion criteria and brief explanations.

Table 6

Inclusion Criteria Chart

No.	Criterion	Definition
1	Is this study designed to empirically test the " <i>performance interference</i> " hypothesis in cognitive ability testing situations?	The hypothesis pertains to test takers' handicapped performance on a cognitive ability test due to stereotype threat.
2	Does this study follow the <i>experimental stereotype threat paradigm</i> ?	The study must include at least an independent variable (stereotype threat) in an experimental design.
3	Is a domain of <i>cognitive abilities</i> measured? (Is the dependent measure a cognitive ability test?)	Cognitive ability test performance must be a dependent variable. Cognitive ability test performance data may include scores on mathematical, verbal, analytical, and/or spatial ability tests.
4	Is test performance operationalized as the <i>number of correctly solved items</i> ?	If test performance was defined differently (e.g., a mean ratio of correct answers to attempted answers), the data should be converted into the <i>number of correct answered</i> if possible.
5	Can <i>within-group data</i> be extracted (at least for a stereotyped group)?	When the stereotype activated is gender-based, for instance, within-group data consist of the test performance of women across levels of stereotype threat.
6	Is an effect size <i>computable</i> ? (Are there sufficient statistics to compute an effect size for this study?)	Convertible statistics include sample sizes, means, standard deviations, correlation estimates, independent sample t-test estimates, and F-test estimates.
7	Is the negative stereotype <i>race-based</i> and/or <i>gender-based</i> ?	Age-based or social class-based stereotypes are not included in this meta-analysis.
8	Is the study written in <i>English</i> ?	<i>Note.</i> Studies written in a foreign language are included (e.g., French, Dutch, German) if English translation is also available.

(1) Performance Interference Hypothesis

Only studies designed to test the within-subgroup “performance interference” hypothesis or the short-term effects of stereotype threat manipulation on cognitive ability test performance were included in this meta-analysis. Studies that investigated the hypothesis of “school disidentification” or the long-term effect of stereotype threat were excluded. (Appendix C lists all excluded papers and unmet criteria.)

(2) Experimental Stereotype Threat Paradigm

Steele and Aronson (1995) provide the basic research paradigm for testing the hypothesis of performance interference due to the manipulation of a situational threat. Therefore, studies to be included in this meta-analysis were those that followed the relevant experimental paradigm: involving random assignment of participants to a stereotype threat-activated (treatment) group where at least one situational stereotype threat cue was introduced and manipulated, and at least a comparison group who received either a stereotype threat-removed intervention or no intervention (control).

Empirical studies that drew inferences from the theory of stereotype threat but were executed within a different research framework were excluded from this review (e.g., a study investigating the effect of mentoring on performance; Good, Aronson, & Inzlicht, 2003; studies using the solo-status paradigm without manipulating stereotype threat directly; Ben-Zeev, Fein & Inzlicht, 2005; Inzlicht & Ben-Zeev, 2000, 2003). Correlation studies (including longitudinal and cross-sectional studies) that tested the hypothesis of school/academic disidentification were not included in this meta-analysis (e.g., Aronson, Fried, & Good 2002). Further, correlation studies that purported to test stereotype threat theory but did not directly measure the correlation of stereotype threat manipulation and

cognitive ability test performance, which is the effect of interest, were also excluded (e.g., Chung-Herrera, Ehrhart, Ehrhart, Hattrup, & Solamon, 2005; Cullen, Hardison, & Sackett, 2004; Osborne, 2002; Roberson, Deitch, Brief, & Block, 2003).

(3) Measuring Cognitive Abilities

Only studies assessing performance on cognitive ability tests were included in this meta-analysis. The cognitive ability domain is narrowly operationalized here as quantitative, verbal, analytic and spatial abilities, or any combination of these cognitive abilities that are typically assessed with standardized cognitive ability tests in educational and employment settings (e.g., SAT; ACT; GRE; WPT). Therefore, studies targeting other ability domains, such as memory, work-related behaviors, athletic sensori-motor skills, and other types of test performance were excluded from the sample (e.g., DeRouin, Fritzsche, & Salas, 2004; Frantz, Cuddy, Burnett, Ray, & Hart, 2004). Table 7 presents cognitive ability tests that contributed data to this meta-analysis. Note that most cognitive ability tests used as dependent measures in the preliminary database are derived from standardized tests such as GRE, GMAT, and SAT. Most of them consist of a subset of test items (specific test item content varying across studies) and rarely is reliability information of these tests reported. On the other hand, some researchers employ standardized intelligence tests the reliability of which is either reported in research reports or retrievable from the broad testing literature (e.g., Wonderlic Personnel Test; The Wide Range Achievement Test; Raven Advanced Progressive Matrices). The sporadic report of dependent measurement reliability information has implications for the method of computing the variation in cross-study d -values (i.e., only a handful of reliability coefficients constituting the artifact distribution in the present meta-analysis).

Table 7

Cognitive Ability Tests Contributing Data to the Meta-Analyses

	Test name	Reliability r_{yy}
	<u>Mathematical ability tests</u>	
1	Canadian Math Competition	n/a
2	GRE mathematical ability	n/a
3	GRE calculus	n/a
3	SAT mathematical ability	n/a
	PSAT mathematical ability	n/a
4	ACT mathematical ability	n/a
5	GMAT mathematical ability	n/a
6	General Equivalency Diploma—mathematical ability	n/a
6	The 3rd International Mathematics & Science Study (Keller, 2002; Keller 2006, 1a-1b)	$\alpha = .80$ (source: literature)
7	AP calculus	n/a
7	Dutch Differential Aptitude Test--mathematic subtest (Wicherts, Dolan & Hessen, 2005)	K/R $r = .94$ (source: literature)
8	Multidimensional Aptitude Battery - Arithmetic subtest (Guajardo, unpublished)	$\alpha = .95$ (source: literature)
9	The Wide Range Achievement Test--Arithmetic subtest (Smith & Hopkins, 2004)	Split-half $r = .94$ (source: literature)
10	The Weschler Adult Intelligence Scale (WAIS-III)— Arithmetic subtest (Pellegrini, 2005)	$\alpha = .84$ (averaged; source: literature)
11	Natural World Form 5 (Math & Science) (Anderson, 2001, 1a-1b)	$\alpha = .61$ to $.75$ (source: literature)
12	Math tests (e.g., simple multiplications; sum; word problems) (Nguyen, Shivpuri, Ryan, & Langset, 2004)	$\alpha = .603$
	<u>Verbal ability tests</u>	
13	GRE verbal ability	n/a
14	SAT verbal ability	n/a
15	Verbal Aptitude Test	n/a
16	Verbal tests (e.g., analogy, reading comprehension, sentence skills)	n/a
	<u>Analytical ability tests</u>	
17	GRE analytical ability	n/a
18	The Weschler Adult Intelligence Scale (WAIS-III)— Similarities subtest (Pellegrini, 2005)	$\alpha = .75$ (average; source: literature)
19	Analytical ability tests	n/a
	<u>Spatial ability tests</u>	
20	Culture Fair Intelligence Matrices (Sawyer & Hollis- Sawyer, 2005, 1a-1b)	$\alpha = .88$ (source: literature)
21	Raven Advanced Progressive Matrices (Brown & Day, in press; McKay, 1999; von Hippel, et al., 2005)	$\alpha = .89$ (source: literature)
22	Mental Rotation Test (Keller & Bless, unpublished; Martens, et al., 2006)	Spearman-Brown $r = .86$ (source: literature)
23	The Weschler Adult Intelligence Scale (WAIS-III)—Block Design subtest (Pellegrini, 2005)	$\alpha = .86$ (averaged; source: literature)

24	Wonderlic Personnel Test (Martin, 2004)	$\alpha = .90$
25	Unobtrusive Knowledge Test (Excellence Scale) (Martin, 2004)	$\alpha = .85$
26	Cognitive ability tests (Nguyen, et al., 2003)	$\alpha = .822$

Note. The reliability of dependent measures (cognitive ability tests) was sporadically reported. I substituted reported reliability coefficients of some well-known cognitive ability tests in the broad testing literature wherever possible. These reliability coefficients constituted the artifact distribution in subsequent meta-analyses.

(4) Number of Correct Responses

Test performance is consistently operationalized as the number of correct answers in the present meta-analysis. Therefore, for studies that used a different index of performance (e.g., a ratio of correct answers to attempted problems), I tried to convert indexes from available reported information or contacted study authors for the information of interest.

(5) Available Within-Group Statistics

The studies retained in the database yielded at least within-group findings (for target group members). For example, when the negative stereotype activated was race-based (i.e., an intellectual inferiority stigma associated with minority subgroups), minority test takers had to be randomly assigned to a stereotype threat-activated condition or a comparison (stereotype threat-removed or control) condition. Likewise, when the stereotype was gender-based (i.e., women's mathematic inferiority stigma), female test takers had to be randomly assigned to experimental conditions. Additionally, a comparison subgroup(s) could be included in the research design, yielding additional within-subgroup findings for the comparison subgroup.

(6) Available Convertible Statistics

The studies included had to yield precise statistics that were convertible to a weighted effect size (e.g., mean performance differences between women in a treatment condition and those in a comparison condition). Therefore, several studies were excluded because of insufficient information (see Appendix C). Note that I had contacted authors of these studies to obtain convertible statistics but either the statistics were not available, researchers could not be reached, or researchers did not respond to the request. I also

consulted with the statistics used in Walton and Cohen's (2003) meta-analysis for some of the missing values.

(7) Race- and/or Gender-Based Stereotypes

For the scope of this meta-analysis, only studies investigating race-based and gender-based stereotypes concerning (the inferiority of) some domain of intellectual abilities of a subgroup were included in the review database. Therefore, studies examining age-based stereotypes about learning and memory, race-based stereotypes about athletic abilities, and class/college major-based stereotypes about intelligence were excluded (see Appendix C).

Several studies were a special case (Aronson, et al., 1999; Smith & White, 2002; von Hippel, von Hippel, Conway, Preacher, Schooler, & Radvansky, 2005). The stereotyped targets in these studies are White males and the negative stereotype was based on another ethnic group's intellectual superiority to Whites (i.e., Asians are better on math tests than Whites). Therefore, these studies contributed estimates of effect size to the meta-analysis database only where appropriate (e.g., used to cumulate an overall mean effect size across samples and conduct some moderator analyses).

(8) Language

Studies written in English (or could be translated into English) were included in the sample. One unpublished paper in Dutch (with an English introduction) was found and included in the sample. My limited language capability may introduce a bias in location of studies to the present meta-analysis. As Egger and Smith (1998) noticed and evidenced, non-English speaking authors "might be more likely to report positive findings in an international, English language journal and negative findings in a local

journal.” The implication is that there may be more null-finding studies existing in the non-English literature than what are gathered in this data set.

Cumulating Results within Studies

Treatment of Independent Data Points

An article or research paper might consist of multiple single studies each of which contributes an independent estimate of effect size. For these studies, I treated each of them as an independent source of effect size estimates.

There were also single studies that produced replication of observations of the stereotype threat-test performance relationship within a study. In this case, I followed Hunter and Schmidt’s (1990) recommendations and decided whether or not to cumulate results within a study. When a single study employed a fully replicated design across demographic subgroups (i.e., a conceptually equivalent but statistically independent design), I treated the data as if they were values from different studies. For example, when cognitive ability test performance data were gathered for separate racial/ethnic subgroups (e.g., Hispanic Americans and African Americans; Sawyer & Hollis-Sawyer, 2005), the cognitive ability test scores from each ethnic subgroup are statistically independent and are treated as such in this meta-analysis.

Treatment of Non-Independent Data Points

When conceptual replication occurs within a study (i.e., each subject provides more than one observation that is relevant to the stereotype threat-test performance relationship), such data points are considered non-independent. There are two approaches in treatment of these data points.

First, I might treat each conceptual replication as yielding a different outcome value or effect size estimate. For example, to assess subgroup differences in cognitive ability test performance, researchers administered a battery of tests to test takers (e.g., quantitative, verbal, analytical, and/or spatial ability tests; Cotting, 2003; Nguyen, O'Neal, & Ryan, 2003) or a general intelligence test with multiple subtests (e.g., the WAIS; Pellegrini, 2005). Each cognitive ability measure could be treated as contributing a separate test performance value in this meta-analysis. According to Hunter and Schmidt (1990), measures of cognitive abilities are likely to be correlated with one another, which may lead to underestimating of variance across studies due to sampling error. The underestimation of variance in turn results in an over-estimate of the true variance of the effect size. Therefore, if the number of studies with multiple cognitive ability tests is large, a meta-analysis may be conservative. (Hunter and Schmidt also noted that a small number of studies using multiple cognitive ability tests in a database might result in little error in the resulting cumulation.) In this meta-analysis, there were eight reports that contained multiple measures of cognitive abilities (see Table 8). Second, to be sensitive to potential problems caused by non-independent data in this meta-analytic data set, I could use only one independent estimate of effect size per study, that is, an average of effect sizes across cognitive ability tests for all sub-samples per study. A limitation of this approach is that a meta-analysis may be liberal in generalizing conclusions.

Table 8

Studies with Multiple Measures of Cognitive Abilities

No.	Study	Cognitive ability tests	Non-independent data points? ^a
1	Cotting (2003)	1 math ability & 1 verbal ability test	Yes
2	Martin (2004)	2 general cognitive ability tests	No
3	Nguyen, O'Neal, & Ryan (2003)	1 math, 1 verbal & 1 analytical ability test	Yes
4	Pellegrini (2005)	3 WAIS subtests	Yes
5	Rivadeneyra (2001)	PSAT & SAT Math & Verbal tests	Yes
6	Stricker & Ward (2004), Study 2	2 math ability tests and 2 verbal ability tests	Yes
7	Wicherts, Dolan & Hessen (2005), Study 1	1 math, 1 verbal & 1 analytical ability test	Yes
8	Wicherts, Dolan & Hessen (2005), Study 3	4 math tests	Yes

Note. ^a Same subjects, multiple outcome values.

There are two single studies in Stricker and Ward (2004) with multiple subsamples; each sub-sample was very large in size and each sub-sample produced multiple cognitive ability test outcomes. If I adopted the first approach in treatment of cognitive ability test outcomes, there would be a substantial portion of non-independent data points in the data set. Therefore, I chose the second approach instead (i.e., using a composite effect size for each single study) to limit a violation of independent variance assumption.

Non-independent data points may also occur when the design of an experiment allows for multiple effect size estimates to be computed across treatment and comparison conditions. For example, in some studies of stereotype threat, the research design consists of one stereotype threat (treatment) condition and at least two or more stereotype threat-removed (comparison) conditions and vice versa, resulting in multiple treatment mean effect estimates. As mentioned above, these effect sizes could be treated as if they were independent. The limitation of such a treatment is that sampling error is under-estimated, which leads to an over-estimation of true variance in effect sizes. Therefore, the meta-analysis conclusion might be conservative about the generalizability of the overall effect size. Following Webb and Sheeran's (2006) procedure, I used the stereotype threat-removed group that produced the largest mean difference in mean cognitive ability test performance compared with the stereotype threat group (and vice versa). The limitation of this treatment is a liberal tendency in interpreting and generalizing conclusions. Table 9 lists the studies with multiple stereotype threat experimental conditions.

Table 9

Studies with Multiple Stereotype Threat Experimental Conditions

Study	Research Design	Selected pair
1 Cadinu, et al. (2003). Study 1 of 2 ^a - High Domain Identification - Low Domain Identification	3 (ST: STA v. STR1 v. STR2) x 2 (Domain Identification: High v. Low)	<i>High DI:</i> STA v. STR2 <i>Low DI:</i> STA v. STR1
2 Cohen & Garcia (2005). Study 2 of 3	3 (ST: STA v. STR1 v. STR2)	STA v. STR2
3 Dinella (2004)	2 (gender) x 3 (ST: STA 1 v STA 2 v. STR)	STA 2 v. STR
4 Gresky, Eyck, Lord, & McIntyre (unpublished) ^a - High Domain Identification - Low Domain Identification	2 (gender) x 2 (domain identification) x 3 (ST: STA v. STR1 v. STR2)	<i>High DI:</i> STA v. STR2 <i>Low DI:</i> STA v. STR1
5 Guajardo (2005)	2 gender x 5 ST (STA v. STR1 v. STR2 v. STR3 v. STR4)	STA v. STR2
6 Johns, Schmader, & Martens (2005)	2 (gender) x 3 (ST: STA v. STR1 v. STR2)	STA v. STR1
7 Martens, Johns, Greenberg, & Schimmel (2006). Study 1 of 2	2 (gender) x 3 (ST: STA v. STR1 v. STR2)	STA v. STR2
8 McIntyre, Lord, Gresky, Eyck, Frye, et al. (2005)	2 (gender) x 5 (ST: STA v. STR1 v. STR2 v. STR3 v. STR4)	STA v. STR4
9 Rivadeneyra (2001)	3 ST (STA1 v. STA2 v. STR)	STA1 v. STR
10 Steele & Aronson (1995). Study 1 of 4	2 (race) x 3 (ST: STA v. STR1 v. STR2)	STA v. STR1
11 Sternberg, et al. (unpublished) Study 1 of 2	2 (gender) x 4 (ST: STA v. STR1 v. STR2 v. control)	STA v. STR2
Study 2 of 2	2 (gender) x 4 (ST: STA v. STR1 v. STR2 v. control)	STA v. STR2

12	van Dijk, Koenders, Korenhof, Mulder, & Vries (unpublished) ^b Study 1a Study 1b	5 (ST: STA 1 v. STA 2 v. STA3 v. STR1 v. STR2)	<i>Study 1a:</i> STA1 v. STR1 <i>Study 1b:</i> STA2 v. STR2
13	Wout, Shih, Jackson, & Sellers (unpublished). Study 2 of 4	3 (ST: STA 1 v. STA 2 v. STR)	STA2 v. STR
14	Wout, et al. (unpublished). Study 3 of 4	3 (ST: STA 1 v. STA2 v. STR)	STA1 v. STR

Note. ST = Stereotype threat conditions. STA = Stereotype threat-activated group. STR = Stereotype threat-removed group. ^a These studies were split into two independent studies: One for high domain identified participants; the other for low domain identified participants (see also Table 10). ^b This study was split into two independent studies: Study 1a consisted of the (STA1 v. STR1) conditions; Study 1b consisted of the (STA2 v. STA3 v. STR2) conditions.

Treatment of Studies with a Control Condition

A related issue involves studies with a “control” condition (i.e., a cognitive ability test was presented to test takers without any special directions) in the stereotype threat design. Although this level of stereotype threat manipulation has been occasionally defined as a stereotype threat activated condition in some research reports, I defined it as a control condition.

Because all studies retained in the database include one stereotype threat-activated condition, each study contributes at least one estimate of effect size to the data set in this review. When a study design consists of two conditions of stereotype threat manipulation: activation and removal, the study contributed one effect size $d_{STA-STR}$ to the data set. When a study design consists of activation and control conditions, the study contributes one effect size $d_{STA-Control}$ to the data set. When all three levels of stereotype threat are present in one study, I would select the effect size $d_{STA-STR}$ to be cumulated toward an overall estimate of effect sizes across studies. Although this approach might result in an upward bias in finding and interpreting the magnitude of stereotype threat effects across studies (i.e., an estimate of $d_{STA-STR}$ might be larger than that of $d_{STA-Control}$), I decided to err on optimizing the probability of detecting stereotype threat effects and supporting the theory tenets, given the important social implications of stereotype threat.

Treatment of Studies with Stereotype Threat x Non-Target Moderator Design

For studies employing a design of “Stereotype Threat x Non-Target Moderator” (i.e., the moderating factor is not hypothesized and investigated in the present meta-analysis), I gathered relevant statistical information across the stereotype threat conditions only (from source reports or by contacting study authors directly; e.g., Brown

& Pinel, 2003). If such information was unavailable, I extracted stereotype threat statistical information from the “control” level of the non-target moderating factor (e.g., Ambady, et al., 2004; Marx & Stapel, in press; Marx, Stapel, & Muller, 2005, Study 4). Alternatively, I took the average of the cell mean effect sizes (as a function of Stereotype threat x Non-target Moderator). For example, in Josephs, et al. (2003), the researchers employed a 2 (stereotype threat) x 2 (gender) x 2 (testosterone) design. Although statistical information of test performance was reported for each cell of this design, no test performance information was reported for Stereotype threat x Gender cells only. Therefore, I took the average of mean effect sizes between high-testosterone and low-testosterone groups across gender and these averages contributed to the final dataset. (Sample sizes were also averaged.)

Treatment of Studies with Stereotype Threat x Target Moderator Design

For primary studies employing either the design of “Stereotype Threat x Domain Identification” or “Stereotype Threat x Test Difficulty,” I split these studies into two or three independent sub-samples according to the amount of domain identification levels or test difficulty levels defined by researchers themselves. Each sub-sample contributed an independent estimate of effect size to the database. Appendix D and Appendix E list these studies.

In the case of Anderson’s (2001) study, the estimates of effect size could be computed either for the full sample (Female $n = 604$; Male $n = 344$) or for the sub-samples of High-Domain identifiers (Female $n = 152$; Male $n = 160$) versus Low-Domain identifiers (Female $n = 302$; Male $n = 71$). In this dissertation review, the effect sizes based on the full sample were used to cumulate the overall effect size across studies,

whereas the effect sizes based on the high/low domain identification subsamples were cumulated in specific moderator analyses.

Treatment of Studies with “Stereotype Threat x Multiple Target Moderators” Design

Keller (in press) employed a factorial design involving both Domain Identification and Test Difficulty—the target conceptual moderators in this review. This study was coded as five separate sub-studies: two studies across levels of Domain Identification and three studies across levels of Test Difficulty. However, to avoid a violation of the independent error variance assumption, I cumulated only the estimates from Stereotype Threat x Test Difficulty studies for the overall mean effect size (because the Domain Identification subsets were nested within the subsets of Test Difficulty studies, based Hunter & Schmidt’s advice, 1990). Further, each set of sub-studies across moderator levels contributed the estimates to respective moderator analyses of Domain Identification or of Test Difficulty.

Treatment of Studies Where Gender Is Nested in Race

Schmader and Johns (2003, Study 2) and Stricker and Ward (2004, Study 2) conducted studies where test takers’ gender was nested within ethnicity (i.e., Latino vs. White, and Black vs. White, respectively). In these studies, subtle race-based stereotype cues were presented to activate stereotype threat among minority test takers (i.e., the tests measuring intelligence; a race inquiry prior to tests). These studies could have been coded separately by race and by gender as previously done (see Walton & Cohen’s coding scheme, 2003). However, because these studies conceptually aimed at assessing race-based stereotype threat effects, I decided that only the study outcome values as a function

of ethnicity and stereotype threat activation contributed data points to the overall meta-analytic data set.

Stricker and Ward (2004, Study 1) was an exception to this rule. Although the study design involved an interaction between race, gender and stereotype threat manipulation, the stereotype threat cues were both race-based and gender-based (i.e., race and gender inquiries prior to tests). Therefore, it was conceptually sound to code the outcomes of this study separately as a function of race or gender; that means, the study contributed some non-independent estimates of effect size to the data set. However, the proportion of these non-independent data points was not large in the data set (i.e., 842 data points altogether, or 10.7%).

Treatment of Studies with Large Sample Sizes

There are a few studies with substantially larger sample sizes than those in the majority of other studies in the meta-analysis (e.g., Anderson, 2001; Stricker & Ward, 2004, Study 1, Study 2). Because of the standard treatment of weighting of studies by sample size in meta-analytic procedure, these studies would receive a weight of multiple times more than the weight of the rest in the study sample. A common procedure for dealing with this issue is to cumulate and report meta-analytic results with and without the estimates of effect sizes from these studies (see Walton & Cohen, 2003).

However, in this meta-analysis, I chose *not* to follow this practice (i.e., I only reported findings including the estimates of effect size from large sample-size studies) because I believed that excluding large sample-size studies might decrease the credibility of meta-analytic findings. Based on Hunter and Schmidt's (1990) opposition to the practice of excluding "weak" design studies (hence, yielding non-significant findings

and/or being unpublished), I argued that a sample size is also a design aspect and should be consistently treated as such in meta-analytic exclusion criteria. Further, arbitrarily determining what constitutes a “large” sample size and excluding large sample size studies from meta-analyses may change the conclusions of the present review, because *inconsistent exclusion* of studies may take place across subsets. For instance, if large sample sizes were used as an exclusion criterion, a study sample size of 150 could be excluded in a meta-analysis with a subset of studies the sample sizes of which range from 20 to 50, but the same study might be included in a different meta-analysis where a subset of studies had more comparable sample sizes. Therefore, all studies were meta-analyzed wherever appropriate. (Nevertheless, for readers’ convenience, I reanalyzed and reported key meta-analytic findings with “sensitive” subsets or subsets without large sample-size studies and reported the results in Appendix H.)

Coding Studies

Three coders coded information in studies in the database. I was the first coder. The other two coders were undergraduate research assistants and seniors majoring in psychology, who had received “A” grades in prerequisite statistics and research design and measurement courses. The undergraduate coders received training on the theory of stereotype threat, the basic and complex stereotype threat research paradigms, and the procedural steps in coding studies for this meta-analysis (i.e., using a coding form and following a manual; see Appendices F & G). The undergraduate coders also had practice sessions where they coded between five and six studies independently, cross-checking with each other, and receiving my feedback, i.e., comparing their coding results with mine and discussing disagreement cases if any. Actual coding sessions were

subsequently held in a lab where two or three coders worked independently for two to three hours each session. The undergraduate coders were encouraged to take notes of anything that needed conceptual or methodological clarification while coding the relevant characteristics of studies. These notes were later used to discuss and resolve inter-rater disagreement in periodic meetings among coders.

The variables in all studies in the meta-analysis sample were coded by at least two coders and cross-checked. Specifically, I coded all studies. The other two coders each coded approximately 50% of studies in the database (randomly assigned).

Continuous variables consisted of statistics of interest (e.g., means test performance and standard deviations). The inter-rater agreement rates for these variables in any subset of coded studies were between 91% and 100% per pair of coders. Some of the disagreement in coding of continuous variables was caused by clerical errors; others were caused by inconsistencies in research reports. Pairs of coders discussed the disagreement cases, double-checked the content of a report, or, if necessary, directly contacted study authors for clarification of statistical information reported (e.g., Spicer, 1999; Walsh, et al., 1999)

For categorical variables in subsets of coded studies, I computed the inter-rater agreement index Kappa, following Landis and Koch's (1977) rules: a Kappa value of 0.8 or greater is very satisfactory; a Kappa value of 0.6 till 0.8 is good; a Kappa value of 0.4 till 0.6 is moderately good, and a Kappa value of less than 0.4 is considered poor agreement. For categorical variables in any subsets of coded studies, Kappa values ranged from 0.49 to 0.95, indicating moderately good to very satisfactory inter-rater agreement levels. The lower Kappa values were mainly associated with the coding task of

research design characteristics; specifically, the classification of treatment-comparison conditions (stereotype threat, stereotype threat-removed, and control conditions) because the coding scheme for these conditions might be different from how researchers define their own levels of stereotype threat manipulation, and the presentation mode of stereotype threat cues (e.g., across levels of explicitness of the stereotype threat manipulation). In other words, any lower agreement on these characteristics could be explained in part by coders' different expertise in research design, and in part by the inconsistency in how research conditions and/or treatment cues were operationally defined by researchers. Again, although no Kappa values were low enough to indicate poor agreement in this meta-analysis, all cases with disagreement were discussed to reach coders' consensus. (Note that my coding judgment calls were often but not always a final coding decision in cases with disagreement.)

Coding Form and Coding Manual

A data coding manual and a coding form were developed (see Appendix F and Appendix G). The coding manual includes a list of all relevant continuous and categorical variables, an operational definition for each variable (i.e., a brief explanation), and the respective category assignments. The coding form is identical to the coding manual minus variable definitions. When the information given in a particular study did not allow for a definite coding judgment, coders marked the data as missing.

Coding Statistics and Continuous Variables

Depending on the results reported in each particular study, objective statistics and continuous variables coded include sample size, variable cell means and standard deviations, *t*-test values, and/or *F*-test values. One coder recorded the information of

interest, and another coder cross-checked the information by comparing it with that in source papers at a later point. When statistical information was insufficient to compute the estimate of effect size for a study, coders tried to contact source authors for additional information before marking the data as missing.

Coding Descriptive Information

All coders coded the descriptive information of studies included in the sample of this meta-analysis, such as name(s) of author(s), year of publication, and publication status (published or unpublished). Study design was also described for reference. Another coder later cross-checked the information with source papers for accuracy.

Coding Study Characteristics

For each subset of the review database, all coders coded the relevant characteristics of studies, such as whether participants in a sample were pre-selected on domain identification, and whether a stereotype threat was primed subtly or explicitly, and if explicitly, whether it was done moderately or blatantly.

Coding Moderators

Methodological Moderators

There were several methodological moderators to be tested in the present study. For the *presentation modes of stereotype threat-activation cues*, I coded data on three levels: blatant, explicit and subtle (see Table 1 above for definitions and examples of these levels; also see Appendix A). A similar coding practice was used for the moderator of *presentation modes of stereotype threat-removing strategies* (see Table 2 and Appendix B). The “control” condition was reserved to code the condition where a

cognitive ability test without any special directions had been administered regardless of what this condition was labeled in original reports.

Although there were no formal hypotheses regarding categories of *group-based stereotypes*, as mentioned above, studies were also coded for demographic characteristics of samples, such as whether the stereotype activated was race-based (i.e., ethnic stereotyped samples and/or comparison samples) or gender-based (i.e., female stereotyped samples and/or male samples). Because stereotype threat manipulation and test takers' race/ethnicity or gender was correlated in many studies, a hierarchical moderator analysis was needed to assess the potential impact of confounding on the moderator analyses. To accomplish this, I would first break down the stereotype threat effect estimates for manipulation conditions by test takers' race/ethnicity or by gender, and then I would undertake a moderator analysis by race/ethnic samples of test takers (minorities vs. Whites) or by gender of test takers (women vs. men) within the stereotype threat manipulation conditions. Note that only studies that included both stereotyped and comparison groups contributed data to these hierarchical moderator analyses.

Conceptual Moderators

Domain identification. Levels of domain identification (high, medium high and low) can be coded from most stereotype threat studies. They were sub-samples or samples that had been prescreened on some criteria of (academic/mathematic) domain identification. Where no information about the construct was available, I marked the data as missing. See Appendix D for a list of studies that contributed data to this moderator analysis.

Note that a few studies in the data base which assessed test takers' domain identification tendency on a continuous scale, such as individuals' endorsement of items in a domain identification measure (e.g., Bailey, 2004; Edwards, 2004; Ployhart, et al., 2003). Unfortunately, I did not find sufficient statistical information in these reports to convert it into binary subgrouping information. For example, only mean domain identification is reported for the whole sample, not broken down by levels of domain identification. Therefore, coders also marked the data as missing in these cases.

Test difficulty. Stereotype threat theorists recommend that researchers should not define levels of test difficulty objectively. Spencer, et al. (1999) posit that the extent to which a test is judged difficult is not objective but contingent on target test takers' ability. In their study, Spencer, et al. lowered the degree of difficulty when testing math performance of a less math-able sample (compared with a highly math-able sample; Study 3). In other words, the level of test difficulty (difficult; moderately difficult; easy) is a sample-dependent issue in coding. Therefore, in the present meta-analysis, I coded levels of test difficulty based on researchers' self-report (e.g., describing a test as very difficult, moderately difficult, or easy) based on the assumption that researchers have the best knowledge of their sample's ability. When no such data was available, I marked the data as missing. See Appendix E for a list of studies that contributed data to this moderator analysis.

Summary of the Meta-Analytic Data Set

The literature search identified a total of 151 published and unpublished empirical reports on stereotype threat effects that could be potentially included in the review. Of these, 75 reports were excluded because they did not meet at least one or more inclusion

criteria for the meta-analysis (see Appendix C), whereas 76 reports were retained in the final database because they met the inclusion criteria.

The 76 reports contained 116 primary studies (three studies contributing non-independent data points; Stricker & Ward, 2004); 67 of which were from published peer-reviewed articles. Sixty-five of these primary studies included a comparison sample (e.g., Whites or men). The study database yielded a total of 8277 data points from stereotyped groups and a total of 6789 data points from comparison groups.

Please note that, because of my decisions in what estimate of effect size each study would contribute to subsequent meta-analyses (i.e., $d_{STA-STR}$ or $d_{STA-Control}$), the actual data set for an overall estimate of stereotype threat effects across studies may consist of the same number of primary studies but fewer data points than those reported here.

Reports that were included in this meta-analysis are preceded with an asterisk on the reference list.

Table 10 presents an overview of the characteristics of studies included in the full database.

Table 10

Overview of the Meta-Analysis Database: Characteristics of Included Studies (K = 116)

Study no.	Study	Publish status ^a	Stereotyped group	Sample size	Effect size	Comparison group? ^b	Do. Id. pre-selected? ^c
1	Ambady, Paik, Steele, Owen-Smith, & Mitchell (2004)	Published	Female undergrads	20	-.57	No	No
2	Ambady, Paik, Steele, Owen-Smith, & Mitchell (2004)	Published	Female undergrads	20	-.67	No	No
3	*Anderson (2001)	Unpublished	Female undergrads	604	-.96	Yes	No
4	Aronson, Lustina, Good, & Keough (1999)	Published	White undergrads	23	-1.46	No	Yes
5	Aronson, et al. (1999)	Published	White undergrads	26	-.99	No	Yes
6	Aronson, et al. (1999)	Published	White undergrads	23	-2.74	No	Yes
7	Bailey (2004)	Unpublished	Female undergrads	44	-.09	Yes	No
8	Brown & Day (in press)	Published	African American undergrads	34	.38	Yes	No
9	Brown & Josephs (1999)	Published	Female undergrads	65	-.09	Yes	No
10	Brown & Josephs (1999)	Published	Female undergrads	35	-1.17	Yes	No
11	Brown & Pintel (2003)	Published	Female undergrads	46	-.53	No	Yes
12	*Brown, Steele, & Atkins (unpublished)	Unpublished	African American undergrads	28	-.62	Yes	No
13	Cadinu, et al. (2003)	Published	Female undergrads (Italian)	25	-.10	No	Yes
14	Cadinu, et al. (2003)	Published	Female undergrads (Italian)	38	.023	No	Yes
15	Cadinu, et al. (2003)	Published	African American soldiers	50	-.19	No	No
16	Cadinu, Maass, Rosabianca, & Kiesner (2005)	Published	Female undergrads (Italian)	60	.015	No	No
17	Cohen & Garcia (2005)	Published	African American undergrads	41	.74	No	No
18	Cotting (2003)	Unpublished	Female undergrads	51	-.53	No	No
19	Cotting (2003)	Unpublished	African American undergrads	55	.44	No	No
20	*Davies, Spencer, Quinn, &	Published	Female undergrads	25	-.71	Yes	Yes

21	Gerhardstein (2002)	2 of 2	Published	Female undergrads	34	.27	Yes	Yes
22	Davies, Spencer, Quinn, & Gerhardstein (2002)	1 of 1	Unpublished	Female high school students	232	.11	Yes	No
23	Dodge, Williams, & Blanton (2001)	1 of 1	Unpublished	African American undergrads	93	.045	Yes	No
24	Edwards (2004)	1 of 1	Unpublished	Female undergrads and graduates	79	-.78	No	No
25	Elizaga & Markman (unpublished)	1 of 1	Unpublished	Female undergrads	145	-.38	No	No
26	Foels (1998)	1 of 1	Unpublished	Female undergrads	33	-.78	No	No
27	Foels (1998)	1b of 1	Unpublished	Female undergrads	32	-.3	No	No
28	*Foels (2000)	1 of 1	Unpublished	Female undergrads	71	-.7	Yes	No
29	Ford, Ferguson, Brooks, & Hagadone (2004)	2 of 2	Published	Female undergrads	31	-1.7	No	No
30	Gamet (2004)	1 of 1	Unpublished	Female undergrads	51	-1.51	No	No
31	Gresky, Eyck, Lord, & McIntyre (unpublished)	1 of 1	Unpublished	Female undergrads	23	-.32	Yes	Yes
32	Gresky, Eyck, Lord, & McIntyre (unpublished)	1b of 1	Unpublished	Female undergrads	37	-.45	Yes	Yes
33	Guajardo (2005)	1 of 2	Unpublished	Female undergrads	56	.03	Yes	No
34	Guajardo (2005)	2 of 2	Unpublished	Female undergrads	30	-.52	Yes	No
35	Harder (1999)	1 of 2 (pilot)	Unpublished	Female undergrads	36	-.66	Yes	No
36	Harder (1999)	2 of 2	Unpublished	Female undergrads	19	-.04	No	Yes
37	Johns, Schmader, & Martens (2005)	1 of 1	Published	Female undergrads	46	.27	Yes	No
38	*Josephs, Newman, Brown, & Beer (2003)	1 of 1	Published	Female undergrads	39	-.79	Yes	Yes
39	Keller & Bless (unpublished)	2 of 3	Unpublished	Female undergrads (German)	66	-.57	No	No
40	*Keller & Dauenheimer (2003)	1 of 1	Published	Female secondary school students (German)	33	-.38	Yes	No
41	*Keller (2002)	1 of 1	Published	Female secondary school students (German)	37	-.62	Yes	No
42	Keller (in press)	1 of 1	Published	Female secondary school students (German)	19	-.10	Yes	No
43	Keller (in press)	1b of 1	Published	Female secondary school students (German)	18	-.95	Yes	No

44	Keller (in press)	1c of 1	Published	students (German) Female secondary school students (German)	18	-1.11	Yes	No
45	Lewis (1998)	1 of 1	Unpublished	African American undergrads	71	-.12	Yes	No
46	Martens, Johns, Greenberg, & Schimel (2006)	1 of 2	Published	Female undergrads	22	-.67	No	Yes
47	*Martens, Johns, Greenberg, & Schimel (2006)	2 of 2	Published	Female undergrads	38	-.11	Yes	No
48	Martin (2004)	2 of 2	Unpublished	African American undergrads	100	.54	No	No
49	Martin (2004)	2b of 2	Unpublished	African American undergrads	102	-.93	No	No
50	Marx & Stapel (2005)	1 of 1	Published	Female undergrads (Dutch)	48	-1.07	Yes	No
51	Marx & Stapel (in press)	1 of 3	Published	Female undergrads (Dutch)	29	-1.22	Yes	No
52	Marx & Stapel (in press)	3 of 3	Published	Female undergrads (Dutch)	28	-1.24	Yes	No
53	Marx, Stapel, & Muller (2005)	3 of 4	Published	Female undergrads (Dutch)	27	.56	No	No
54	Marx, Stapel, & Muller (2005)	3b of 4	Published	Female undergrads (Dutch)	25	-.16	No	No
55	Marx, Stapel, & Muller (2005)	4 of 4	Published	Female undergrads (Dutch)	25	-.14	No	No
56	McFarland, Kemp, Viera, & Odin (2003)	1 of 1	Unpublished	Female undergrads	126	-.035	No	No
57	McFarland, Lev-Arey, & Ziegert (2003)	1 of 1	Published	African American undergrads	50	-.22	Yes	No
58	McIntyre, Lord, Gresky, Eyck, Frye, et al. (2005)	1 of 1	Published	Female undergrads	81	-.98	Yes	No
59	McIntyre, Paulson, & Lord (2003)	1 of 2	Published	Female undergrads	116	-.52	Yes	No
60	McIntyre, Paulson, & Lord (2003)	2 of 2	Published	Female undergrads	74	-.49	Yes	No
61	*McKay (1999)	1 of 1	Unpublished	African American undergrads	103	.91	Yes	No
62	Nguyen, O'Neal, & Ryan (2003)	1 of 1	Published	African American undergrads	80	.05	Yes	No

63	Nguyen, Shivpuri, Ryan & Langset (2004)	1 of 1	Unpublished	Female undergrads	114	.057	Yes	No
64	O'Brien & Crandall (2003)	1 of 1	Published	Female undergrads	58	-.305	Yes	No
65	Oswald & Harvey (2000/2001)	1 of 1	Published	Female undergrads	34	-.06	No	No
66	Pellegrini (2005)	1 of 1	Unpublished	Hispanic undergrads (female)	60	-1.03	No	No
67	Philipp & Harton (2004)	1 of 1	Unpublished	Female undergrads	38	-1.21	Yes	No
68	Ployhart, Ziegert & McFarland (2003)	1 of 1	Published	African American undergrads	48	-.59	Yes	No
69	Ployhart, et al. (2003)	1b of 1	Published	African American undergrads	48	-.57	Yes	No
70	Prather (2005)	1 of 1	Unpublished	Female undergrads	114	-.67	No	No
71	Rivadeneyra (2001)	1 of 1	Unpublished	Latino high school students	116	-.96	No	No
72	Rosenthal & Crisp (2006)	2 of 3	Published	Female undergrads (British)	24	-1.46	No	No
73	Rosenthal & Crisp (2006)	3 of 3	Published	Female undergrads (British)	29	-.99	No	No
74	Rosenthal & Crisp (2006)	3b of 3	Published	Female undergrads (British)	27	-2.74	No	No
75	*Salinas (1998)	1 of 2	Unpublished	Mexican American undergrads	27	-.09	Yes	No
76	*Salinas (1998)	2 of 2	Unpublished	Mexican American undergrads	56	.38	Yes	No
77	Sawyer & Hollis-Sawyer (2005)	1 of 1	Published	African American undergrads	66	-.09	Yes	No
78	Sawyer & Hollis-Sawyer (2005)	1b of 1	Published	Hispanic undergrads	47	-1.17	Yes	No
79	Schimmel, Arndt, Banko, & Cook (2004)	2 of 3	Published	Female undergrads	46	-.53	No	Yes
80	Schmader & Johns (2003)	1 of 3	Published	Female undergrads	28	-.62	Yes	Yes
81	Schmader & Johns (2003)	2 of 3	Published	Latino American undergrads	33	-.10	Yes	No
82	Schmader & Johns (2003)	3 of 3	Published	Female undergrads	28	.023	No	Yes
83	*Schmader (2002)	1 of 1	Published	Female undergrads	32	-.19	Yes	No
84	Schmader, Johns & Barquissau (2004)	2 of 2	Published	Female undergrads	68	.015	No	No
85	Schneeberger & Williams (2003)	1 of 1	Unpublished	Female undergrads	61	.74	Yes	No

86	*Schultz, Baker, Herrera, & Khazian (unpublished)	1 of 2	Unpublished	Hispanic American undergrads	44	-.533	Yes	No
87	*Schultz, Baker, Herrera, & Khazian (unpublished)	2 of 2	Unpublished	Hispanic American undergrads	40	.44	Yes	No
88	Seagal (2001)	6 of 6	Unpublished	African American and Latino undergrads	101	-.71	Yes	No
89	*Sekaquaptewa & Thompson (2002)	1 of 1	Published	Female undergrads	80	.27	Yes	No
90	Smith & Hopkins (2004)	1 of 1	Published	African American undergrads	160	.11	No	No
91	Smith & White (2002)	1 of 2	Published	White undergrads (male)	47	.045	No	No
92	Smith & White (2002)	2 of 2	Published	Female undergrads	23	-.78	No	No
93	Spencer (2005)	1 of 1	Unpublished	Female undergrads	40	-.38	No	No
94	*Spencer, Steele, & Quinn (1999)	2 of 3	Published	Female undergrads	30	-.78	Yes	Yes
95	Spicer (1999)	2 of 2	Unpublished	African American undergrads	39	-.3	No	Yes
96	Spicer (1999)	2b of 2	Unpublished	African American undergrads	39	-.7	No	Yes
97	*Steele & Aronson (1995)	1 of 4	Published	African American undergrads	38	-1.7	Yes	No
98	*Steele & Aronson (1995)	2 of 4	Published	African American undergrads	20	-1.51	Yes	No
99	*Steele & Aronson (1995)	4 of 4	Published	African American undergrads	22	-.32	Yes	No
100	*Sternberg, Jarvin, Leighton, Newman et al. (unpublished)	1 of 2	Unpublished	Female high school students	27	-.45	Yes	No
101	*Sternberg, Jarvin, Leighton, Newman et al. (unpublished)	2 of 2	Unpublished	Female high school students	96	.03	Yes	No
102	*Stricker & Ward (2004)	1 of 2	Published	African American high school students	122	-.52	Yes	No
103	(*) Stricker & Ward (2004)	1b of 2	Published	Female high school students	730	-.66	Yes	No
104	*Stricker & Ward (2004)	2 of 2	Published	African American undergrads	468	-.04	Yes	No
105	Tagler (2003)	1 of 1	Unpublished	Female undergrads	136	.27	Yes	No
106	van Dijk, Koenders, Korenhof, Mulder, & Vries (unpublished)	1 of 1	Unpublished	Female undergrads (Dutch)	38	-.79	Yes	No
107	van Dijk, Koenders, Korenhof,	1b of 1	Unpublished	Female undergrads	38	-.57	Yes	No

108	Mulder, & Vries (unpublished) von Hippel, von Hippel, Conway, Preacher et al. (2005)	4 of 4	Published	(Dutch) White undergrads (Australian)	56	-.38	No	No
109	Walsh, Hickey, & Duffy (1999)	2 of 2	Published	Female undergrads (Canadian)	96	-.62	Yes	No
110	Walters (2000)	1 of 2	Unpublished	African American undergrads	49	-.10	No	Yes
111	Wicherts, Dolan, & Hessen (2005)	1 of 3	Published	Minority high school students (Dutch)	138	-.95	Yes	No
112	Wicherts, et al. (2005)	3 of 3	Published	Female undergrads (Dutch)	95	-1.11	Yes	No
113	Wout, Shih, Jackson, & Sellers (unpublished)	1 of 4	Unpublished	African American undergrads	57	-.12	No	No
114	Wout, Shih, Jackson, & Sellers (unpublished)	2 of 4	Unpublished	African American undergrads	29	-.67	No	No
115	Wout, Shih, Jackson, & Sellers (unpublished)	3 of 4	Unpublished	African American undergrads	24	-.11	No	No
116	Wout, Shih, Jackson, & Sellers (unpublished)	4 of 4	Unpublished	African American undergrads	26	.54	No	No

Note. Studies with an asterisk ($k = 23$; $n = 2810$) overlap with studies in Walton and Cohen's (2003) meta-analysis of stereotype threat effects ($k = 43$). There is an exception: in the case of two studies in Stricker and Ward (2003), Walton and Cohen meta-analyzed data in both studies by race and by gender, whereas the data in Study 2 were analyzed by race in the present meta-analysis.

^a *Status*: "Published" refers to peer-reviewed journal articles including in-press ones; "unpublished" refers to dissertations, theses, conference papers, and working manuscripts.

^b Whether a comparison group was included in the study design.

^c Whether participants from a stereotyped group were pre-selected based on a domain identification criterion(a).

Meta-Analytic Procedure

Meta-analysis is a rigorous alternative to the traditional review process because it involves statistical integration of results. The basis of this methodology for experimental studies is the effect size, a standardized statistic that quantifies the magnitude of an effect.

In the present review, I employed the meta-analysis procedure by Hunter and Schmidt (1990) and conducted an overall meta-analysis to cumulate the findings from all independent samples, as well as separate meta-analyses to examine moderator effects.

Correction for Measurement Unreliability

Hunter and Schmidt (1990) recommend that effect sizes should be corrected by criterion measurement unreliability by dividing each observed effect size by the square root of the reliability of the dependent variable measure(s) (i.e., cognitive ability tests). However, as mentioned above the reliability information on cognitive ability tests used to assess stereotype threat effects on target test takers' performance was sporadically reported in source reports (see Table 7 above). I had put forth an effort to locate the reliability estimates of several established cognitive ability measures in the broad testing literature. However, I could only identify the reliability estimates of tests employed in 20 primary studies (about 17 percent of the data set; see Table 7 above). Therefore, study effect sizes could not be corrected individually for measurement error. In these cases, Hunter and Schmidt's advice is to use artifact distributions for meta-analyses. Note that most of the reported or identified reliability coefficients of cognitive ability tests in the data set tended to range from satisfactory to excellent. That means, the correction for artifact distributions may not account for much variance across studies. An implication is that the meta-analysis may provide liberal estimates (that is, upper-bound estimates) of

the reliability of the majority of tests used in this database. However, given that most of the measurement instruments of cognitive abilities were adapted and/or modified from standardized tests such as GRE, SAT, and GMAT, there is a possibility that the unreported reliability of tests in stereotype threat studies is not substantially poorer than the estimates used for artifact distributions in the present review. Further, Hunter and Schmidt recommend using a uniform artifact distribution in meta-analyzing subsets of d -values instead of adjusting artifact distributions for each subset.

Computing Effect Size

I used the effect size (Cohen's d) which corresponds to the mean difference between cell means in standard score form (i.e., the ratio of the difference between the means to the pooled within-group standard deviation). Based on an extensive survey of statistics reported in the literature in the social sciences, Cohen (1988) operationally defines standardized effect sizes as “small, $d = 0.20$,” “medium, $d = 0.50$,” and “large, $d = 0.80$ ” (p. 25). Cohen (2002) explains that “medium ES (effect sizes) represent an effect of a size likely to be apparent to the naked eye of a careful observer, that small ES be noticeably smaller yet not trivial, and that large ES be the same distance above medium as small is below it.” Although the guidelines do not take into account specific research contexts, they do correspond to the distribution of effects across meta-analyses found by Lipsey and Wilson (1993). Therefore, readers might want to use these suggested guidelines to evaluate cross-study stereotype threat effects in the present study.

If descriptive statistics (i.e., sub-sample sizes, cell means and standard deviations) were not reported in a study, I applied transformation formulas to compute d values from other statistical information available, such as t -test estimates, or F -test estimates.

To calculate effect sizes, I used Thalheimer and Cook's (2002) Excel-based program called "Calculating effect sizes from published research: A simplified spreadsheet" (updated in 2003). This program consists of built-in formulas of effect size conversion in Rosnow and Rosenthal's (1996) and Rosnow, Rosenthal, and Rubin's (2000) articles. The program allows the calculation of effect sizes from descriptive statistics (means, standard deviations or standard errors, and sample sizes), t -test values (with or without standard deviation or standard error estimates) and F -test values (with or without Mean Squared Error values).

Meta-Analytic Computations and Moderator Analyses

Following Hunter and Schmidt (1990), I cumulated the average population effect size δ (corrected for measurement error) and computed variance $\text{var}(\delta)$ across studies, weighted by sample size.

I categorized data into moderating categories and conducted a separate meta-analysis for each category. Meta-analytic evidence for moderator effects is established when true estimates are different across moderator categories and when the mean corrected standard deviation within categories is smaller than the corrected standard deviation computed for combined categories (see, for example, Judge, Colbert, & Ilies, 2004).

To judge whether substantial variation due to moderators exists, I use the standard deviation SD_{δ} estimated from $\text{var}(\delta)$ to construct the 90% credibility intervals around δ as an index of true variance due to moderators (Whitener, 1990). When the credibility intervals are large (e.g., greater than 0.11; Koslowsky & Sagie, 1993) and overlap zero (0), these intervals suggest the presence of moderators (Hunter & Schmidt, 1990),

provided that the “rule of thumb” test for moderator variables is also met. Specifically, $V\%$ or the ratio of sampling error variance to the observed variance in the corrected effect size is calculated. According to Hunter and Schmidt (1990), if $V\%$ is equal to or greater than 75%, most of the observed variance is due to sampling error; therefore, it is less likely that a true moderator exists and explains the observed variance in effect sizes. This method is able to detect the existence of unsuspected moderators and is advantageous “when the number of studies was small (4, 8, 16, 32, or 64) and the sample size of each study was small (50 or 100)” (p. 415). Hunter and Schmidt (1990) also recommend another approach, binary subgrouping, to testing for conceptually hypothesized moderators as those in the present review.

Software program. I used the “Hunter-Schmidt Meta-Analytic Programs Package, v1.1” (Schmidt & Le, 2004) to cumulate data across studies. This software package includes programs that implement all basic types of Hunter-Schmidt psychometric meta-analysis methods. For the present review, I utilized the “Meta-analysis of d -values using artifact distributions” program (Meta-analysis Type 4) because information on the reliability of cognitive ability tests was only sporadically reported in most of the primary studies in the database.

Testing for Publication Bias

Rosenthal (1979) describes a potential threat to the validity of a meta-analysis: publication bias or the file-drawer problem. This problem may lead to an overestimation of effect sizes because non-significant findings tend to be attributed to design artifacts by peer-reviewers and, thus, less likely to get published than studies with significant findings. As mentioned above, to resolve this problem, I included all studies that satisfy

the inclusion criteria regardless of publication status (implying an inclusion of null results).

The meta-analysis database consists of a relatively balanced number of published and unpublished reports (54.8% and 45.2%, respectively). Nevertheless, “fail-safe N ” analyses were conducted to test a potential file-drawer bias in each meta-analysis. That is to say, the actual mean effect size may not be as high as the one generated by meta-analysis, due to the fact any studies reporting weak or zero effect sizes are less likely to be published, and thus less likely to be included in the meta-analysis. Hunter and Schmidt (1990) provide a formula to calculate Fail-Safe N which indicates the number of missing studies with zero effect size that would have to exist to bring the mean effect size down to a specific level. In the present review, mean $d_{critical}$ is arbitrarily set to 0.10, which constitute a negligible effect size (see Cohen, 1988). According to Hunter and Schmidt, the lower the value of the “Fail-Safe N ,” the more uncertainty exists regarding whether the mean effect size may be biased by the number of file-drawer studies due to weak, non-existent, or contrary effects.

Additionally, I tested for the presence of publication bias by using Light and Pillemer’s (1984) “funnel graph” technique, which means simply plotting sample size versus effect size. The funnel graph is based on the fact that the precision in estimating stereotype threat effects increases as the sample size of included studies increases. Results from small sample size studies would scatter widely at the bottom of the graph. The spread would narrow as precision increases among larger sample size studies.

In the absence of bias, the plot should resemble a symmetrical inverted funnel. A problem of publication bias is graphically demonstrated if there is a cutoff of small

effects for studies with a small sample size. In other words, because only large effects reach statistical significance in small samples, a publication bias or other types of location biases are present when only large effects are reported by studies with a small sample size (i.e., an asymmetrical and skewed shape). On the contrary, there are no biases if an exclusion of null results is not visible on the funnel graph.

Summary

This section described the meta-analytic steps of literature search, the process of setting inclusion criteria and selecting studies based on these criteria, the coding scheme of relevant variables, the coding process and judgment calls that coders made in this review, approaches of treatment of meta-analytic data, and the meta-analytic procedure to investigate the research questions and hypotheses of interest. Potential biases in the data set and procedures to detect them, including conducting fail-safe N analyses (Hunter & Schmidt's 1990 formula) and graphing an overall funnel plot (Light & Pillemer, 1984) were described. In the next chapter, the meta-analytic results of interest are presented.

Chapter 3

RESULTS

Within-Group Meta-Analytic Findings

Overall Within-Group Stereotype Threat Effects

Hypothesis 1 predicts that a stereotype threat manipulation negatively affects stereotyped test takers' cognitive ability performance across studies. For stereotyped test takers' performance, effect sizes were computed as Cohen's d where a positive effect size would represent stereotyped individuals' overperformance on cognitive ability tests and a negative effect size represents underperformance due to the activation of stereotype threat as the theory predicts.

As shown in the top column of Table 11, the mean effect size (d) and mean true effect size (δ) for the total set of $K = 116$ effect size values and a total sample size $N = 7964$ are $|.26|$ and $|.28|$ respectively, which are comparable to the finding of mean $d = .29$ in Walton and Cohen (2003). The observed variance is $.23$ and the true variance is $.20$. The study artifacts of sampling error in cognitive ability tests explain about 26 percent of the cross-study variance of observed d -values. In other words, after correction for sampling error, the true effect size (δ) slightly increases, and the true variance became slightly smaller than the observed variance.

Table 11

Hierarchical Meta-Analytical Findings (Within-Group) ^a

<u>Overall Findings ^b</u>			
	<i>K</i>		116
	<i>N</i>		7964
	Mean <i>d</i>		-.258
	Var <i>d</i>		.227
	Var <i>e</i>		.06
	Mean δ		-.281
	Var δ		.198
	% var SE		26.26
	% var acc. for (V%)		26.33
	90% CI		(-.85) - (.29)
	Fail safe <i>N</i> ^c		415

<u>Race/ethnicity Stereotype Subset Findings ^d</u>		<u>Female Stereotype Subset Findings</u>	
<i>K</i>	44	<i>K</i>	72
<i>N</i>	2988	<i>N</i>	4935
Mean <i>d</i>	-.324	Mean <i>d</i>	-.208
Var <i>d</i>	.186	Var <i>d</i>	.241
Var <i>e</i>	.060	Var <i>e</i>	.059
Mean δ	-.353	Mean δ	-.227
Var δ	.149	Var δ	.216
% var SE	32.50	% var SE	24.64
% var acc. for (V%)	32.63	% var acc. for (V%)	24.68
90% CI	(-.85) - (.14)	90% CI	(-.82) - (.37)
Fail safe <i>N</i>	187	Fail safe <i>N</i>	222

Note. *K* = Number of effect sizes (*d*-values). *N* = Total sample size. Mean (*d*) = Sample size weighted mean effect size. Var (*d*) = Sample size weighted observed variance of *d*-values. Var (*e*) = Variance attributed to sampling error variance. Mean (δ) = Mean true effect size. Var (δ) = True variance of effect sizes. % var SE = Percent variance in observed *d*-values due to sampling error variance. % var acc. for (V%) = Percent variance in observed *d*-values due to all corrected artifacts. 90% CI = 90 percentile of (δ) (credibility interval).

^a I defined the dependent measure (cognitive ability test performance) more narrowly and rigorously than Walton and Cohen (2003). Therefore, the review findings do not strictly replicate those in Walton and Cohen's study.

^b The observed *d*-values were yielded from mean test score comparisons between the stereotype threat activated condition and a comparison condition.

^c Hunter and Schmidt's (1990) effect size file drawer analysis or fail safe *N*: number of missing studies averaging null findings needed to bring Mean (*d*) down to .10.

^d If five primary studies that had used White test takers as the stereotyped group were excluded from this subset (Aronson, et al., 1999; Smith & White, 2002; von Hippel, et al., 2005) so that only ethnic minority-based studies were meta-analyzed, mean *d* would be slightly smaller (-.32) than the above value, whereas var *d* (.18) and var *e* (.057) were comparable to the above values.

However, the variance of effect sizes is non-zero ($V\%$ is about 26 percent); the credibility interval shows that there is a 90% probability that the true effect size is between $(-.85)$ and $(.29)$ —a range of d -values overlapping zero. These values indicate that true moderators exist, and that the interpretation of an overall mean effect size without considering any moderator effects might be misleading. In other words, Hypothesis 1 was not supported due to the inconclusive meta-analytic finding.

A fail-safe N analysis addresses the potential file drawer problem, that is, how certain conclusions would be about a mean effect size if studies that do not find significant effects had been published or included in the data set. If a lower value of fail-safe N is found, there is more uncertainty (see Hunter & Schmidt, 1990). The last line on the top column in the top row of Table 11 lists 415 as the value of nonsignificant studies that would be necessary to reduce the effect size to a nonsignificant value, defined in this meta-analysis as a mean effect size of $d_{critical} = .10$. Because this is a very high fail-safe N value, it is unlikely that there are more than 400 “file drawer” studies of stereotype threat effects on cognitive ability test performance existing to reduce the meta-analyzed overall mean effect size to $.10$.

Moderator Analysis: Group-Based Stereotypes

The data set in the present review consists of two methodologically different subsets of studies: (a) studies that manipulate an ethnic/racial group-based stereotype of intellectual inferiority, and (b) studies that manipulate a gender-based stereotype of mathematical ability inferiority. Therefore, before proceeding with testing the conceptualized hypotheses of moderators, I logically examined whether these two subsets produced equivalent mean stereotype threat effect sizes. If there were such an

equivalency, the group-based stereotypes would not mitigate the observed variance in the overall effect size.

The columns in the bottom row of Table 11 shows that the activation of an ethnic/racial group-based intellectual stereotype (salient to minority test-takers or some Whites) and the activation of a gender-based math ability stereotype (salient to women) produce *differential* stereotype threat effects at least at the mean level. The mean effect size is larger by .11 in the ethnicity/race-based stereotype subset (mean $d = .32$, var $d = .19$, $k = 44$, $n = 2988$) than that in the gender-based stereotype subset (mean $d = .21$, var $d = .24$, $k = 72$, $n = 4935$). After correction for sampling error, the true mean effect sizes for the race-based stereotype group and gender-based stereotype group are slightly larger (mean $\delta = .35$ and $.23$, respectively) whereas true variance is slightly smaller for both subsets. The difference between the standardized mean effect sizes of these two subsets was $.12$.

Although the subset variance values are reduced compared with the variance of the entire set of d -values, they are still non-zero. There is a 90% probability that the true mean effect size of the race-based subset is between a wide range of $(-.85)$ and $(.14)$ which overlaps zero, whereas there is a 90% probability that the true mean effect size of the gender-based subset is between a wide range of $(-.82)$ and $(.37)$ which also overlaps zero. (Subset $V\%$ values are 33 percent and 25 percent for the race-based subset and the gender-based subset, respectively.) Taken together, these values suggested that one should not interpret these mean effect size findings before conducting further moderator meta-analyses. Large fail-safe Ns (on the last line of each column in Table 11) indicate that the file-drawer problem is unlikely for these subsets.

The following sections present the meta-analytic findings of the hypothesized moderator analyses for within-group effects of stereotype-threat activating cues, stereotype threat-removing strategies, test takers' domain identification, and test difficulty.

Moderator Analysis: Stereotype Threat-Activating Cues

Hypothesis 2 predicted that, in a stereotype threat-activated condition, the presentation mode of stereotype threat cues moderates stereotype threat mean effect sizes: studies using blatant cues would produce the smallest mean effect size compared with those using explicit cues or subtle cues, whereas studies using explicit cues would yield the largest effect size. Table 12 presents the meta-analytic findings of interest.

Hypothesis 2 was not supported at the threat-activating cue level. As shown in the columns in the bottom row of Table 12, all stereotype-threat activating cues produced comparable mean effect sizes for stereotyped groups (mean d s = $|.29|$, $|.25|$, and $|.25|$ for blatant, explicit, and subtle cues, respectively). In addition, these estimates do not substantially differ from one another and from the overall cross-study mean effect size (mean $d = |.26|$; see the top row of Table 12). Further, the variance of the entire set of d -values is non-zero. There is a 90% probability that the true mean effect size of the race-based subset lies in a wide range of $(-.85)$ and $(.29)$, overlapping zero. (V% values are between 22 and 44 percent.) The information suggests that there are true moderators that explain the variance of these d -values and that one should not interpret the mean effect sizes detected. The large fail-safe N values indicate that the file-drawer problem is unlikely for these subsets. Therefore, additional hierarchical meta-analyses were conducted across levels of stereotype-threat activating cues and group-based stereotypes.

Table 12

Hierarchical Moderator Analyses of Stereotype Threat Activating Cues

<u>Overall (ST-Activating Cues)^a</u>	
<i>K</i>	116
<i>N</i>	7964
Mean <i>d</i>	-.258
Var <i>d</i>	.227
Var <i>e</i>	.06
Mean δ	-.281
Var δ	.198
% var SE	26.26
% var acc. for (V%)	26.33
90% CI	(-.85) - (.29)
Fail safe <i>N</i> ^c	415

	<u>STA Blatant^b</u>	<u>STA Explicit^b</u>	<u>STA Subtle^b</u>
<i>K</i>	33	27	56
<i>N</i>	1930	1432	4603
Mean <i>d</i>	-.289	-.253	-.246
Var <i>d</i>	.312	.177	.204
Var <i>e</i>	.07	.077	.05
Mean δ	-.315	-.275	-.268
Var δ	.292	.119	.183
% var SE	22.14	43.44	24.36
% var acc. for (V%)	22.2	43.52	24.43
90% CI	(-1.0)-(.38)	(-.72)-(.17)	(-.82) - (.28)
Fail safe <i>N</i>	128	95	194

Note. *K* = Number of effect sizes (*d*-values). *N* = Total sample size. Mean (*d*) = Sample size weighted mean effect size. Var (*d*) = Sample size weighted observed variance of *d*-values. Var (*e*) = Variance attributed to sampling error variance. Mean (δ) = Mean true effect size. Var (δ) = True variance of effect sizes. % var SE = Percent variance in observed *d*-values due to sampling error variance. % var acc. for (V%) = Percent variance in observed *d*-values due to all corrected artifacts. 90% CI = 90 percentile of (δ) (credibility interval).

^a The observed *d*-values were yielded from mean test score comparisons between the stereotype threat activated condition and a comparison condition.

^b Levels of cues that may activate stereotype threat: "Blatant" = blatantly explicit cues about group differences in cognitive ability test performance (specified direction). "Explicit" = explicit cues about possible group differences in cognitive ability test performance (unspecified direction). "Subtle" = subtle cues that may prime stereotype threat without directly referring to group ability differences.

^c Hunter and Schmidt's (1990) effect size file drawer analysis or fail safe *N*: number of missing studies averaging null findings needed to bring Mean (*d*) down to .10.

Table 13 presents the findings for the subset of race-based stereotype (targeting minorities) and the subset of gender-based stereotype (targeting female test takers) across stereotype threat-activating cues. The columns in the top row of Table 13 show that stereotype threat-activating cues differentially affect minority test takers (mean $d = |.30|$; var $d = .19$; $k = 38$, $n = 2724$) and women test takers' cognitive ability test performance (mean $d = |.21|$; var $d = .24$; $k = 73$, $n = 4947$). The effect was slightly greater for minorities than for women.

The left columns in the bottom row in Table 13 show that when the negative stereotype is race-based (i.e., invoking minorities' fear of confirming social stereotype about group intellectual inferiority), explicit stereotype threat-activating cues yield the largest mean effect size mean $d = |.64|$ (var $d = .07$, $k = 7$, $n = 277$) as compared with other types of threat-activating cues (blatant cues: mean $d = |.41|$, var $d = .08$, $k = 6$, $n = 436$; subtle cues: mean $d = |.22|$, var $d = .20$; $k = 25$, $n = 2011$).

For blatant and explicit cue subsets, study artifacts explain most or all of the variance in d -values (large $V\%$ values). The 90% credibility interval for blatant cues does not overlap zero, indicating that no other moderators explain the variation in d values for blatant cues of stereotype threat were activated. In other words, the findings of mean effect size estimates for the race-based subset in these stereotype threat-activating cue conditions are conclusive. However, for the "subtle cues" subset, the small $V\%$ and the overlapping-zero 90% credibility interval indicate that further moderator analyses are still needed and the mean effect size result is not conclusive.

Table 13

Hierarchical Meta-Analytic Findings: Stereotype Threat-Activating Cues by Group-Based Stereotypes

<u>Minority test takers - Overall (ST-Activating Cues)</u>					<u>Women test takers - Overall (ST-Activating Cues)</u>				
<i>K</i>	38				<i>K</i>	73			
<i>N</i>	2724				<i>N</i>	4947			
Mean <i>d</i>	-.295				Mean <i>d</i>	-.205			
Var <i>d</i>	.185				Var <i>d</i>	.24			
Var <i>e</i>	.057				Var <i>e</i>	.06			
Mean δ	-.322				Mean δ	-.223			
Var δ	.151				Var δ	.214			
% var SE	30.89				% var SE	25.06			
% var acc. for (V%)	31				% var acc. for (V%)	25.1			
90% CI	(-.82) - (.18)				90% CI	(-.82) - (.37)			
Fail safe <i>N</i>	150				Fail safe <i>N</i>	223			

<u>Minority test takers - Overall (ST-Activating Cues)</u>					<u>Women test takers - Overall (ST-Activating Cues)</u>				
<i>K</i>	6	STA Blatant	STA Explicit	STA Subtle	<i>K</i>	22	STA Blatant	STA Explicit	STA Subtle
<i>N</i>	436		277	2011	<i>N</i>	1279		1138	2564
Mean <i>d</i>	-.405		-.639	-.224	Mean <i>d</i>	-.172		-.184	-.239
Var <i>d</i>	.077		.058	.201	Var <i>d</i>	.39		.181	.193
Var <i>e</i>	.057		.108	.051	Var <i>e</i>	.07		.072	.051
Mean δ	-.441		-.696	-.244	Mean δ	-.188		-.201	-.261
Var δ	.024		0	.179	Var δ	.381		.13	.169
% var SE	73.37		100	25.1	% var SE	17.9		39.63	26.42
% var acc. for (V%)	73.86		100	25.16	% var acc. for (V%)	17.92		39.67	26.49
90% CI	(-.64) - (-.24)		n/a	(-.79) - (.3)	90% CI	(-.98) - (.60)		(-.66) - (.26)	(-.8) - (.27)
Fail safe <i>N</i>	30		52	81	Fail safe <i>N</i>	60		57	108

As shown in the right columns in the bottom row of Table 13, the negative stereotype concerning women's weaker mathematical ability (than men), when activated, yields a different pattern of findings from the race-based stereotype. Studies using moderately explicit cues yield a comparable mean effect size (mean $d = |.18|$, var $d = .18$, $k = 20$, $n = 1138$) to that in studies using blatant cues (mean $d = |.17|$, var $d = .39$, $k = 22$, $n = 1279$). Studies employing subtle stereotype threat cues yield the largest mean effect size (mean $d = |.24|$, var $d = .19$, $k = 32$, $n = 2564$), although the effect size differences among subsets were trivial. However, $V\%$ values are between 18 and 40 percent, and the 90% credibility intervals overlap zero, suggesting that other moderators would further explain the variance in these d values and that the findings are not conclusive. In terms of fail safe N analyses, the relatively larger N of file-drawer studies indicate that the file-drawer problem is not probable for these subsets. Therefore, Hypothesis 2 was only partially supported.

Moderator Analysis: Stereotype Threat-Removing Strategies

Hypothesis 3 predicted that, among studies with a stereotype threat-removed condition, the more explicit stereotype threat-removal cues or strategies of stereotype threat were, the smaller a mean stereotype threat effect would be found. As shown in the top row of Table 14, the mean effect size (d) and mean true effect size (δ) for the subset of 93 d -values and a total sample size n of 5075 are $|.3|$ and $|.33|$, respectively. The observed variance is .31 and the true variance is .28. The study artifacts of sampling error in cognitive ability tests explain about 24 percent of the variance of observed d -values in this subset. However, the 90% credibility interval overlaps zero, indicating true moderator effects and inconclusive findings of this mean effect size.

As shown in the bottom row of Table 14, stereotype threat-removing strategies produce differential mean effect size estimates in that explicit removal works more effectively than subtle removals in reducing stereotype threat effects, at least at the mean effect size level. Specifically, for studies using explicit removal strategies ($k = 38$; $n = 1886$), the mean effect size (d) and mean true effect size (δ) are both $|.24|$. The observed and true variance estimates are both $.26$. The study artifacts of sampling error in cognitive ability tests explain about 28 percent of the variance of observed d -values in this subset. For studies using subtle removal strategies ($k = 55$; $n = 3189$), the mean effect size (d) and mean true effect size (δ) are $|.35|$ and $|.38|$, respectively (slightly larger than the overall mean d for this subset and the explicit mean d). The observed and true variance values are $.32$ and $.29$, respectively. However, the study artifacts of sampling error in cognitive ability tests explain between 23 and 28 percent of the variance of observed d -values. The 90% credibility intervals still overlap zero, indicating that true moderator effects exist and one should not interpret the findings of mean effect size for these subsets as conclusive evidence. Large fail-safe N s indicate that a file-drawer problem is unlikely in these cases. Therefore, the subsets of stereotype threat-removal strategies were analyzed by group-based stereotypes.

Table 14

Hierarchical Moderator Analyses of Stereotype Threat Removal Strategies

<u>Overall (ST-Removal Cues)</u>		
<i>K</i>	93	
<i>N</i>	5075	
Mean <i>d</i>	- .30	
Var <i>d</i>	.311	
Var <i>e</i>	.075	
Mean δ	- .33	
Var δ	.28	
% var SE	24.14	
% var acc. for (V%)	24.21	
90% CI	(-1.01)-(.35)	
Fail safe <i>N</i>	372	

	<u>STR Explicit^a</u>	<u>STR Subtle^a</u>
<i>K</i>	38	55
<i>N</i>	1886	3189
Mean <i>d</i>	- .24	- .35
Var <i>d</i>	.26	.317
Var <i>e</i>	.082	.071
Mean δ	- .22	- .33
Var δ	.258	.286
% var SE	27.48	22.74
% var acc. for (V%)	27.52	22.83
90% CI	(- .89) - (.41)	(-1.07) - (.30)
Fail safe <i>N</i>	129	248

Note. *K* = Number of effect sizes (*d*-values). *N* = Total sample size. Mean (*d*) = Sample size weighted mean effect size. Var (*d*) = Sample size weighted observed variance of *d*-values. Var (*e*) = Variance attributed to sampling error variance. Mean (δ) = Mean true effect size. Var (δ) = True variance of effect sizes. % var SE = Percent variance in observed *d*-values due to sampling error variance. % var acc. for (V%) = Percent variance in observed *d*-values due to all corrected artifacts. 90% CI = 90 percentile of (δ) (credibility interval).

^a Levels of cues that may remove stereotype threat: "Explicit" = explicitly refuting group differences in cognitive ability test performance or implementing interventions to boost stereotyped group members' performance. "Subtle" = subtle and indirect strategies that aim at changing stereotyped test takers' mental frame of reference (of a test, a testing purpose or a testing situation).

The columns in the top row of Table 15 show that stereotype threat-removal strategies differentially affect minority test takers (mean $d = |.42|$, var $d = .25$, $k = 30$, $n = 1661$) and women test takers' cognitive ability test performance (mean $d = |.23|$, var $d = .34$, $k = 61$, $n = 3310$), in that removal strategies worked better on women's math test performance than on minorities' test performance, at least at the mean level. The fail-safe N values are large, indicating no file-drawer problems.

The left columns in the bottom row of Table 15 show that, when broken down by levels of explicitness in type of removal strategies or interventions, minority test takers seem to benefit more from subtle or indirect strategies (mean $d = |.38|$, var $d = .25$, $k = 25$, $n = 1504$) than from direct, explicit ones (mean $d = |.80|$, var $d = .05$, $k = 5$, $n = 157$). Study artifacts explain all variance in the explicit-removal strategy subset of d -values, indicating that this finding of interest is conclusive although one should be cautious about generalizing this finding because the sample size was small ($k = 5$) and there was a smaller fail safe N of 45. However, study artifacts explain only 28 percent of the variance in the subtle-removal strategy subset of d -values and the 90% credibility interval overlaps zero, indicating true moderators and inconclusive findings. The right columns in the bottom row of Table 15 show that female test takers benefit more from a direct, explicit type of stereotype threat-removal strategies (mean $d = |.14|$, var $d = .29$, $k = 31$, $n = 1626$) than from subtle strategies (mean $d = |.33|$, var $d = .37$, $k = 30$, $n = 1684$). However, low $V\%$ values and zero-overlapping 90% credibility intervals indicate the effects of other true moderators or inconclusive findings of mean effect sizes. Large fail safe N s indicate unlikely file-drawer problems. Therefore, Hypothesis 3 was only partially supported.

Table 15

Hierarchical Meta-Analytic Findings: Stereotype Threat Removal Strategies by Group-Based Stereotypes

<u>Minority test takers - Overall (ST-Removal Cues)</u>				<u>Women test takers - Overall (ST-Removal Cues)</u>			
<i>K</i>	30			<i>K</i>	61		
<i>N</i>	1661			<i>N</i>	3310		
Mean <i>d</i>	-.415			Mean <i>d</i>	-.233		
Var <i>d</i>	.245			Var <i>d</i>	.337		
Var <i>e</i>	.075			Var <i>e</i>	.075		
Mean δ	-.452			Mean δ	-.254		
Var δ	.201			Var δ	.311		
% var SE	30.53			% var SE	22.34		
% var acc. for (V%)	30.69			% var acc. for (V%)	22.38		
90% CI	(-1.03) - (.12)			90% CI	(-.97) - (.46)		
Fail safe <i>N</i>	155			Fail safe <i>N</i>	203		

<u>STR Explicit</u>				<u>STR Subtle</u>			
<i>K</i>	5			<i>K</i>	31		
<i>N</i>	157			<i>N</i>	1626		
Mean <i>d</i>	-.8			Mean <i>d</i>	-.135		
Var <i>d</i>	.053			Var <i>d</i>	.285		
Var <i>e</i>	.14			Var <i>e</i>	.078		
Mean δ	-.87			Mean δ	-.147		
Var δ	0			Var δ	.245		
% var SE	100			% var SE	27.17		
% var acc. for (V%)	100			% var acc. for (V%)	27.19		
90% CI	n/a			90% CI	(-.78) - (.49)		
Fail safe <i>N</i>	45			Fail safe <i>N</i>	73		

<u>STR Explicit</u>				<u>STR Subtle</u>			
<i>K</i>	5			<i>K</i>	31		
<i>N</i>	157			<i>N</i>	1626		
Mean <i>d</i>	-.8			Mean <i>d</i>	-.135		
Var <i>d</i>	.053			Var <i>d</i>	.285		
Var <i>e</i>	.14			Var <i>e</i>	.078		
Mean δ	-.87			Mean δ	-.147		
Var δ	0			Var δ	.245		
% var SE	100			% var SE	27.17		
% var acc. for (V%)	100			% var acc. for (V%)	27.19		
90% CI	n/a			90% CI	(-.78) - (.49)		
Fail safe <i>N</i>	45			Fail safe <i>N</i>	73		

Moderator Analysis: Domain Identification

Hypothesis 4 predicted that studies targeting participants classified as highly domain identified would produce the largest mean threat effect size compared with studies with other levels of domain identification, whereas studies with a sample low in domain identification would produce the smallest mean threat effect size, with the medium level of domain identification being in between.

As shown in the top row of Table 16, the mean effect size (d) and mean true effect size (δ) for the total subset of 25 effect size values and a sample size n of 1119 are $|.32|$ and $|.35|$, respectively. The observed variance is .24 and the true variance is .18. However, the study artifacts of sampling error in cognitive ability tests explain about 38 percent of the cross-study variance of observed d -values, and the 90% credibility interval overlaps zero, indicating true moderator effects and inconclusive results.

As shown in the columns in the middle row of Table 16, low domain identifiers suffered the least in terms of cognitive ability test performance where stereotype threat is manipulated (mean $d = |.11|$, var $d = .19$, $k = 4$, $n = 307$), as compared with high domain identifiers (mean $d = |.32|$, var $d = .21$, $k = 12$, $n = 478$) and medium domain identifiers (mean $d = |.37|$, var $d = .29$, $k = 9$, $n = 313$). At the mean level, contrary to the theory, high domain identification yields only a similar mean effect size to that produced by moderate domain identification. However, all smaller $V\%$ values and zero-overlapping credibility intervals indicate that the findings are inconclusive because of the presence of true moderators that explain the variance in the data set.

Table 16

Hierarchical Moderator Analyses of Domain Identification^a

<u>Overall Findings</u>			
<i>K</i>	25		
<i>N</i>	1119		
Mean <i>d</i>	- .323		
Var <i>d</i>	.243		
Var <i>e</i>	.092		
Mean δ	- .353		
Var δ	.179		
% var SE	37.8		
% var acc. for (V%)	37.9		
90% CI	(- .9)-(.19)		
Fail safe <i>N</i>	106		

	<u>High Domain Ident.</u>	<u>Medium Domain Ident.</u>	<u>Low Domain Ident.</u>
<i>K</i>	12	9	4
<i>N</i>	478	313	307
Mean <i>d</i>	- .316	- .371	- .11
Var <i>d</i>	.21	.29	.194
Var <i>e</i>	.103	.12	.053
Mean δ	- .344	- .404	- .12
Var δ	.127	.203	.168
% var SE	49.21	41	27.24
% var acc. for (V%)	49.32	41.1	27.26
90% CI	(- .8)-(.11)	(- .98)-(.17)	(- .65)-(.40)
Fail safe <i>N</i>	50	42	8

	<u>Women^b</u> <u>High Domain Ident.</u>	<u>Women</u> <u>Medium Domain Ident.</u>	<u>Women</u> <u>Low Domain Ident.</u>
<i>K</i>	9	6	4
<i>N</i>	380	212	307
Mean <i>d</i>	- .287	- .518	- .111
Var <i>d</i>	.201	.204	.194
Var <i>e</i>	.097	.119	.053
Mean δ	- .313	- .565	- .12
Var δ	.123	.1	.168
% var SE	48.44	58.29	27.24
% var acc. for (V%)	48.54	58.59	27.26
90% CI	(- .76) - (.14)	(- .97) - (- .16)	(- .65) - (.40)
Fail safe <i>N</i>	35	37	8

Note. *K* = Number of effect sizes (d-values). *N* = Total sample size. Mean (*d*) = Sample size weighted mean effect size. Var (*d*) = Sample size weighted observed variance of d-values. Var (*e*) = Variance attributed to sampling error variance. Mean (δ) = Mean true effect size. Var (δ) = True variance of effect sizes. % var SE = Percent variance in observed d-values due to sampling error variance. % var acc. for = Percent

variance in observed d -values due to all corrected artifacts. 90% CI = 90 percentile of (δ) (credibility interval).

^a Domain identification levels: “High” = strongly identified with academic or cognitive ability domains. “Medium” = moderately identified. “Low” = weakly identified.

^b Only the subsets of female test takers were meta-analyzed here because there were too few minority d -values in this moderator category.

When the subsets d -values for female test takers were meta-analyzed (there was insufficient number of race-based studies contributing effect size estimates to conduct the moderator analyses of interest), as shown in the columns in the bottom row of Table 16, low math domain identified women do not suffer much in terms of their cognitive ability test performance where stereotype threat is manipulated (mean $d = |.11|$, var $d = .19$, $k = 4$, $n = 307$), as compared with high math domain identified females (mean $d = |.29|$, var $d = .20$, $k = 9$, $n = 380$) and moderately math identified women (mean $d = |.52|$, var $d = .20$, $k = 6$, $n = 212$). (The small sample size of the “low domain identification” subset may be the cause for being cautious in interpreting the findings though.) Nevertheless, smaller V% values and zero-included credibility intervals indicate that these findings are inconclusive because of other moderators that may explain the variance in the data over and above study artifacts. Further, lower fail-safe N values show that the conclusions based on these meta-analytic findings may be uncertain. Therefore, Hypothesis 4 was not supported.

Moderator Analysis: Test Difficulty

Hypothesis 5 predicted that studies using highly difficult cognitive ability tests would yield a larger mean effect size than that in studies using moderately difficult and easy tests. As shown in the top row in Table 17, the mean effect size (d) and mean true effect size (δ) for the total set of 81 effect size values and a total sample size n of 4029 are $|.28|$ and $|.30|$, respectively. The observed variance is $.31$ and the true variance is $.27$. However, the smaller V% and the zero-included credibility interval indicate inconclusive findings because of the presence of true moderators explaining data variance.

Table 17

Hierarchical Moderator Analyses of Test Difficulty ^a

<u>Overall Findings</u>			
<i>K</i>	81		
<i>N</i>	4029		
Mean <i>d</i>	-.279		
Var <i>d</i>	.307		
Var <i>e</i>	.082		
Mean δ	-.304		
Var δ	.267		
% var SE	26.81		
% var acc. for (V%)	26.87		
90% CI	(-.97)-(.35)		
Fail safe <i>N</i>	307		

	<u>High</u>	<u>Medium</u>	<u>Low</u>
<i>K</i>	48	24	9
<i>N</i>	2161	1560	308
Mean <i>d</i>	-.394	-.19	.083
Var <i>d</i>	.396	.153	.199
Var <i>e</i>	.092	.063	.119
Mean δ	-.429	-.208	.091
Var δ	.361	.107	.095
% var SE	23.2	40.86	59.74
% var acc. for (V%)	23.29	40.92	59.74
90% CI	(-1.2)-(.34)	(-.63)-(.21)	(-.3)-(.49)
Fail safe <i>N</i>	237	70	2

Note. *K* = Number of effect sizes (d-values). *N* = Total sample size. Mean (*d*) = Sample size weighted mean effect size. Var (*d*) = Sample size weighted observed variance of d-values. Var (*e*) = Variance attributed to sampling error variance. Mean (δ) = Mean true effect size. Var (δ) = True variance of effect sizes. % var SE = Percent variance in observed d-values due to sampling error variance. % var acc. for = Percent variance in observed d-values due to all corrected artifacts. 90% CI = 90 percentile of (δ) (credibility interval).

^a Test difficulty levels: "High" = very difficult to difficult. "Medium" = average, mixed difficult. "Low" = easy.

As shown in the columns in the bottom row of Table 17, test takers seem to suffer the most from situational stereotype threat when cognitive ability tests are difficult (mean $d = |.39|$, var $d = .40$, $k = 48$, $n = 2161$), followed by moderately difficult tests (mean $d = |.19|$, var $d = .15$, $k = 24$, $n = 1560$) and easy tests (mean $d = |.08|$, var $d = .20$, $k = 9$, $n = 308$). However, smaller V% values and zero-overlapping credibility intervals still indicate inconclusive findings and moderator effects. Large fail safe N s for difficult and moderately difficult tests suggest no file-drawer bias. However, the small fail safe N for the easy test subset suggest one should be cautious in interpreting and generalizing this meta-analytic finding.

Additional hierarchical meta-analyses across subsets of minority and female test takers and levels of math test difficulty qualified these findings. The left columns in the bottom row of Table 18 demonstrate that minorities perform more poorly compared with their true ability when cognitive ability tests are highly difficult (mean $d = |.43|$, var $d = .16$, $k = 12$, $n = 549$) than when tests are moderately difficulty (mean $d = |.18|$, var $d = .07$, $k = 10$, $n = 647$). (There were no studies using easy tests to investigate stereotype threat effects among minority test takers.) The credibility intervals did not overlap zero, meaning that the findings were conclusive.

Table 18

Hierarchical Meta-Analytic Findings: Test Difficulty Mitigating Stereotype Threat Effects by Group-Based Stereotypes

<u>Minority test takers - Overall</u>		<u>Women test takers - Overall</u>	
<i>K</i>	22	<i>K</i>	55
<i>N</i>	1196	<i>N</i>	2706
Mean <i>d</i>	-.294	Mean <i>d</i>	-.25
Var <i>d</i>	.126	Var <i>d</i>	.387
Var <i>e</i>	.075	Var <i>e</i>	.083
Mean δ	-.32	Mean δ	-.373
Var δ	.06	Var δ	.361
% var SE	59.72	% var SE	21.45
% var acc. for (V%)	59.88	% var acc. for (V%)	21.49
90% CI	(-.63) - (-.01)	90% CI	(-1.04) - (-.5)
Fail safe <i>N</i>	87	Fail safe <i>N</i>	193

<u>Difficult</u>		<u>Medium</u>		<u>Easy</u>	
<i>K</i>	12	<i>K</i>	33	<i>K</i>	9
<i>N</i>	549	<i>N</i>	1508	<i>N</i>	308
Mean <i>d</i>	-.425	Mean <i>d</i>	-.363	Mean <i>d</i>	.083
Var <i>d</i>	.157	Var <i>d</i>	.5	Var <i>d</i>	.199
Var <i>e</i>	.091	Var <i>e</i>	.09	Var <i>e</i>	.119
Mean δ	-.464	Mean δ	-.395	Mean δ	.091
Var δ	.078	Var δ	.487	Var δ	.095
% var SE	57.84	% var SE	18.04	% var SE	59.74
% var acc. for (V%)	58.11	% var acc. for (V%)	18.1	% var acc. for (V%)	59.74
90% CI	(-.82) - (-.11)	90% CI	(-1.29) - (-.50)	90% CI	(-.3) - (-.49)
Fail safe <i>N</i>	63	Fail safe <i>N</i>	153	Fail safe <i>N</i>	2

The right columns in the bottom row of Table 18 shows that women tend to underperform when a math test is highly difficult (mean $d = |.36|$, var $d = .46$, $k = 36$, $n = 1705$) more than when a math test is only moderately difficult (mean $d = |.18|$, var $d = .20$, $k = 13$, $n = 890$), or when it is easy (previously reported). Study artifacts explain between 18 and 60 percent of the variance in these subsets. Zero-included credibility intervals further suggest moderating effects. Except for the “difficult” subset of d -values, the smaller file-drawer N values indicate that the findings for women in terms of test difficulty may not be conclusive. Therefore, Hypothesis 5 was only partially supported.

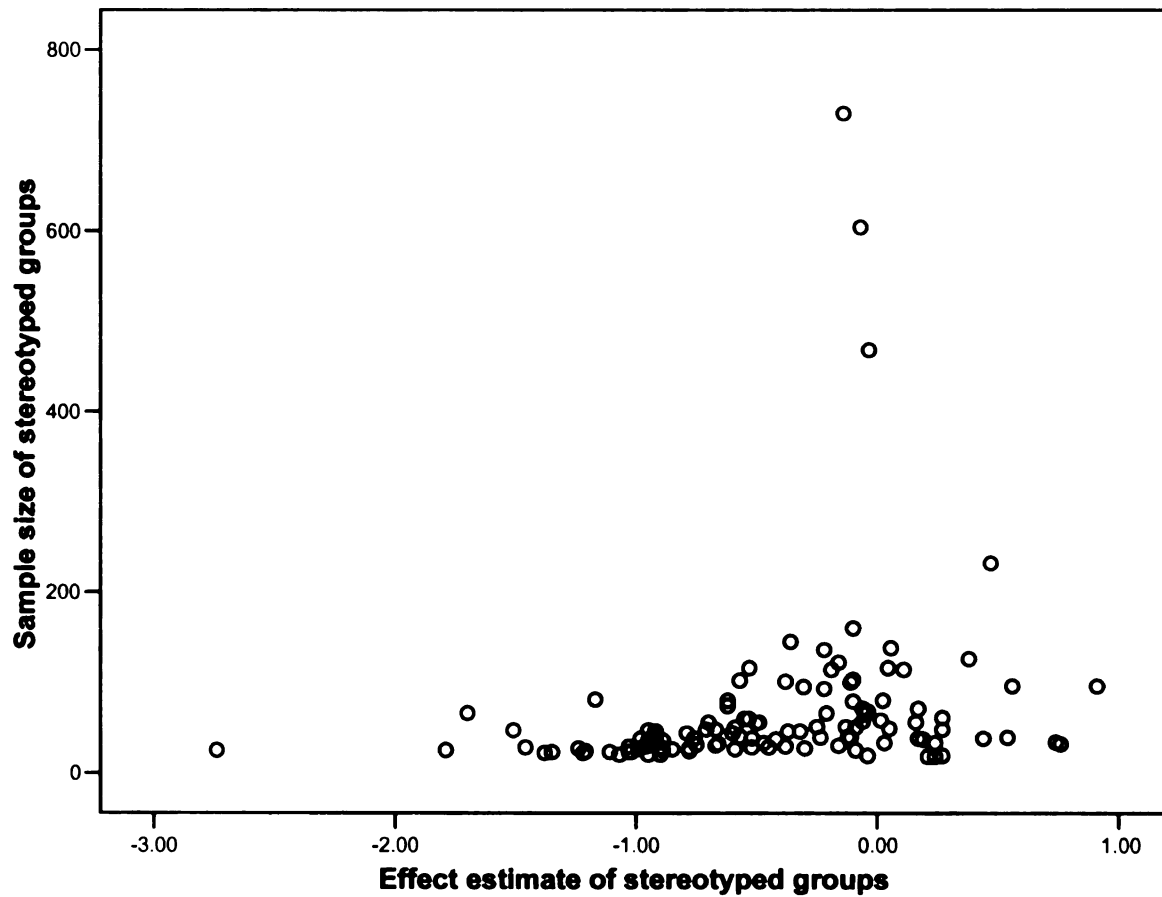
Supplemental Bias Analysis

As mentioned, a different method of detecting potential publication biases in a meta-analytic data set is graphing a funnel plot (i.e., plotting effect size estimates against study sample size; see Light & Pillemer, 1984). The principle of this graphing technique is that a publication bias or other types of location biases are present when only large effects are reported by studies with a small sample size. On the contrary, there are no biases if an exclusion of null results is not visible on the funnel graph. As shown in Figure 2, the funnel plot for the full meta-analytic data set resembles a relatively symmetrical inverted funnel (i.e., results from smaller sample-size studies being scattered widely at the bottom of the graph). This plot indicates the absence of location and/or publication bias in the data set.

Further, the relationship between effect size estimates and study sample sizes was positive and statistically significant ($r = .23$, $p < .05$). That means, the larger study sample sizes were, the greater effect size estimates were (i.e., increasingly non-negative values).

Figure 2

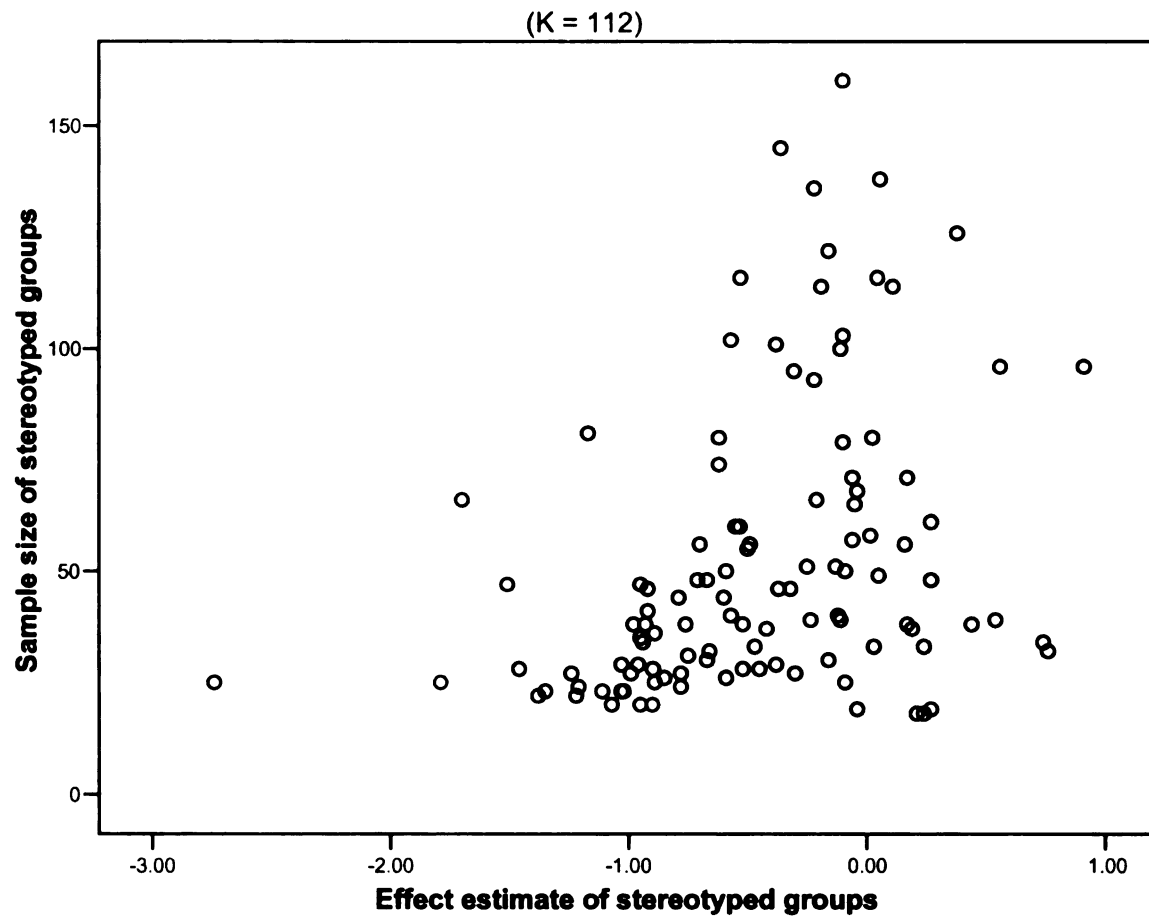
The funnel graph of stereotype threat effects on target test takers' cognitive ability test performance. The effect size estimates are plotted against study sample sizes ($r = .23, p < .05$).



If four primary studies each with a sample size larger than 200 were excluded from the data set (Anderson, 2001; Dinella, 2004; Stricker & Ward, 2004, Study 1b & Study 2), a similar pattern of findings was also found in terms of bias analysis and correlation analysis (see Figure 3). The correlation coefficient of effect estimates and sample sizes was positive and significant ($r = .35, p < .01$).

Figure 3

The funnel graph of stereotype threat effects on target test takers' cognitive ability test performance in the absence of large sample studies. The effect size estimates are plotted against study sample sizes ($r = .35$, $p < .01$).



One additional question is whether studies that yielded either positive effect size estimates or estimates clustering around the zero point ($k = 29$; 25% of the data set) have differential characteristics from studies the d -values of which supported the hypothesis of performance interference (i.e., a negative effect size). A peruse of the general characteristics of samples in subsets of studies at different levels of effect size estimates revealed the fact that the patterns of study characteristics were mixed.

Specifically, among studies yielding d -estimates greater than or equal to .05 (a rounded value; $k = 20$), in a majority of studies a race-based stereotype was activated ($k = 12$ or 60% of this subset). Further, most of these studies did not pre-screen participants for high domain identifiers ($k = 19$; 95%). Among studies that yielded d -estimates which were smaller than .05 but greater than -.05 (rounded values; $k = 9$), in a majority of these studies a gender-based stereotype was activated ($k = 8$; 89%). Most of these studies did not pre-screen participants for those high on cognitive ability domain identification ($k = 6$; 66.7%). About half of the whole group of “non-effect” studies were unpublished papers ($k = 15$; 52%).

In comparison, the majority of the data set consisted of d -estimates that were in line with a stereotype threat prediction (i.e., negative values equal to or smaller than -.05; $k = 87$ or 75%). Overall, in one-third of these studies a race-based stereotype was activated ($k = 30$; 34.5%), a smaller proportion than that in the “non-effect” subset; the rest were gender-based stereotype studies. Participants were not prescreened for high domain identifiers in most studies ($k = 68$; 78.16%), although this ratio was slightly smaller than that among the group of null-finding and positive-finding studies (86.21%). About one-third of the whole group of “stereotype threat effect” studies were unpublished

papers ($k = 34$; 39%), which was a smaller proportion than that in the first subset of studies. In each subset, participants were recruited from either private and elite universities or public universities. In other words, there are no clearly defining characteristics that can distinguish studies that found no stereotype threat effects or positive effects from studies that found the effects of interest.

The only clear difference is that, whereas there was only one non-American sample in the “non-effect” group of studies (3.5%), there was 23 non-American samples (26.5%) in the “stereotype threat effect” group, suggesting that non-American authors (or American authors who used non-American samples) might be more likely to publish significant findings that were consistent with the hypothesis of performance interference in American journals than non-significant findings or findings that were contradictory to the hypothesis.

Between-Group Meta-Analytic Findings

Overall Between-Group Stereotype Threat Effects

Of interest to some readers is how stereotyped test takers may empirically fare in terms of performance on cognitive ability tests across levels of stereotype threat manipulation, not only in comparison with their own groups, but also in comparison with the performance of members of non-stereotyped, reference groups. In fact, this research question has been the focal one in a portion of stereotype threat studies in this review because it may have important applied implications for educational and employment testing. There is a lack of conceptual premises in the theory of stereotype threat based on which one could generate meaningful hypotheses for between-group test performance comparisons within a stereotype threat framework. Nor could potential moderating effects be hypothesized from the theory. Therefore, only exploratory between-group meta-analyses were conducted across type of group-based stereotypes in the present study.

Table 19 presents the meta-analytic findings regarding the relationships of interest. Up to 62 primary studies that followed a between-group stereotype threat research design (e.g., women vs. men; ethnic minority test takers vs. majority ones) contributed effect size estimates to the meta-analyses.

Table 19

Hierarchical Meta-Analytic Findings of Between-Group Mean Test Performance across Stereotype Threat Levels

In Control Conditions				In STA Conditions				In STR Conditions			
K	23	K	62	K	62	K	46	K	62	K	46
N	3620	N	5937	N	5937	N	2603	N	5937	N	2603
Mean d	-.44	Mean d	-.53	Mean d	-.53	Mean d	-.28	Mean d	-.53	Mean d	-.28
Var d	.08	Var d	.16	Var d	.16	Var d	.13	Var d	.16	Var d	.13
Var e	.03	Var e	.04	Var e	.04	Var e	.07	Var e	.04	Var e	.07
Mean δ	-.48	Mean δ	-.58	Mean δ	-.58	Mean δ	-.3	Mean δ	-.58	Mean δ	-.3
Var δ	.06	Var δ	.13	Var δ	.13	Var δ	.07	Var δ	.13	Var δ	.07
% var SE	34.88	% var SE	27.8	% var SE	27.8	% var SE	54.76	% var SE	27.8	% var SE	54.76
% var acc. for (V%)	35.48	% var acc. for (V%)	28.22	% var acc. for (V%)	28.22	% var acc. for (V%)	54.9	% var acc. for (V%)	28.22	% var acc. for (V%)	54.9
90% CI	(-.79)(-.17)	90% CI	(-1.05)(-.11)	90% CI	(-1.05)(-.11)	90% CI	(-.64)(.04)	90% CI	(-1.05)(-.11)	90% CI	(-.64)(.04)
Fail safe N	124	Fail safe N	391	Fail safe N	391	Fail safe N	175	Fail safe N	391	Fail safe N	175

Minority v. Majority ^a				Minority v. Majority ^a				Minority v. Majority ^a			
K	10	K	23	K	23	K	14	K	23	K	14
N	1695	N	2498	N	2498	N	848	N	2498	N	848
Mean d	-.564	Mean d	-.686	Mean d	-.686	Mean d	-.377	Mean d	-.686	Mean d	-.377
Var d	.0195	Var d	.065	Var d	.065	Var d	.182	Var d	.065	Var d	.182
Var e	.025	Var e	.039	Var e	.039	Var e	.068	Var e	.039	Var e	.068
Mean δ	-.615	Mean δ	-.747	Mean δ	-.747	Mean δ	-.41	Mean δ	-.747	Mean δ	-.41
Var δ	0	Var δ	.065	Var δ	.065	Var δ	.135	Var δ	.065	Var δ	.135
% var SE	100	% var SE	60.29	% var SE	60.29	% var SE	37.42	% var SE	60.29	% var SE	37.42
% var acc. for (V%)	100	% var acc. for (V%)	61.95	% var acc. for (V%)	61.95	% var acc. for (V%)	37.6	% var acc. for (V%)	61.95	% var acc. for (V%)	37.6
90% CI	n/a	90% CI	(-.97)(-.53)	90% CI	(-.97)(-.53)	90% CI	(-.88)(.06)	90% CI	(-.97)(-.53)	90% CI	(-.88)(.06)
Fail safe N	66	Fail safe N	181	Fail safe N	181	Fail safe N	67	Fail safe N	181	Fail safe N	67

Minority v. Women				Minority v. Women				Minority v. Women			
K	13	K	39	K	39	K	32	K	39	K	32
N	1803	N	3330	N	3330	N	1765	N	3330	N	1765
Mean d	-.264	Mean d	-.392	Mean d	-.392	Mean d	-.232	Mean d	-.392	Mean d	-.232
Var d	.025	Var d	.186	Var d	.186	Var d	.101	Var d	.186	Var d	.101
Var e	.029	Var e	.048	Var e	.048	Var e	.074	Var e	.048	Var e	.074
Mean δ	-.288	Mean δ	-.428	Mean δ	-.428	Mean δ	-.252	Mean δ	-.428	Mean δ	-.252
Var δ	0	Var δ	.163	Var δ	.163	Var δ	.032	Var δ	.163	Var δ	.032
% var SE	100.00	% var SE	26.08	% var SE	26.08	% var SE	73.02	% var SE	26.08	% var SE	73.02
% var acc. for (V%)	100	% var acc. for (V%)	26.27	% var acc. for (V%)	26.27	% var acc. for (V%)	73.15	% var acc. for (V%)	26.27	% var acc. for (V%)	73.15
90% CI	n/a	90% CI	(-.94)(.09)	90% CI	(-.94)(.09)	90% CI	(-.48)(.02)	90% CI	(-.94)(.09)	90% CI	(-.48)(.02)
Fail safe N	66	Fail safe N	192	Fail safe N	192	Fail safe N	106	Fail safe N	192	Fail safe N	106

Note. K = Number of effect sizes (d -values). N = Total sample size. $\text{Mean}(d)$ = Sample size weighted mean effect size. $\text{Var}(d)$ = Sample size weighted observed variance of d -values. $\text{Var}(e)$ = Variance attributed to sampling error variance. $\text{Mean}(\delta)$ = Mean true effect size. $\text{Var}(\delta)$ = True variance of effect sizes. % var SE = Percent variance in observed d -values due to sampling error variance. % var acc. for (V%) = Percent variance in observed d -values due to all corrected artifacts. 90% CI = 90 percentile of (δ) (credibility interval).

^a Five primary studies from the entire data set that used White test takers as the stereotyped group were excluded in these analyses (Aronson, et al., 1999; Smith & White, 2002; von Hippel, et al., 2005) so that only minority subgroups' d -values were meta-analyzed.

As shown in the columns in the top row of Table 19, the between-group effect values increase from a mean effect size $d = .44$ (var $d = .08$, $k = 23$, $n = 3620$) in test-only control conditions to a mean effect size $d = .53$ (var $d = .16$, $k = 62$, $n = 5937$) in stereotype threat-activated conditions. When interventions or threat-removing strategies are implemented, stereotyped test takers tend to underperform on cognitive ability tests compared with reference test takers (mean $d = .28$, var $d = .13$; $k = 46$, $n = 2603$). After correction for study artifacts, the true mean effect sizes increase slightly to $.48$, $.58$, and $.30$ across the conditions of control, threat-activation, and threat-removal, respectively, and the corresponding variance values are slightly reduced.

Nevertheless, the variance values are non-zero and V% estimates for these subsets are lower than 75 percent, indicating further true moderator effects. The credibility intervals for mean d s in stereotype threat-activated conditions and in stereotype threat-removed conditions overlap zero, meaning that these findings are not conclusive. The credibility interval for mean d in control conditions does not overlap zero, however. Large fail safe N values indicate that a file-drawer problem is unlikely for these subsets.

Potential Moderator: Group-Based Stereotypes

Again, the types of group-based stereotypes may play a mitigating role explaining the variance in d -values as evidenced with the within-group findings. Therefore, subsequent hierarchical meta-analyses across group-based stereotypes were conducted. As shown in the left columns in the bottom row of Table 19, in test-only control conditions, ethnic minority test takers underperform compared with majority test takers and the between-group mean effect size is $.56$ (var $d = .02$; $k = 10$, $n = 1695$). The value of true mean effect size after correction for artifacts is slightly larger, mean $\delta = .62$,

whereas the true variance is slightly smaller. In other words, on the average, ethnic minority test takers' cognitive ability test scores are approximately at the 30th percentile of majority groups' mean test scores, which is relatively consistent with the literature on subgroup mean differences in cognitive ability test performance (i.e., the overall mean standardized differences for g are 1.10 for the Black-White difference and .72 for the Latino-White difference; Roth, Bevier, Bobko, Switzer, & Tyler, 2001).

Note that study artifacts such as sampling error explain all observed variance in the d values in this subset, suggesting that no further moderator analyses should be conducted for this subset. Although the number of d -values in this subset is small ($k = 10$), the fail-safe N value of 66 is sufficiently large, and similar minority-majority group mean differences in test performance are generally observed in the broad cognitive ability testing literature. Therefore, a file-drawer problem is not probable.

In test-only control conditions, female test takers underperform compared with men on mathematical ability tests and the between-group mean effect size is $|.26|$ ($\text{var } d = .03$; $k = 13$, $n = 1803$), which is consistent with the literature (see a review by Hyde & Kling, 2001). The value of true mean effect size after correction for artifacts is slightly larger, mean $\delta = |.29|$, whereas the true variance is slightly smaller. In other words, on the average, women's mean math test scores are approximately at the 40th percentile of men's mean math test scores. Study artifacts explain all of the variance in d -values, suggesting no other moderator for this subset. Although the fail safe N value is not very large in this case (47), similar overall gender mean differences in math test performance are generally observed in the testing literature; therefore, a file-drawer problem is possible but not plausible.

As shown in the middle columns in the bottom row of Table 19, in stereotype threat-activation conditions, ethnic minority test takers underperform compared with majority test takers and the between-group mean effect size was $|.69|$ ($\text{var } d = .07$; $k = 23$, $n = 2498$). True mean effect size after correction for artifacts is slightly larger, mean $\delta = |.75|$, whereas the true variance is slightly smaller. In other words, when stereotype threat is activated, on the average, ethnic minority test takers' mean cognitive ability test scores are approximately at the 25th percentile of majority groups' mean test scores, which is worse than the results in test-only control conditions. Study artifacts explain about 62 percent of the observed variance in d values, suggesting that further moderator effects should be investigated for this subset. The credibility interval does not overlap zero, indicating a conclusive result.

In stereotype threat-activated conditions, female test takers underperform compared with men on mathematical ability tests and the between-group mean effect is $|.39|$ ($\text{var } d = .19$; $k = 39$, $n = 3330$). The true mean effect after correction for artifacts is slightly larger, mean $\delta = |.43|$, whereas the true variance is slightly smaller. In other words, when stereotype threat is activated, on the average, women's mean math test scores are approximately at the 34th percentile of men's mean math scores. Study artifacts explain only 26 percent of the variance in d , suggesting other true moderator effects. The zero-included credibility interval indicates an inconclusive finding. For both minorities and women, the large fail-safe Ns indicate that a file-drawer problem is not likely.

The right columns in the bottom row of Table 19 present the meta-analytic results for between-group cognitive ability test performance in experimental conditions where researchers actively introduce various strategies to remove a negative stereotype (i.e.,

stereotype threat-removed conditions). Ethnic minority test takers underperform compared with majority test takers; the between-group mean effect size is $|.38|$ (var $d = .18$; $k = 14$, $n = 848$). The true effect size value after correction for study artifacts is $|.41|$. This value represents a sharp decrease by $|.34|$ in between-group mean test performance, compared with the true mean effect of $|.75|$ in stereotype threat-activated conditions, and a decrease by $|.21|$ compared with the true mean effect of $|.62|$ in test-only control conditions. On the average, when stereotype threat effects are removed, ethnic minority test takers' mean test scores are approximately at the 34th percentile of majority groups' mean test scores. However, study artifacts explain about 38 percent of the observed variance in the d values in the subset findings, suggesting moderator effects. The overlapping-zero credibility interval indicates that this finding is not conclusive.

In stereotype threat-removed conditions, women underperform compared with men on mathematical ability tests and the between-group mean effect size is $|.23|$ (var $d = .10$; $k = 32$, $n = 1765$). The true mean effect size is $|.25|$. On the average, women's math test scores are approximately at the 41st percentile of men's mean math scores when stereotype threat-removing strategies are implemented. Study artifacts explain 73 percent of the variance in d , suggesting no true moderator(s) that can further explain this finding. The credibility interval does not include zero, indicating a meaningful effect. The large fail safe Ns indicate that a file-drawer problem is not likely for both minority and female subsets of studies.

Supplemental Meta-Analyses

Cognitive Ability Tests = Stereotype Threat?

One interesting design question that can be explored meta-analytically is whether stereotype threat is inherently embedded in any cognitive ability testing situation. In other words, is the fear of confirming a group stereotype of intellectual inferiority really “a threat in the air” for members of a stereotyped group as stereotype threat theorists posit (see Steele, 1997; Steele & Davis, 2003)? Would facing the prospect of taking an evaluative test of cognitive abilities be enough to invoke stereotype threat effects for some people? There is some empirical evidence to support this position (e.g., Spencer, et al., 1999). However, Sackett, Schmitt, Ellingson and Kabin (2001) imply that observed minority-majority stereotype threat effects might be more a product of laboratory manipulation than an actual phenomenon of cognitive ability tests and testing situations.

To explore the answer to this issue, I proceeded with a couple of additional meta-analyses. First, as a direct test, I meta-analyzed the mean effect size of a subset of d values, which is the standardized mean difference between stereotyped test takers in stereotype threat-activated conditions (subtle cues of test diagnosticity and social group status inquiries only) and those in test-only control conditions. The stereotype threat theorists’ position would be supported if the distribution of cognitive ability test scores for the stereotype threat-activated group overlaps completely or almost completely with the distribution of scores for the control group (i.e., approaching 0% of non-overlap). However, because this subset is very small ($k = 8$), I additionally compared a subset of d -values (stereotype threat-activated conditions vs. stereotype threat-removed conditions) with another a subset of d -values (control conditions vs. stereotype threat-removed

conditions). Again, the stereotype threat-activated conditions should employ subtle threat activating cues as described above. The prediction of interest would be supported when the standardized mean stereotype threat effects yielded from the two subsets are approximately equal with each other.

As shown in the left column in Table 20, for the subset of studies using subtle threat-activating cues, which is the most similar situation to a real-life cognitive ability test setting (i.e., emphasizing the diagnostic nature of a test and/or asking test takers to fill out a demographic question before tests) ($k = 8$; $n = 1011$), the mean effect size (d) and mean true effect size (δ) are $-.18$ and $-.20$, respectively ($\text{var } d = .17$). The negative sign of these values indicates that mean test scores of stereotype threat-activated groups are lower than those in control conditions, and there is approximately 14% of non-overlap between the two score distributions. The observed variance estimate is $.17$. The study artifacts of sampling error in cognitive ability tests explain all variance in the observed d -values, showing that there is no further moderator and that the finding is conclusive.

Table 20

Meta-Analytic Evidence for the Equivalency between Control and Subtle Stereotype Cues Conditions

	<u>STA - Control</u>		<u>Subtle STA - STR</u>	<u>Control - STR</u>
<i>K</i>	8	<i>K</i>	19	17
<i>N</i>	1011	<i>N</i>	1241	808
Mean <i>d</i>	-.178	Mean <i>d</i>	-.286	-.184
Var <i>d</i>	.169	Var <i>d</i>	.264	.236
Var <i>e</i>	.032	Var <i>e</i>	.063	.086
Mean δ	-.194	Mean δ	-.313	-.201
Var δ	0	Var δ	.239	.179
% var SE	100	% var SE	23.68	36.31
% var acc. for (V%)	100	% var acc. for (V%)	23.76	36.34
90% CI	n/a	90% CI	(-.94) - (.31)	(-.74) - (.34)
Fail safe <i>N</i>	73	Fail safe <i>N</i>	73	48

Note. *K* = Number of effect sizes (*d*-values). *N* = Total sample size. Mean (*d*) = Sample size weighted mean effect size. Var (*d*) = Sample size weighted observed variance of *d*-values. Var (*e*) = Variance attributed to sampling error variance. Mean (δ) = Mean true effect size. Var (δ) = True variance of effect sizes. % var SE = Percent variance in observed *d*-values due to sampling error variance. % var acc. for (V%) = Percent variance in observed *d*-values due to all corrected artifacts. 90% CI = 90 percentile of (δ) (credibility interval).

As shown in the right columns, for the subset of studies using subtle threat activating cues that are the most similar to a real-life cognitive ability test setting (i.e., emphasizing the diagnostic nature of a test and/or asking test takers to fill out a demographic question before tests) ($k = 19$; $n = 1241$), the mean effect size (d) and mean true effect size (δ) are $|.29|$ and $|.31|$, respectively. The observed and true variance estimates are $|.26|$ and $|.24|$, respectively. However, the study artifacts explain about 24 percent of the variance of observed d -values in this subset, suggesting other moderators. The 90% credibility intervals overlap zero, indicating an inconclusive finding.

Additionally, for studies that administered cognitive ability tests without any special instructions ($k = 17$; $n = 808$; not shown in Table 20), the mean effect size (d) and mean true effect size (δ) are $|.18|$ and $|.20|$, respectively. The observed and true variance values are $.24$ and $.18$, respectively. The study artifacts explain between 24 and 36 percent of the variance in the observed d -values, suggesting moderator effects. The zero-included credibility interval indicates an inconclusive finding.

Reference Group Members' Test Performance

I conducted parallel exploratory meta-analyses for members of reference groups across type of group-based stereotypes (subtle stereotype threat cues only). Table 21 presents the relevant meta-analytic findings.

Table 21

Meta-Analytic Findings for Reference Groups (Within-Group)

	<u>STA^a - Control</u>	<u>STA^a - STR</u>	<u>Control - STR</u>
<i>K</i>	12	18	4
<i>N</i>	3566	889	270
Mean <i>d</i>	-.031	.14	.30
Var <i>d</i>	.032	.12	.006
Var <i>e</i>	.01	.08	.06
Mean δ	-.03	.15	.32
Var δ	.22	.47	0
% var SE	42.75	67.36	100
% var acc. for (V%)	42.75	67.39	100
90% CI	(- .22) - (.15)	(- .13) - (.43)	n/a
Fail safe <i>N</i>	16	7	8

<u>Whites Only</u>			
	<u>STA^a - Control</u>	<u>STA^a - STR</u>	
<i>K</i>	7	10	
<i>N</i>	2178	618	
Mean <i>d</i>	-.05	.04	
Var <i>d</i>	.02	.09	
Var <i>e</i>	.01	.07	
Mean δ	-.06	.04	
Var δ	.005	.03	
% var SE	75.83	75.08	
% var acc. for (V%)	75.87	75.08	
90% CI	(- .15) - (.03)	(- .17) - (.25)	
Fail safe <i>N</i>	11	6	

<u>Men Only</u>			
	<u>STA^a - Control</u>	<u>STA^a - STR</u>	
<i>K</i>	5	8	
<i>N</i>	1388	271	
Mean <i>d</i>	.003	.36	
Var <i>d</i>	.053	.13	
Var <i>e</i>	.015	.12	
Mean δ	.003	.39	
Var δ	.05	.09	
% var SE	27.5	94.8	
% var acc. for (V%)	27.5	95.02	
90% CI	(- .27) - (.28)	(.28) - (.51)	
Fail safe <i>N</i>	5	21	

Note. ^a Subtle stereotype threat-activating cues only.

As shown in the columns in the top row of Table 21, subtle cues activating an out-group stereotype (i.e., targeting out-group stereotyped members) do not affect reference group members' cognitive ability test performance. In other words, reference individuals (i.e., Whites; men) may perform as well on a test without special directions as on a test with a statement about test diagnosticity or some pre-test demographic inquiries (mean $d = .03$, var $d = .03$). However, study artifacts explain only 43 percent of the variance in d -values, suggesting moderators. The credibility interval overlaps zero, indicating an inconclusive finding.

When stereotype threat-removing strategies are implemented in order to reduce stereotype threat effects for target stereotyped groups, these strategies inadvertently cause members of reference groups to *underperform* compared with other reference group members in stereotype threat conditions (mean $d = .14$). However, $V\%$ shows that there are other moderators explaining the variance in d -values, and the zero-included credibility interval indicates an inconclusive finding.

Reference members also underperformed in stereotype threat-removed conditions, compared with those in test-only control conditions (mean $d = .30$). $V\%$ values and credibility intervals show this finding is conclusive and artifacts account for all variance. Therefore, I conducted the hierarchical analyses of the "STA – Control" and "STA – STR" conditions across levels of sub-group stereotypes.

The columns in the middle row of Table 21 show that Whites' cognitive test performance do not change much regardless of stereotype threat conditions when stereotype threat activation is subtle (mean $ds = .05$ & $.04$). Although artifacts account for most variance in d -values, indicating no meaningful moderator effects need to be

further investigated, the credibility intervals overlap zero, suggesting that these findings are not conclusive. As shown in the bottom row of Table 21, like Whites, men's math test performance is not affected by subtle stereotype threat activation targeting females compared with those in control conditions (mean $d = .003$). This finding is not conclusive (an zero-included credibility interval) and moderators may account for more variance in the effect estimates (small $V\%$). Unlike Whites, men's math performance is negatively affected when out-group stereotype threat-removing strategies are implemented, compared with the performance of those in subtle threat activation conditions (mean $d = .36$). Study artifacts explain most of the variance in d -values and the credibility interval does not include zero, indicating a meaningful effect across studies.

Nevertheless, because fail-safe N values are very low for these subsets of d -values, the above conclusions are tentative and readers should exercise caution when generalizing these findings.

Summary

I conducted a series of hierarchical meta-analytic tests of the hypotheses, investigating stereotype threat effects on cognitive ability test performance of stereotyped group members in comparison with themselves and with members of reference groups under certain circumstances. The findings were mixed: the hypotheses of moderating effects of stereotype threat-activating cues, stereotype threat-removing strategies, and test difficulty were partially supported but only for minority test takers. The hypotheses of an overall effect and of the moderating effect of domain identification were not supported. Although most mean effects were negative values (i.e., indicating stereotype threat

effects), one could not rule out the possibility that the effects were zero or even positive values (i.e., indicating a stereotype reactance effect). Study artifacts accounted for a small proportion of the variance in d in most moderator meta-analyses, suggesting the presence of other potential moderators that have yet to be investigated in the present study due to insufficient sample sizes.

Table 22 summarizes the key within-group meta-analytic results. Table 23 summarizes the key between-group findings.

Table 22

Summary of Key Hypothesis Within-Group Findings

Hypothesis	Level 2	Level 3	Magnitude	Variance % explained by error	True moderator exists?	CI overlaps zero? ^a	File drawer bias?
H1. Overall effect			-.26	26	yes	yes	no
	<i>Group-based stereotype</i>						
		Race/ethnicity stereotype	-.32	33	yes	yes	no
		Female stereotype	-.21	25	yes	yes	No
H2. Moderator: ST-activating cues							
		STA blatant	-.29	22	yes	yes	no
		STA explicit	-.25	44	yes	yes	no
		STA subtle	-.25	24	yes	yes	no
		<i>Race/ethnicity stereotype</i>					
		STA blatant	-.41	74	no	no	yes
		STA explicit	-.64	100	no	no	no
		STA subtle	-.22	25	yes	yes	no
		<i>Female stereotype</i>					
		STA blatant	-.17	18	yes	yes	no
		STA explicit	-.18	40	yes	yes	no
		STA subtle	-.24	27	yes	yes	no
H3. Moderator: ST-removing strategies							
		STR explicit	-.24	28	yes	yes	no
		STR subtle	-.35	23	yes	yes	no
		<i>Race/ethnicity stereotype</i>					
		STR explicit	-.80	100	no	no	yes
		STR subtle	-.38	28	yes	yes	no
		<i>Female stereotype</i>					
		STR explicit	-.14	27	yes	yes	no
		STR subtle	-.33	20	yes	yes	no
H4. Moderator: Domain identification							
		High DI	-.32	49	yes	yes	no
		Medium DI	-.37	41	yes	yes	yes
		Low DI	-.11	27	yes	yes	yes
		<i>Female stereotype</i>					
		High DI	-.29	48	yes	yes	yes
		Medium DI	-.52	59	yes	yes	yes
		Low DI	-.11	27	yes	yes	yes
H5. Moderator: Test difficulty							
		High difficulty	-.39	23	yes	yes	no
		Medium difficulty	-.19	41	yes	yes	no
		Low difficulty	.08	60	yes	yes	yes
		<i>Race/ethnicity stereotype</i>					
		High difficult	-.43	58	yes	no	no
		Medium diff.	-.18	86	no	no	yes
		<i>Female stereotype</i>					
		High difficult	-.36	18	yes	yes	no
		Medium diff	-.18	30	yes	yes	yes
		Low difficult	.08	60	yes	yes	yes

Note. ST = Stereotype threat. STA = Stereotype threat activation conditions. STR = Stereotype threat removal conditions. DI = Domain identification. ^a When the 90% credibility interval overlaps zero, a mean effect size estimate is considered inconclusive and should not be interpreted as providing meta-analytic evidence to support the hypothesis of interest.

Table 23

Summary of Between-Group Findings

Level 1	Level 2	Magnitude	% explained by error	True moderator exists?	CI overlaps zero? ^a	File drawer bias?
<i>Experimental condition</i>						
	Control condition	-.44	36	yes	no	no
	STA condition	-.53	28	yes	no	no
	STR condition	-.28	55	yes	yes	no
<i>Minorities-Majority</i>						
	Control condition	-.56	100	no	no	no
	STA condition	-.69	62	yes	no	no
	STR condition	-.38	38	yes	yes	no
<i>Women-Men</i>						
	Control condition	-.26	100	no	no	possible
	STA condition	-.39	26	yes	yes	no
	STR condition	-.23	73	no	no	no

Note. STA = Stereotype threat activation conditions. STR = Stereotype threat removal conditions. ^a When the 90% credibility interval overlaps zero, a mean effect size estimate is considered inconclusive and should not be interpreted as providing evidence to support the hypothesis of interest.

Chapter 4

DISCUSSION

The present dissertation aims at providing a qualitative and quantitative review of stereotype threat effects on target test takers' cognitive ability test performance. Partially replicating and extending the meta-analytic paradigm on stereotype threat effects by Walton and Cohen (2003), I conducted a series of meta-analyses to investigate the general stereotype threat effects on stereotyped social groups' test performance, the exploratory effects across levels of group-based stereotypes, as well as several hypothesized moderator effects.

I operationally defined the target dependent measure of cognitive ability test performance and moderators of stereotype threat effects somewhat differently than Walton and Cohen (2003), taking into account empirical evidence in the literature as well as other broad stereotype theories. Further, there was only a percentage of the data set that overlapped with the studies meta-analyzed by Walton and Cohen ($k = 23$ or 20%). Therefore, meta-analytic findings in the present manuscript were only a partial replication of those in Walton and Cohen's meta-analysis of stereotype threat effects.

In this chapter, I discuss (1) the theoretical implications of the meta-analytic findings as well as implications for future research based on these findings, (2) some practical implications of the meta-analytic results, including implications for test takers in educational and employment testing settings, and (3) the limitations of the present meta-analytic review.

Theoretical Implications and Implications for Research

Within-Group Meta-Analytic Findings

Based on this meta-analysis integrating 10 years of research on stereotype threat effects on stereotyped test takers' cognitive ability test performance, there seems to be no affirmative answer to the question of whether stereotyped test takers' performance generally suffers from a situational stereotype threat, at least at the overall level.

Although the overall mean effect size of $-.26$ may be suggestive of the existence of an effect consistent with the theory of stereotype threat, the wide variability across studies (i.e., one fourth of studies showing zero or positive effects) indicated that there were true moderators further explaining the variance of effect size estimates over and above study artifacts. In other words, interpreting this overall mean effect size as supportive evidence for overall stereotype threat effects in the literature without first considering other moderators would lead to misleading conclusions.

Based on Hunter and Schmidt's (1990) recommendations, all moderators were meta-analyzed in a hierarchical order as fully as possible (i.e., levels of one moderator nested in levels of another moderator). Therefore, subsequent result interpretation would focus on key meta-analytic results: those at the lower-order level of these hierarchical moderating relationships.

Moderator Meta-Analytic Findings

Stereotype threat theorists propose that stereotype threat activation does not equally affect all members of a stereotyped group. Those who strongly identify themselves with a cognitive ability domain are the most motivated to avoid confirming a negative group-based stereotype; they are thus ironically affected by a situational

stereotype threat in terms of their test performance. Further, stereotyped individuals may perform well on a non-challenging task or test; their performance only suffers when a cognitive ability test is at the upper level of their cognitive capability because only at this level that an activated stereotype can interfere with their true ability. Therefore, domain identification and test difficulty were proposed and meta-analyzed as conceptual moderators in the present review. Methodologically, stereotype threat effects may also be contingent on how a group-based stereotype is either presented to test takers (as in experimental manipulation) or removed from a cognitive ability testing situation.

Before investigating these hypothesized moderators, I first considered a descriptive factor that might mitigate mean effect sizes: *group-based stereotypes*. Although there were no specific theoretical grounds to suspect that different types of negative stereotypes employed to present a threatening testing situation to certain target social groups might play a moderating role, anecdotal evidence and logical reasons led to a closer meta-analytic examination of whether group-based stereotypes were confounded with stereotype threat in influencing test takers' scores. Specifically, gender and ethnicity of target test takers are differentially correlated with mathematic/cognitive ability test performance in the broad testing literature, and mean effect size estimates seem to differentially vary across levels of group-based stereotypes in the stereotype threat literature. Therefore, different types of stereotype targeting different social groups might yield variability in mean effect sizes across social groups. The meta-analytic findings subsequently supported this theory.

Specifically, the data set was split into two subsets by type of group-based stereotypes. These subsets were meta-analyzed for within-group mean effect sizes. Under

situational stereotype threat, minority test takers tended to perform poorly on various cognitive ability tests compared with themselves (mean $d = -.32$), whereas stereotype threatened women also performed worse on math tests than non-threatened women but to a lesser extent than minorities (mean $d = -.21$). The zero-included credibility intervals of these mean effects did not allow a direct interpretation though.

Other hypothesized moderators were further analyzed hierarchically. In terms of the methodological moderator of *stereotype threat-activating cues* (“blatant,” “moderately explicit” and “subtle”), some conclusive results on stereotype threat effects were found for minority test takers only, whereas no affirmative answers were found for women test takers.

Examining the detected mean effects, moderately explicit threat-activating cues produced a larger mean effect size ($-.64$) than even a direct, blatant way to convey a negative stereotype to minority test takers ($-.41$). As hypothesized, these meta-analytic findings lend partial credence to the theory of stereotype reactance which posits that stereotyped individuals may perceive a blatant negative stereotype as a limit to their freedom and ability to perform, thereby ironically invoking behaviors that are *inconsistent* with the stereotype (see Kray, et al., 2001). By definition, the presentation mode of moderate explicitness is direct enough to make individuals aware of a negative stereotype in the testing environment and thus getting distracted, but the message is not blatant enough to invoke behaviors that are inconsistent of the stereotype (e.g., motivating targets to overperform). The credibility intervals of these mean effects did not overlap zero (i.e., no null findings or positive findings in these subsets); study artifacts

also explained all or most of the variance in the subsets. In other words, these findings may be considered conclusive.

Subtle cues were the weakest presentation mode in producing stereotype threat effects on minorities' test performance (mean $d = -.22$). At the mean effect level, this finding seemed to be in disagreement with Levy's (1996) theory about how subtle priming of negative stereotypes might have a more direct and stronger effect on task performance through the activation of associated behavioral tendencies. Instead, this finding tended to be in line with Walton and Cohen's (2003) meta-analytic result regarding weak stereotype threat effects caused by "implicit" primes (i.e., the more implicit or subtle a stereotype threat cue is, the weaker stereotype threat effects are). However, the zero-overlapping credibility interval of this mean effect made the finding inconclusive because one cannot rule out the possibility that there might be no stereotype threat effects when subtle stereotype threat-activating cues were employed in experiments.

At the mean effect level, for women test takers, explicit threat-activation cues (blatant and moderate) generally produced smaller mean effect sizes than subtle cues (- .18, -.17, and -.24, respectively), supporting Levy's (1996) position that explicit threat activation may weaken the effect of group-based stereotypes because it indirectly affects task or test performance through some psychological mediators. However, all credibility intervals of the mean effects overlapped zero, making these findings inconclusive. In other words, regardless of the explicitness degree in manipulated threat-activation cues, one cannot rule out the possibility that under certain circumstances, stereotype threat

would not diminish women's math performance, or threat cues might even enhance the performance of interest.

In terms of *stereotype threat-removing strategies*, I kept Walton and Cohen's (2003) classification scheme of "explicit" and "subtle" modes, but redefined these categories in that "explicit" strategies meant any interventions that explicitly refuted the negative stereotype message, and "subtle" strategies referred to either implicit interventions or some subtle manipulation of test/task purpose statements. Again, minorities and women reacted differently to levels of threat-removals at the mean effect level.

On the one hand, for minorities, explicit strategies unexpectedly increased stereotype threat effects to a larger magnitude (mean $d = -.80$) compared with subtle strategies (mean $d = -.34$). One possible explanation is that, for minorities, implementing explicit interventions such as telling test takers outright that minorities perform better than Whites on a certain cognitive ability test may accomplish the same thing as stereotype threat could, with a twist: the negative effect of performance interference. Direct and explicit statements about how good minorities can be in intellectual performance situations may raise an illogical fear or performance pressure for these individuals: should they do poorly, they would *not* be able to confirm the positive in-group image associated with a particular cognitive ability test(s). Therefore, their performance might suffer—probably because of the same psychological mechanisms as those of a "model minority" status, which can sometimes invoke debilitated intellectual performance among target minorities (see Cheryan & Bodenhausent, 2000 for empirical evidence of model minority effects).

On the other hand, for women, explicit threat-removal strategies were more effective than subtle ones in reducing stereotype threat effects (mean $d = -.14$ and $-.33$, respectively). The meta-analytic results seem to support the theory of stereotype susceptibility (Shih, et al., 1999), positing that the direct activation of a positive in-group stereotype (e.g., introducing a statement that Asians are superior at math; women are better on a specific math test than men) might cause a performance boost for women test takers.

However, except for the finding regarding mean stereotype threat effect on minorities' test performance in the condition of explicit stereotype threat-removing strategies, all other credibility intervals of the above mean effect estimates overlapped zero. In other words, for the moderating category of threat removals, one can be only conclusive about the finding that explicit interventions aiming at refuting a negative stereotype about minorities' intellectual inferiority might backfire, inadvertently worsening stereotype threat effects instead of alleviating them.

Why does stereotype threat, either activated or refuted, differentially affect minorities than women? Do minority test takers as a social group possess some unique characteristics that women do not (or not at the same intensity)? Do stereotype threat presentation modes of activation or removal somehow tap into these unique characteristics (or not) and invoke differential behavioral reactions between minorities and women?

One such group-based characteristic may be minorities' race-based learned expectations of rejection, or rejection sensitivity. Rejection sensitivity is defined "as a cognitive-affective processing dynamic (Mischel & Shoda, 1995) whereby people

anxiously expect, readily perceive, and intensely react to rejection in situations in which rejection is possible” (Mendoza-Denton, Downey, Davis, Purdie, & Pietrzak, 2002; p. 897). Race-based rejection expectations among ethnic minorities are originated in a lifetime history of being subjected to group-based discrimination, mistreatment, prejudice and exclusion from a certain salient domain (e.g., higher academic education), either directly or vicariously (c.f., Essed, 1991). Under certain circumstances where the outcome is important and where one would possibly experience rejection based on one’s group membership (i.e., in a personally salient and self-applicable situation; c.f. Higgins, 1996), one would more readily recognize and/or interpret these situational factors as acceptance-rejection cues than out-group members without the same life experiences. In other words, these situational cues would trigger an intense reaction in terms of affect, cognition, and behavior, unique to members of a stigmatized group (Mendoza-Denton, et al., 2002).

Pietrzak, Downey, and Ayduk (2005) conceptualize rejection sensitivity as a defensive motivational system or a “better safe than sorry” self-preservation strategy. According to Pietrzak et al.’s model, individuals who are high on reaction sensitivity are vulnerable to even mild situational threatening cues (i.e., hypersensitive) whereas those low on rejection sensitivity do not perceive such cues as personally threatening. Highly rejection sensitive individuals may overreact to protect themselves from realizing the perceived rejections. This cognitive-motivational process is activated *automatically* (i.e., without one’s conscious awareness). It can be inferred that, at the risk of realizing a perceived status-based rejection, members of a socially stigmatized subgroup are more likely to become hypersensitive and overreacting than members of a mainstream

subgroup. In fact, Mendoza-Denton, Page-Gould, and Pietrzak (2005) identify the construct of status-based rejection expectations as an important aspect of stereotype threat theory as well as playing a central role in other stigmatization-related dynamics (e.g., stigmatization; Miller & Kaiser, 2001; stigma conscientiousness; Pinel, 1999; rejection sensitivity; Pietrzak, 2004).

The implication is that minority test takers may have a higher tendency of rejection sensitivity than female test takers. Therefore, the more explicit situational cues were, regardless whether they were threat-activating cues or removing ones, the more strongly minorities might react to the cues, producing greater stereotype threat effects. In the literature, a single study found that rejection sensitivity at the individual level was *not* significantly correlated with Black students' cognitive ability test performance under stereotype threat (Williams, 2004). However, there needs to be more research on the relationship of interest to yield a more conclusive understanding and facilitate a future quantitative review. (As a cautionary note, one might not want to read too much into the findings regarding the effect of explicit threat removals on minority test takers' performance because the subset of *d*-values meta-analyzed is small.)

In terms of *domain identification*, the stereotype threat conceptual framework and previous empirical evidence suggest a linear relationship between higher level of intellectual domain identification and stronger stereotype threat effects (or lower cognitive ability test performance). At the mean effect level, meta-analytic findings in the present study showed a non-linear relationship among the variables of interest for women test takers. Specifically, stereotype threat surprisingly affected moderately math identified women more severely (mean $d = -.52$) than highly identified women (mean $d =$

-.29). Low math identified women suffered the least from stereotype threat though (mean $d = -.11$). Again, one cannot be certain about these findings given their zero-included credibility intervals.

Supposing that the true mean effect sizes had been in the hypothesized direction (i.e., negative findings), one explanation for such findings would be that certain highly domain identified women might not be strongly concerned about a dominance status in mathematic ability. In other words, they were not influenced by a fear induced by situational stereotype threat activation. This explanation is based on Josephs, et al.'s (2003) empirical evidence on the relationship between women's tendency in dominance status concern, mathematic identification, and stereotype threat effects. I further speculate that women who were moderately identified themselves with the math ability domain might be more likely than high identifiers to prove themselves and disconfirm the negative stereotype activated, thus experiencing greater stereotype threat effects. Theory-wise, these findings might cast some doubts on one of the most accepted boundary conditions of stereotype threat effects. Research-wise, the meta-analytic results imply that some stereotype threat researchers might have inadvertently lost informative data when pursuing the strongest experimental design possible by purposefully selecting only high domain identifiers for their studies.

Another explanation might be in the inconsistency in operational definitions of domain identification in the literature. Stereotype threat researchers tended to arbitrarily and inconsistently define and categorize domain identification levels of stereotyped test takers. For example, test takers' domain identification might be directly assessed using self-report measures (e.g., Brown & Pinel, 2003; Spicer, 1999), indirectly inferred from

objective measures such as standardized cognitive ability test scores (e.g., Anderson, 2001; Quinn & Spencer, 2001; Schmader & Johns, 2003), or via both approaches (e.g., Davies, et al., 2002; Harder, 1999). This is a pre-existing research design limitation beyond the scope of the present meta-analysis; therefore, one should be cautious in generalizing the interpretation of these findings. Future research needs to reach a consensus on the operational definition of domain identification.

Note that defining individuals' domain identification indirectly from their prior standardized cognitive ability test scores may be problematic. The performance interference hypothesis would predict that a negative stereotype may negatively affect target stereotyped individuals in a highly diagnostic testing situation (e.g., a high-stake standardized cognitive ability test). Prescreening participants based on their good performance on prior tests in the hope that these high performers would subsequently underperform on another cognitive ability test may result not only in a restriction of range but also in a circularity of conceptualizing of the construct.

There are insufficient *d*-values in the minority subset to meta-analyze. However, given the previous patterns of findings for stereotyped minority test takers, I speculate that intellectual domain identification might moderate stereotype threat effects on intellectual test performance of minorities in the predicted direction and magnitudes, assuming that design artifacts do not confound with the experimental effects.

Could stereotype threat effects be manifested only when *test difficulty* level is high? Stereotype threat theory predicts so; the conceptual rationale is that only at the challenging, upper-bound level of their ability do target stereotyped group members underperform. The “choking under pressure” theory in cognitive psychology further

provides a possible explanation for the moderating effect of test difficulty on stereotype threat effects. Beilock and Carr (2001, 2005) found that the performance pressure of an actual cognitive ability testing situation might interact with the cognition-demanding level of test items to influence participants' performance, such that performance pressure "choked" participants' performance on highly cognition-demanding math problems but not on simpler, less cognition-demanding math items. This "choking under pressure" phenomenon was further explained by individual differences in working memory capacity: only participants with high working memory capacity would experience the suboptimal performance when solving highly cognition-demanding math problems (i.e., performing better than those with low working memory capacity on these math items in low pressure conditions, but performing as poorly as the latter in high pressure conditions). The researchers attributed the findings to the fact that performance pressure associated with high cognition-demanding mental tasks may induce worries about the situation and its consequences, thereby reducing working memory capacity available for performance (distraction theories; c.f., Lewis & Linder, 1997). Note that Engle (2002) defined working memory capacity as the ability to control attention to maintain (relevant) or suppress (irrelevant) information. In other words, performance failure due to insufficient working memory capacity is a loss of ability to suppress irrelevant information (i.e., worries, distractions).

Integrating the cognitive theory and stereotype threat theory concerning the moderating role of test difficulty, one may speculate that being exposed to both a performance pressure situation (e.g., taking a test) and a cognition-demanding task (e.g., difficult, complex test items), any test taker may experience more diminished test

performance than when undertaking a less cognition-demanding test. However, when a negative stereotype is further introduced into the high-difficulty testing situation, consequently compounding the degree of performance pressure for target individuals, members of the stereotyped group may suffer from a performance failure to a greater extent than that of comparison group members, because the former may suffer more from reduced executive attention caused by the extra stimulus of stereotype threat priming. On the other hands, low cognition-demanding mental tasks (e.g., easy tests) do not deplete working memory capacity in the first place, which may leave targets sufficient executive attention to handle the extra pressure of stereotype threat activation effectively (e.g., not being overly worried or distracted). Therefore, stereotype threat effects would be less likely to manifest under these circumstances.

In the present meta-analysis, in terms of test difficulty, at the mean effect level, the hypothesis seemed to be supported for both minorities and women (i.e., the more difficult tests, the larger effect sizes). However, for women, the zero-overlapping credibility intervals rendered these mean effect findings inconclusive. In other words, with difficult tests, the greatest stereotype threat effects were found for minorities and most of the variance in the data was accounted for. For women, one could not rule out the likelihood of null findings or positive findings, thus these mean effects should not be interpreted as supporting evidence.

Aside from the inconclusiveness of some findings, one should also note the methodological inconsistency in the operationalization of test difficulty levels in the literature. For example, most researchers selected a complex subsection of a standardized cognitive ability test (e.g., GRE Calculus only; Aronson, et al., 1999) or an established

“challenging” test or subtest (e.g., Canadian Math Competition; Ambady, et al., 2004) based on an assumption that such a subtest or test was at the upper ability level of test takers. Alternatively, simpler types of cognitive ability tests might constitute “moderately difficult” or “easy” and were labeled as such in research reports (e.g., algebra, trigonometry, & geometry; O’Brien & Crandall, 2003; Spencer, et al., 1999). A problem with this approach is that the construct of test difficulty might be confounded by type of math tests: it remains unclear in these studies whether stereotype threat effects were manifested at a high level of difficulty, or they were observed with certain types of cognitive ability problems (e.g., advanced calculus) but not with other types.

Some other researchers assessed test difficulty levels post-hoc; that is, researchers reviewed a sample’s test score distributions to see whether or not a test was challenging enough to the study sample and reported as such (e.g., certain proportions of test takers answering a test correctly; Schmader & Johns, 2003; Wicherts, et al., 2005). This practice is more desirable than the previous one because readers are able to find how test difficulty levels are quantified in research. However, the best methodological approach is pilot-testing test items with a sample of similar characteristics to the study sample, which a few researchers undertook (e.g., Tagler, 2003). Future studies in this area may need to adopt a methodologically sound and consistent way to define test difficulty.

Other Potential Conceptual Moderators

Because of the scope of the present meta-analytic review, the conceptual salience of hypothesized variables, and the availability of relevant information (or lack of) in study reports, I was able to hypothesize and examine only two key conceptual moderators of domain identification and test difficulty, and three methodological moderators

(subgroup stereotypes; stereotype threat activating cues, and stereotype threat removing strategies) in the present review. The meta-analytic results showed that sampling error explained between 18 percent and 100 percent of the cross-study variation in various subsets of stereotype threat effect sizes across levels of these variables; therefore, there would still be other potential conceptual and methodological moderators that could have explained more stereotype threat effect variance had they been hypothesized and meta-analyzed. (As mentioned, most of the variance of d -values was substantially less than 75%.) As the literature expands in the future, another quantitative review may further contribute to theoretical knowledge of stereotype threat by investigating several other social psychological variables relevant to stereotyped test takers' characteristics, such as defense mechanisms and social group identity (e.g., racial identity; gender identity).

Defense mechanisms. Stereotype threat theorists posit that members of a stereotyped group may engage in certain defense mechanisms against a group-based stereotypic threat; for example, Blacks under stereotype threat influence may disidentify themselves from the domain of intellectual ability (see Steele & Aronson, 1995). Further, stereotyped group members who discount feedback on cognitive ability tests as not reflective of their intelligence may become disengaged from the academic or intelligence testing domain (Major, Feinstein, & Crocker, 1994; Schmader, Major, & Granzow, 2001). Although these tenets constitute a prediction of long-term effects of stereotype threat for stereotyped group members as far as intellectual domains and/or cognitive ability tests are concerned, individuals' tendency of engaging in defense mechanisms, such as intellectual disidentification, discounting and disengagement, might mitigate the effects of stereotype threat on cognitive test performance as a short-term effect. Those

who are high on these defense mechanisms are ironically buffered from the negative influence of a situational stereotype threat (i.e., less likely to experience performance interference) as compared with those who do not engage in these mechanisms, simply because the former do not care much about test performance outcomes and, thus, are not faced with the fear of confirming a group-based negative stereotype activated in an evaluative environment.

Little empirical evidence has been found to support this hypothesis so far. In Nguyen, et al. (2004), female test takers' defense mechanisms (i.e., discounting; disengagement) did not interact with situational stereotype threat to affect math test performance. Nevertheless, future research may incorporate this variable as a conceptual moderator for stereotype threat effects.

Steele, et al. (2002) emphasize the role of multiple social identities in invoking stereotype threat experience and behavior effects. The theorists write, "in particular settings or domains of activity, a person can come to realize that they could be devalued, marginalized or discriminated against, based on one of these identities [sex, age, race, ethnicity, social class, religion, professional identity, etc.] Once this realization happens, we assume that the person becomes vigilant to the possibility of identity threat in the setting..." (pp. 416-417). In other words, given a relevant social identity and threatening contextual cues, individuals may become vigilant in searching for evidence of identity threat, but at the same time they also experience a strong motivation to disconfirming the salient social identity threat. These conflicting motivations subsequently distract individuals and thus undermine their performance in evaluative settings.

Two types of social identity that are most conceptually relevant to the present review and empirical evidence that either refutes or supports the role of these identities as potential moderators of stereotype threat effects are discussed next.

Race identity. Race identity or racial identification refers to individuals' tendency to view race as a core aspect of the self-concept (see Sellers, Rowley, Chavous, Shelton, & Smith, 1997). In the literature, this tendency has been empirically verified to predict how minorities interact with in-group and out-group majority members (e.g., Rowley, Sellers, Chavous, & Smith, 1997; Sellers, et al., 1997). The more strongly African American students see race as central to their self-image, the more likely they would make race-related attributions to racially ambiguous situations (Shelton & Sellers, 2000). Strongly racially identified individuals also tend to perceive others' racially ambiguous behaviors towards them as race-related and respond as such (e.g., Lalonde & Cameron, 1994; Shelton & Sellers, 2000).

In the conceptual framework of stereotype threat, race identity may mitigate effects of stereotype threat on minorities' cognitive ability test performance, in that the effects may be enhanced for strong racial identifiers because they are more likely to detect even ambiguous situational cues of stereotype threat in a diagnostic testing setting, and because these individuals may be quick to be aware of the threat of being judged stereotypically. In other words, subtle stereotype threat-activating cues may be sufficient to invoke the predicted reactions among highly race identified individuals. On the other hand, when stereotype threat-activating cues are subtle, weak racial identifiers may be spared stereotype threat effects because they are less likely than strong identifiers to attribute ambiguous situational cues to a group-based intellectual stereotype, and/or to

perceive that their intellectual self-images are threatened. Nevertheless, when explicit stereotype threat cues are introduced to a testing situation, moderately or blatantly, even weak racial identified individuals may experience a situational pressure and subsequently underperform on tests. In other words, even though levels of racial identity may buffer some individuals from stereotype threat effects, too strong a situational signal might override the effect of this individual characteristic.

So far, empirical evidence regarding race identity as a moderator of stereotype threat effects on minorities' cognitive test ability performance is scarce. Most researchers did not find a significant interaction between race identity and stereotype threat to support the hypothesis (see Table 3a for citations). The potential contribution of a future meta-analytic review is to present a clear picture of whether the magnitudes of stereotype threat effect on minorities' cognitive test performance change as a function of race identity levels.

Gender identity. In the same vein as race identity, gender identity can be conceptualized as a potential moderator of stereotype threat effects on women's mathematic ability performance. Interestingly, there are competitive theories specifically explaining women's performance in math and/or science domains that provide alternative predictions of how gender identity may fit into the theoretical framework of stereotype threat effects.

A prediction which is in line with the above race identity assumptions and stereotype threat tenets is that high gender identification would motivate women to maintain a positive group image in the eyes of others, thus making them vulnerable to stereotype threat (Schmader, 2002). Therefore, the more highly gender identified women

are, the more likely they are to experience threat effects (i.e., handicapped performance). This hypothesis has been investigated. For example, Schmader (2002) found a significant three-way interaction between gender, gender identity relevance (linking gender group identity to math test performance—stereotype threat manipulation) and gender identification (how central gender group membership was to individuals). Subsequent simple slopes analyses supported the researcher's hypothesis. The gender identity relevance manipulation affected only the performance of women who tended to be highly identified with their gender, but not for those with low gender identification. Men were not affected by identity condition regardless of level of gender identification. Therefore, Schmader concluded that individual differences in gender identification moderated stereotype threat effects for women, such that higher identified individuals showed more performance decrements. However, some other researchers were not able to replicate these results (see Table 3a for citations).

The imbalance-dissonance principle of social cognition theory alternatively posits that the degree of gender group identification (e.g., "female" identity) predicts the extent to which women disassociate themselves with a (math) domain socially categorized as non-female (see Nosek, Banaji, & Greenwald, 2002 for evidence). In other words, the more women consider gender group membership important to their self-identity, the more they would move away from incongruent situations where their identity is threatened (i.e., taking math courses and tests), and the less they would invest in a math domain (i.e., becoming low math identification). Although this hypothesis has never been investigated in a stereotype threat paradigm, the theoretical implication is that a higher level of gender identification as defined here might *buffer* women from the debilitating effects of

stereotype threat invoked in math testing situations (i.e., the more highly identified, the less likely stereotype threat effects are detected). Would this prediction also hold for ethnic minority groups? To answer this question, one might first want to answer the questions of whether or not, and/or to what extent intellectual abilities are socially categorized as “non-Black” or “non-Hispanic,” such that a high level of race/ethnic identification might equate to a disidentification with intellectual domains.

Between-Group Meta-Analytic Findings

As mentioned above, stereotype threat theory is commonly believed as providing a partial explanation of the observed between-group gaps in cognitive ability test performance (e.g., minorities vs. majority; women vs. men). This relationship has been directly investigated in the literature (e.g., Keller, 2002; Kray, et al., 2002; McFarland, et al., 2003).

On the one hand, the stereotype threat research assumptions would be that (a) the subgroup performance gaps in standardized cognitive ability testing are partially due to stereotype threat effects and (b) removing or refuting stereotype threat reduces or even eliminates these performance gaps (i.e., women performing equally well as men on difficult math tests; minority test takers performing equally well as Whites on various cognitive ability tests). On the other hand, Sackett and colleagues (2001; 2003; 2005) propose that any observed reduction in mean test score differences between minority and majority test takers under stereotype threat may be a product of artificial experimental treatments of stereotype threat and/or of how cognitive ability test performance is analyzed (i.e., mean scores controlled for prior cognitive ability). The meta-analytic results in the present review seemed to support Sackett, et al.'s position.

At the mean effect level, the minority-majority mean effect sizes were greater in stereotype threat-activated conditions (mean $d = -.69$) than in test-only control conditions (mean $d = -.56$). (There might be other moderators further explaining the variance in the data related to threat activation conditions though.) Because the finding related to test-only conditions relatively reflects the existing cognitive ability test score gap between these social groups in the broad educational and employment testing literature, activating stereotype threat in a laboratory experimental setting seemed to introduce another dose of pressure that might artificially widen the observed test score gap between Whites and minorities, which was consistent with Sackett, et al.'s argument (2001). At the mean effect level, implementing some threat-removed strategies reduced the observed performance gap (mean $d = -.38$), which is consistent with the prediction of stereotype threat theorists; however, the substantial variability in the data set yielded this finding as inconclusive in terms of meta-analytic evidence for stereotype threat effects across the subset of studies at the present time.

Furthermore, the female-male mean math test score differences were practically unchanged from test-only control conditions (mean $d = -.26$) to stereotype threat-removed conditions (mean $d = -.23$) and both findings were conclusive and meaningful. The former finding reflects the real-world observation that women do not perform as well as men on math tests in general, and trying to implement some stereotype threat-removing strategies to change this picture did not seem to work across studies in this data set. No other moderators would further explain the variance in the data of effect size estimates of interest. Although the mean effect size was slightly increased in stereotype threat-activation conditions (mean $d = -.39$), not only would further moderating meta-analyses

be needed to explain the wide variance in this subset of data, but also one cannot rule out the possibility that priming stereotype threat might have no effects or even positive effects on women's math performance across studies. Again, convergence evidence seemed to be consistent with Sackett and colleagues' (2001; 2005) argument that between-group stereotype threat effects might be an artificial product of laboratory experiments. At least, these meta-analytic findings demonstrated that one cannot be very conclusive about what portion of the between-group difference in cognitive ability scores is due to stereotype threat.

As previously mentioned, a lack of theoretical rationales in the theory of stereotype threat does not facilitate a conceptualization of moderating roles of key constructs such as domain identification and test difficulty in the between-group relation of stereotype threat to test performance. Regarding domain identification, cross-study meta-analytic findings in this study showed that high math identified women might or might not suffer from stereotype threat in terms of math performance. Logically, the observed math score difference between men and women (or Whites and minorities) might not be substantial or significant among highly domain identified individuals (as inferred from Cullen, et al.'s 2004 correlational findings). Therefore, introducing a situational stereotype threat into a testing environment would not necessarily be detrimental to highly math identified women's test performance, whereas it might inadvertently boost men's math performance and possibly confound any detected effects. Regarding test difficulty, in this study no meaningful moderating effects were found for within-group test performance under stereotype threat given the wide variability in the data not explained by this moderator in this study. However, at the mean effect level

only, there is a possibility that between-group relationships might mirror within-group relationships, such as the more difficult a cognitive ability test is, the more stereotype threat activation might distract target test takers (i.e., reducing their executive attention to the task at hand), which in turn worsens targets' test performance and increases the score gaps. These research questions need to be empirically tested in future experiments and/or meta-analytically investigated in a future quantitative review.

Practical Implications

As stereotype threat theorists acknowledge, the theory is first and foremost derived from a practical question, "Do social psychological processes play a significant role in the academic underperformance of certain minority groups, and if so, what is the nature of those processes?" (p. 379; Steele, et al., 2002). Therefore, practical implications of the meta-analytic results in the present study may be as fascinating as (or even more than) theoretical ones.

Existence of Stereotype Threat Effects

Do stereotype threat effects exist and, if yes, in what condition would they be manifested? Cross-study meta-analytic findings in the present study did not provide strong evidence to support most of the theory-based within-group predictions. Therefore, some skeptical readers might conclude that there seems to be no meaningful effect existing across studies, even after several key moderators have been taken into account.

Even the few conclusive meta-analytic results regarding the negative effects of blatantly explicit and moderately explicit stereotype threat-activation on minorities' cognitive ability test scores may not hold much interest for some readers as far as their practical implications are concerned. Common sense and normative practices in testing

procedures, particularly in high-stakes and standardized testing situations, typically dictate that a cognitive ability testing setting is devoid of any overt effort to divert test takers' attention away from the test or draw their attention to existing racial group test performance differences. That means, even if stereotype threat effects are actual phenomena in experimental settings, these effects should be rare occurrences in real-life testing situations given the antecedents.

Experimental conditions where subtle threat cues are used could have provided useful information for test program administrators if the hypothesized stereotype threat effects had been meaningfully detected across studies. Even if a meaningful effect were found, the control environment of laboratory studies might still limit the generalizability of stereotype threat meta-analytic findings. Although stand-alone subtle cues of stereotype threat could be pronounced and noticeable in experimental studies, they would have to compete with other salient factors for test takers' attention in real-life testing situations. Therefore, a negative message conveyed by a subtle stereotype cue(s) might be diluted or barely registered in targets' mind; consequently, the effects might be weaker than those found in this study.

One fascinating finding, albeit counter-intuitive and contradictory to stereotype threat rationales, is that explicit stereotype threat-removing strategies might do more harm than good to minority test takers (a result consistent with that in Walton & Cohen's 2003). In other words, stereotype threat effects were manifested for minorities where they should have been alleviated or eliminated. The implication is that any well-intentioned intervention programs of a "quick-fix" nature may fail, or even serve as a catalyst furthering stereotype threat effects. For test program administrators who are concerned

about potential stereotype threat effects on targets' test performance, the lesson learned from these findings is that the best thing to do is perhaps doing nothing or at least nothing of a "stereotype-refuting" nature.

Although meta-analytic evidence did not support the boundary condition of domain identification on stereotype threat effects, a characteristic of cognitive ability tests, test difficulty, behaved as predicted, although only for minority test takers. Generally, when tests are increasingly difficult, ethnic minority test takers under stereotype threat tend to experience more diminished performance, compared with the performance of those in non-stereotype threat conditions. These findings have several practical implications.

First, the meta-analytic results indicate that taking highly challenging tests is the most likely predictor of stereotype threat effects for minority test takers (and for women to a less certain extent). These findings have implications for institutions and/or organizations who use highly difficult screen-in cognitive ability tests to select top candidates (e.g., for prestigious scholarships; for top employment positions). There is a possibility that ethnic minorities or women might underperform on these tests compared with their true ability due to some subtle stereotype threat cues in the testing environment, whereas the performance of their competitors might not be negatively affected. Therefore, such high-stakes decisions should not be made solely based on candidates' cognitive ability test scores.

Second, stereotype threat effects are also manifested when cognitive ability tests are only moderately difficult (e.g., consisting of mixed difficult items); this finding has real-life implications because it may apply to a category of actual tests used in

educational or employment testing settings, such as the GRE, SAT, Raven Advanced Progressive Matrices, and the Wonderlic Personnel Test. Even when a mean stereotype threat effect size d is as small as .18 across studies, it still indicates that almost a one-fifth standard deviation separates two group means (e.g., stereotyped-activated group means vs. stereotype-removed group means). If a minority student took the SAT and his or her true cognitive ability were at the national mean level, he or she might underperform by about 50 points due to stereotype threat effects. Combining this fact with the meta-analytic results concerning situational cues and threat removal strategies, at-risk test takers (stereotyped group members) should be aware of the possibility of underperformance in high-stakes testing situations.

Third, minority test takers can actually handle possible stereotype threat effects which may incur when a cognitive ability test is difficult or moderately difficulty. The best counter-stereotype threat strategy may be simply “practice makes perfect.” Indeed, according to research findings in cognitive psychology, repeatedly practicing a type of cognitive ability problem-solving, even those of a high cognition-demanding nature, may reduce or eliminate the detrimental effect of test difficulty pressure on test performance (see, for example, Beilock, Holt, Kulp, & Carr, 2004). The reason is because heavy practices at least may help one to resume the attention allocated to task execution (e.g., taking tests), instead of being distracted by thoughts about the testing situation and its importance. Although the moderating effect of practice on stereotype threat effects has not been directly investigated in the literature, one can logically deduce that encouraging targets to repeatedly practice solving complex, difficult type of cognitive ability test

items may indirectly reducing the likelihood of situational stereotype threat effects via reducing the perceived performance pressure and distraction in testing situations.

In sum, stereotype threat effects seem to be an elusive phenomenon in most conditions and they may be applicable to ethnic minorities sometimes, but not to women. Nevertheless, treating the meta-analytic knowledge about stereotype threat effects may need a similar approach as treating information about a medical pathology. Even when the information might be inconclusive (e.g., tests randomly coming out as negative, neutral or positive), one might still want to err on the safe side and treat the phenomenon as if it were a real occurrence, particularly when most mean effects are suggestive of the existence of stereotype threat effects and a substantial proportion of the data tends to align with the theory. However, one should be cautious about generalizing these implications or adopting applications based on these findings because of the wide variability existing in these mean effect estimates.

Magnitudes of Stereotype Threat Effects

Of importance to some readers is the magnitude of the interaction effects between situational stereotype threat and some methodological and conceptual moderators investigated in the present study. As noted above, Cohen's (1988) effect size guidelines could be used to interpret stereotype threat effects across moderators' levels. However, small effect sizes can still be practically important in the domains of educational and/or employment testing. Cohen himself cautions researchers against indiscriminating usage of the labels of "large, medium, or small" effect sizes within particular social science disciplines or topic areas. Cohen writes, "Many effects sought in personality, social, and clinical-psychological research are likely to be small... because of the attenuation in

validity of the measures employed and the subtlety of the issue frequently involved” (p. 13). In other words, smaller effect sizes may be the norm for research, even experimental, in certain areas of research, such as education (Valentine & Cooper, 2003). In organizational research, Aguinis, Beaty, Boik, and Pierce (2005) reviewed 30 years of published articles in prominent journals (Academy of Management Journal, Journal of Applied Psychology, and Personnel Psychology). Aguinis, et al. found that the median effect size for interaction effects involving a categorical moderator was only $f^2 = 0.002$ (i.e., the moderator explaining only .02% of the variance in effect sizes). Aguinis, et al. attributed the trivial finding to research design artifacts in these non-experimental studies.

A Threat Might Be in the Air, or Not?

Could it be true that the situational stereotype threat effects found in laboratory settings would also be observable in real-world testing settings? In other words, is stereotype threat inherently embedded in real-life testing situations, jeopardizing test performance of stereotyped group members as the theorists posit (Steele, 1997; Steele & Aronson, 1995; Steele & Davis, 2003)? Small subsets of effect size estimates were meta-analyzed—those from experimental stereotype threat conditions (e.g., with some subtle manipulation of threat cues such as mentioning the diagnostic nature of a test in a test direction) with those from more or less real-world testing conditions (e.g., a control condition without any special instructions). The findings thus were inconclusive.

Some results tentatively show that the answer to the “threat in the air” question might not be affirmative. Contrary to the theorists’ position, even the most subtle form of stereotype threat manipulation that is commonly believed to be almost equivalent to a

real-life testing situation was still capable of worsening target test takers' performance as compared with control conditions (mean $d = -.18$).

However, one may argue that test takers in control conditions might underperform due to stereotype threat inherent in any testing situations. Threat-removals did improve test takers' performance compared with those in control groups (mean $d = -.18$). In other words, the improvement in cognitive ability test performance among stereotyped individuals in the threat-removing conditions might be deductive evidence for the existence of stereotype threat in real-life testing situations: Only when stereotype threat is removed can members of stereotyped groups perform at their true ability level. In other words, a threat in the air might be real. Or the threat in the air phenomenon might even be contingent on other situational and/or individual factors, as evidenced in some large variance in the subsets of studies in the meta-analyses reported in this paper. Lacking sufficient sample sizes, I could not test this hypothesis meta-analytically in the present study to arrive at a better understanding.

Group-Based Stereotypes Matter

The theory of stereotype threat has been long acknowledged for its robustness and generalizability, explaining stereotyped individuals' performance on various types of ability tasks and across various group-based stereotypes. Based on the meta-analytic findings in the present review, I conclude that the results on stereotype threat effects might not be generalizable that well from one stereotyped social group to another (e.g., from female test takers to minority ones). Therefore, I suggest that stakeholders such as researchers, practitioners, educators, social policy-makers and test takers themselves should be cautious in generalizing and applying stereotype threat-based knowledge

acquired from one target social group (e.g., women performing on math tests under the influence of a math inferiority stereotype) to generate research ideas or make decisions regarding another target group in a different testing setting (e.g., ethnic minorities taking intellectual tests).

Note that cognitive ability test types may be confounded with social groups. For example, women are almost always given mathematical ability tests in a stereotype threat research paradigm. However, minorities' test performance was investigated either in a single or multiple ability domains. For example, examining Blacks' cognitive ability test performance, Steele and Aronson (1995) tested stereotype threat effects using a GRE verbal test; Smith and White (2002) used a GRE math test; whereas some other researchers employed a mixed-ability domain test (e.g., McFarland, et al., 2003; Stricker & Ward, 2004). Could different types of math tests yield different stereotype threat effects for women? Could specific domains of cognitive ability tests moderate stereotype threat effects for minorities? Unfortunately, as explained before, the issue of non-independent data points does not allow these research questions to be examined meta-analytically.

Another related note is the unbalance in subset samples: although stereotype threat theory is commonly known for its theoretical and practical implications about minorities' academic underperformance, there are many more studies investigating the gender-based stereotype of women's math deficiency than that of minority test takers in the literature. Because of small subsets of *d*-values, the interpretation of some meta-analytic findings regarding minorities may be inconclusive in the present study. Another

implication is that knowledge generated from gender-based stereotype studies may not generalize perfectly to applications involving minorities.

There are other practical implications based on the meta-analytic findings in the present review. First, there has to be some situational manipulation of stereotype threat, even subtle ones, for stereotype threat effects to be clearly detected. As stereotype threat theorists suggest, test administrators should take cautions to avoid priming stereotype threat inadvertently (e.g., not emphasizing the evaluative, high-stakes nature of cognitive ability tests in test directions; leaving demographic questions until the end or assessing the information at a different time instead of immediately before tests). Because explicit interventions of stereotype threat removals such as those used in laboratory research (e.g., forewarning test takers about possible stereotype threat effects; emphasizing the fairness of tests if true) might do minority test takers harm, whereas women tended to benefit from them, test administrators and organizations might need to exercise caution in implementing these strategies.

Second, stereotype threat-related scientific knowledge generated from research using members of a social group might not generalize well to members of a different social group because these groups may react differently to stereotype threat manipulation, at least in the case of ethnic minority test takers and females. Therefore, any social and legal interpretations of stereotype threat effects regarding subgroup performance gaps in cognitive ability test performance should take these differences into account.

Limitations

The present meta-analytic review has several limitations. First, the inclusion criteria might be defined and applied too strictly. Consequently, only about half of

empirical papers in the preliminary database were retained in the final meta-analytic set. Therefore, one might wonder about the generalizability of the meta-analytic conclusions when stereotype threat effects seem to affect a variety of outcomes besides cognitive ability test performance in the literature (e.g., other types of task outcomes; affective and/or attitudinal outcomes). However, the smaller set of *d*-values in the present review can be justified. First, the scope of the present review is focused on the most socially meaningful dependent variable, namely cognitive ability test performance, the most conceptually interesting hypothesis of “performance interference,” and the most commonly used experimental research paradigm. Second, by upholding a consistent set of decision rules in selecting studies to include in the meta-analyses, I controlled for potential variation in dependent variables, tested concepts and methodology, and thus avoiding conceptual and methodological ambiguities in conducting meta-analyses, and subsequently arriving at results which could be interpreted clearly (see Bobko, Roth, & Rotosky, 1999).

Potential biases in meta-analyses might cause findings to be inconclusive. Therefore, I conducted a vigilant process of literature search, not only relying on traditional search engines and methods to find stereotype threat reports but also taking advantage of the world wide web to locate studies, statistical information, and/or stereotype threat researchers. However, publication biases still existed in a few subsets of the data, somewhat limiting the interpretation of these meta-analytic findings. These cases were clearly described so that readers could exercise caution in drawing conclusions about the findings.

Following Hunter and Schmidt's (1990) advice, I conducted hierarchical moderator analyses as fully as the dataset allowed. The interpretation of the meta-analytic results was focused on lower-order moderating findings. For example, regarding the moderating effects of group-based stereotypes, such that minorities and women reacted differently to stereotype threat manipulation, I analyzed and interpreted all hypothesized moderators at the stereotyped group level. However, the non-zero variance in some findings still showed that further moderator analyses might be necessary; for instance, domain identification might be nested in test difficulty, which might be nested in stereotype threat-activating cue levels. Testing such full hierarchical moderator analyses is unfortunately impossible given very small data sets (or no data at all) across these moderators' levels. Similarly, there was only a small subset of studies investigating minority test takers' performance across these moderators' levels, limiting the conclusiveness of some meta-analytic results in the present review. There might be other potential moderators as discussed above which may further explain the variance in the data.

Another limitation is that I coded and analyzed only binary/categorical moderating variables (e.g., three levels of domain identification). Several studies measuring continuous domain identification were not included. However, I unsuccessfully tried to locate sufficient information to convert data in these studies into mean effect sizes for moderator meta-analyses.

Other potential methodological limitations include the treatment of meta-analytic data and study artifacts. Wherever appropriate, non-independent data points and/or large-size studies were included in meta-analyses. Consequently, I might have made a few

liberal conclusions based on study findings. However, any decision regarding data treatment in the present review was rooted in conceptual and/or methodological grounds. Readers are warned about a possible upward bias in some stereotype threat effects though. Further, following Hunter and Schmidt's (1990) advice, I used a consistent set of measurement reliability estimates as artifact distributions in all meta-analytic subsets. Because these reliability estimates ranged from acceptable to excellent values (see Table 7 above), the artifact distribution might overcorrect some variance, particularly when a subset of *d*-values was small (i.e., not accounting for much variance across studies).

As a side note, results from studies employing other group-based stereotypes (e.g., associated with ageism, social class, study majors) and/or measuring other domains of ability (e.g., athletic ability; work-related abilities; working memory capacity) were not cumulated in the present review. Therefore, it is unclear whether stereotype threat theory might or might not explain stereotyped individuals' performance in these domains, and if yes, how strongly, and under what circumstances. Given the differential patterns of findings between women and ethnic minorities in the present study, it is possible to speculate that members of other stereotyped groups might also react differently to stereotype threat, either underperforming and confirming stereotype threat theoretical predictions, or overperforming and supporting the theory of stereotype reactance effects. This research question remains to be explored in future studies.

APPENDICES

Appendix A - *Stereotype Threat-Activating Cues*

Appendix B - *Stereotype Threat-Removing Strategies*

Appendix C - *Excluded Studies*

Appendix F - *Coding Manual*

Appendix G - *Coding Form*

Appendix H - *Hypothesized Within-Group Meta-Analytic Findings from "Sensitive"
Subsets*

Appendix A

Stereotype Threat-Activating Cues

SOURCE	STEREOTYPE THREAT-ACTIVATING CUE	LEVEL ^a
Ambady Paik, Steele, Owen-Smith, & Mitchell (2004)	Gender prime: "women" word series before test	1
Ambady, et al. (2004)	Gender prime: "women" word series before test	1
Anderson (2001)	Gender prime: Demographic (gender) prior to tests	1
Aronson, et al. (1999)	Asians outnumbered Whites in math majors; Asians scored higher than Whites on math tests	3
Aronson, Lustina, Good, Keough, Steele, & Brown (1999)	Asians are better at math than Whites	3
Bailey (2004)	A handout with information favoring males	3
Brown & Day (in press)	The test measures individual's intelligence and ability	1
Brown & Josephs (1999)	Math test purpose: assessing weak performance	1
Brown & Pinel (2003)	Men and women perform differently on standardized math tests	2
Brown, Steele, & Atkins (unpublished)	Diagnostic of verbal reasoning ability	1
Cadinu, Maass, Frigerio, et al. (2003)	Racial group questionnaire before test; Whites perform better than Blacks; Belonging to a lower status group	3
Cadinu, Maass, Frigerio, Impagliazzo, & Latinotti (2003)	Women obtain lower scores on logical-math ability tasks than men.; Bar graphs showing gender differences	3
Cadinu, Maass, Rosabianca, & Kiesner (2005)	Differences in test scores between men & women	2
Cohen & Garcia (2005)	A Black peer was performing poorly on an IQ test	3
Cotting (2003)	Study to better understand what makes some people better at math and English than others	2
Dinella (2004)	Gender prime: Gender inquiry before test	1
Dinella (2004)	Gender inquiry before test; Gender differences on test	2
Dodge, Williams, & Blanton (2001)	Diagnostic test of verbal and intellectual ability; Race inquiry before test	1
Edwards (2004)	Demographic inquiry about gender before test; Study investigates gender differences in math performance	2
Elizaga & Markman (unpublished)	Diagnostic of math ability (strengths & weaknesses)	1

Foels (1998)	A test of mathematical abilities and limitations	1
Foels (2000)	The test is difficult; Diagnostic test of math ability	1
Ford, Ferguson, Brooks, & Hagadone (2004)	Diagnostic test of true ability and limitations; Gender differences in test performance	2
Gamet (2004)	A discussion about separating women and men in math classes because of men's higher math abilities	3
Gresky, Eyck, Lord, & McIntyre (unpublished)	Men outperform women on math tests; A diagnostic test of math ability	3
Guajardo (unpublished)	Gender prime: Demographic (gender) prior to tests	1
Harder (1999)	Diagnostic of math ability	1
Johns, Schmader, & Martens (2005)	Studying gender differences in math; Reported gender on the test	2
Josephs, Newman, Brown, & Beer (2003)	Questionnaire to prime gender-based stereotype threat (specific questions)	3
Keller & Bless (unpublished)	The test produced gender differences	2
Keller & Dauenheimer (2003)	The test produced gender differences	2
Keller (2002)	The test produced gender differences; Males outperform women	3
Keller (in press)	The test produced gender differences	2
Lewis (1998)	Race prime: Race inquiry before test	1
Martens, Johns, Greenberg, & Schimel (2006)	A direct measure of math intelligence	1
Martens, Johns, Greenberg, & Schimel (2006)	Men & women differ in performance on this test; Women have more trouble with spatial rotation tasks; Evaluate abilities and limitations	3
Martin (2004)	"You will be individually evaluated in your performance on this test.	1
Marx & Stapel (in press)	Diagnostic of math ability (strengths and weaknesses); Label "Diagnostic Exam"; A testing center	1
Marx, Stapel, & Muller (2005)	Diagnostic of math ability (strengths & weaknesses)	1
Marx, Stapel, & Muller (2005)	Diagnostic of math ability (strengths & weaknesses); Read about negative female role model (bad at math)	1
Marx, Stapel, & Muller (2005)	Diagnostic test (+ read about a neutral role model)	1
McFarland, Kemp, Viera, & Odin (2003)	Women score lower than men on math test; Test is diagnostic of math ability; Gender inquiry before test	3
McFarland, Lev-Arey, & Ziegert (2003)	Race prime: Racial identity questionnaire before test; Race inquiry before test. A test of intelligence	1
McIntyre, Lord, Gresky, Eyck, Frye, et al. (2005)	Women perform worse than men on math test; Reading biographies of NO successful women (all successful	3

	corporations)	
McIntyre, Paulson, & Lord (2003)	Research shows men outperform women in math, but evidence is mixed	3
McIntyre, Paulson, & Lord (2003)	Research showed men outperform women in math, but evidence is mixed; Reading about neutral corporation profiles	3
McKay (1999)	An IQ test diagnostic of strengths and weaknesses	1
Nguyen, O'Neal, & Ryan (2003)	Difficult test of ability and limitations; Race inquiry before tests	1
Nguyen, Shivpuri, Ryan & Langset (2004)	Men performed better than women; Test of math ability	3
O'Brien & Crandall (2003)	Test produced gender differences	2
Oswald & Harvey (2000)	Gender prime: Hostile environment: A sexist cartoon (no information about gender differences)	1
Pellegrini (2005)	Ethnic & gender inquiry before tests; Test measured intellectual abilities (including math ability) of Hispanic females on a measure of White-normed intelligence test	2
Philipp & Harton (2004)	Stereotype regarding women math ability	3
Ployhart, Ziegert & McFarland (2003)	A diagnostic test of strengths and weaknesses	1
Prather (2005)	An indicator of mathematical ability	1
Rosenthal & Crisp (2006)	Gender prime: Creating thoughts of gender differences in general ("generate things that can distinguish men from women") before test	1
Rosenthal & Crisp (2006)	Comparing math performance to see whether there is gender difference	2
Salinas (1998)	A pretest questionnaire about the "level of bias" as one of reasons for difficulty level of test	2
Sawyer & Hollis-Sawyer (2005)	Diagnostic test of general ability; Race inquiry before test	1
Schimel, Arndt, Banko, & Cook (2004)	Test of mathematical intelligence; Gender inquiry before test	1
Schmader & Johns (2003)	Test highly predictive of intelligence performance: Ethnicity inquiry before test	1
Schmader & Johns (2003)	Collecting normative data on men and women	1
Schmader & Johns (2003)	A measure of quantitative capacity; Gender differences in math performance and quantitative capacity	2
Schmader (2002)	Male researcher voice; Research purpose: Diagnostic test of personal math ability	1
Schmader, Johns & Barquissau (2004)	Studying how women score on the test relative to men (indicating gender math ability); Gender inquiry before test	2

Schneeberger & Williams (2003)	Men scored higher than women on math test	3
Schultz, Baker, Herrera, & Khazian (unpublished)	Whites score better than Hispanics on verbal tests	3
Seagal (2001)	Completing ST questionnaire to prime negative stereotypes about group intelligence	3
Sekaquaptewa & Thompson (2002)	Test described as traditional math to which gender stereotypes apply	2
Smith & Hopkins (2004)	African Americans do not do well on standardized math tests	3
Smith & White (2002)	Men were superior in math test; Information explains why men are better at math than women	3
Smith & White (2002)	Asians were superior in math to Whites; Information explains why Asians are better at math than Whites	3
Spencer (2005)	Women do not do well on math tests as men.	3
Spencer, Steele, & Quinn (1999)	Test showed gender differences	2
Spicer (1999)	Public threat: experimenter scores their test; Race prime: by an Implicit Association Test before test	1
Steele & Aronson (1995)	Diagnostic test of verbal abilities (strengths and weaknesses)	1
Steele & Aronson (1995)	Race prime: Race inquiry before tests	1
Sternberg, Jarvin, Leighton, Newman et al. (unpublished)	Boys outperform girls on math test	3
Stricker & Ward (2004)	Race prime: Race inquiry before tests	1
Stricker & Ward (2004)	Gender prime: Demographic (gender) prior to tests	1
Tagler (2003)	Certain groups of people perform better than others on math exams; Gender differences on standardized tests; Gender inquiry before test	2
van Dijk, et al. (unpublished)	Diagnostic test of math strengths and weaknesses; Group membership: A "We" sense	1
van Dijk, et al. (unpublished)	Diagnostic test of math strengths and weaknesses; Group membership: A "We" sense; Gender inquiry before test	1
van Dijk, Koenders, Korenhof, Mulder, & Vries (unpublished)	Diagnostic test of math strengths and weaknesses	1
von Hippel, von Hippel, Conway, Preacher et al. (2005)	Asians outscored Whites on IQ test; Race inquiry before test	3
Walsh, Hickey, & Duffy (1999)	Men got higher scores than females on this test	3
Wicherts, Dolan, & Hessen	Intelligence tests; Race prime: Ethnicity questionnaire	1

(2005)	before test	
Wicherts, Dolan, & Hessen (2005)	Gender differences: Females score lower on math tests than males	3
Wout, Shih, Jackson, & Sellers (unpublished)	Test performance would be assessed.	1
Wout, Shih, Jackson, & Sellers (unpublished)	Black Test evaluator endorsed the belief that there are group differences in the test	2
Wout, Shih, Jackson, & Sellers (unpublished)	White Test evaluator endorsed the belief that there are group differences in the test	2
Wout, Shih, Jackson, & Sellers (unpublished)	Evaluator used an unbiased test to investigate group differences on standardized tests ; pre-test racial questionnaire	2

Note. The stereotype threat condition in a study may consist of multiple stereotype threat cues. The explicitness of one of these cues decides the overall categorization. ^a Level: 1 = "Subtle;" 2 = "Explicit;" 3 = "Blatant."

Appendix B

Stereotype Threat-Removing Strategies

STUDY	STEREOTYPE THREAT-REMOVING STRATEGY	LEVEL ^a
Bailey (2004)	Explicit Intervention: A handout with information favoring females	2
Brown & Day (in press)	Task purpose: A series of puzzle-solving tasks	1
Brown & Josephs (1999)	Test purpose: Math test purpose: scoring above cutoff point = exceptionally strong in math	1
Brown & Josephs (1999)	Test purpose: Math test purpose: assessing weak performance. But weak performance has an excuse (a computer crash)	1
Brown & Pinel (2003)	Explicit Intervention: Math test is free of gender bias (men = women)	2
Brown, et al. (2001)	Explicit Intervention: Test is racially and ethnically unbiased	2
Cadinu, et al. (2005)	Explicit Intervention: Women obtain higher scores than men	2
Cadinu, et al. (2005)	Explicit Intervention: No differences between men & women	2
Cadinu, et al. (2005)	Explicit Intervention: Blacks perform better than Whites	2
Cadinu, et al. (2005)	Explicit Intervention: No differences between men and women	2
Cohen & Garcia (2005)	Indirect intervention: A Black peer was performing poorly on an ART test (not relevant to IQ)	1
Cotting (2003)	Task purpose: Study to better understand the psychological factors involved in the problem solving process	1
Davies, et al. (2002)	Indirect intervention: TV commercials show women in astereotypical role (engineering and healthcare)	1
Dinella (2004)	Explicit Intervention: Gender differences NOT found for this test; no gender inquiry	2
Dodge, et al. (2001)	Test purpose: A test of tolerance of uncertainty & no race inquiry before test	1
Edwards (2004)	Test purpose: A pilot test for math items only	1
Elizaga & Markman (unpublished)	Task purpose: A reasoning exercise	1
Foels (1998)	Explicit Intervention: This test has not shown gender differences	2
Foels (2000)	Explicit Intervention: No gender differences in this test	2
Ford, et al. (2004)	Explicit Intervention: Problem-solving strategies; Research showed no gender difference	2
Gresky, et al. (unpublished)	Indirect Intervention: ST + Individuality (elaborated "Me")	1
Gresky, et al. (unpublished)	Indirect Intervention: ST + Individuality ("Me")	1
Guajardo (unpublished)	Explicit Intervention: Educating about stereotype threat phenomenon	2
Harder (1999)	Explicit Intervention: Males and females perform equally well on the test	2
Harder (1999)	Explicit Intervention: The gender stereotype is irrelevant	2
Johns, Schmader, & Martens (2005)	Task purpose: Problem-solving exercise	1
Keller & Bless (unpublished)	Explicit Intervention: The test didn't produce gender differences (men = women)	2
Keller & Dauenheimer (2003)	Explicit Intervention: The test didn't produce gender differences (men = women)	2

Keller (in press)	Explicit Intervention: The test didn't produce gender differences (men = women)	2
Keller (in press)	Explicit Intervention: The test didn't produce gender differences (men = women)	2
Martens, et al. (2006)	Indirect Intervention: ST as above + Self-affirmation (personal importance)	1
Martin (2004)	Test purpose: "You will not be individually evaluated in your performance on this test."	1
Marx & Stapel (2005)	Task purpose: A reasoning exercise	1
Marx & Stapel (in press)	Task purpose: A reasoning exercise	1
Marx, Stapel, & Muller (2005)	Task purpose: A reasoning exercise	1
Marx, Stapel, & Muller (2005)	Indirect Intervention: Diagnostic test but read about a positive female role model (excellent at math)	1
Marx, Stapel, & Muller (2005)	Task purpose: A reasoning exercise	1
McFarland, Kemp, Viera, & Odin (2003)	Task purpose: Psychological factors involved in problem solving	1
McFarland, Lev-Arey, & Ziegert (2003)	Task purpose: A personality test and problem-solving task	1
McIntyre, Lord, Gresky, Eyck, Frye, et al. (2005)	Indirect Intervention: Reading 4 successful women biographies	1
McIntyre, Paulson, & Lord (2003)	Explicit Intervention: Women perform better than men in psychology experiments	2
McIntyre, Paulson, & Lord (2003)	Explicit Intervention: Research showed men outperform women in math, but evidence is mixed; Then read successful women profiles	2
McKay (1999)	Test purpose: Test not be used to evaluate ability	1
Nguyen, et al. (2004)	Task purpose: A problem-solving task	1
O'Brien & Crandall (2003)	Explicit Intervention: Test did not produce gender differences	2
Oswald & Harvey (2000)	Explicit Intervention: Males and females do equally well on this test	2
Pellegrini (2005)	Task purpose: Taking psychological measure; no demographic inquiry	1
Ployhart, et al., (2003)	Test purpose: A test of retail management skills	1
Prather (2005)	Test purpose: A test examining problem-solving strategies	1
Rivadeneyra (2001)	Test purpose: A test to choose types of problems for a future test	1
Rosenthal & Crisp (2006)	Indirect Intervention: Creating thoughts of common things between genders (focusing away from gender differences in general)	1
Rosenthal & Crisp (2006)	Explicit Intervention: There are gender differences in math. Creating thoughts of common things between genders (focusing away from gender differences in general)	2
Rosenthal & Crisp (2006)	Explicit Intervention: Creating thoughts of common things between genders (focusing away from gender differences in general. Then, there are gender differences in math.	2
Sawyer & Hollis-Sawyer (2005)	Task purpose: An interest measure; no race inquiry	1
Sawyer & Hollis-Sawyer (2005)	Task purpose: An interest measure; no race inquiry	1
Schimmel, Arndt, Banko, & Cook (2004)	Task purpose: A problem-solving exercise; no gender inquiry before test	1

Schmader & Johns (2003)	Task purpose: Non-diagnosticity: A measure of working memory capacity	1
Schmader & Johns (2003)	Task purpose: A measure of working memory capacity	1
Schmader & Johns (2003)	Task purpose: A problem solving exercise	1
Schmader, et al. (2004)	Task purpose: Studying individual performance (indicating personal math ability); no gender inquiry before test	1
Schneeberger & Williams (2003)	Explicit Intervention: Men & women perform the same	2
Schultz, et al. (unpublished)	Task purpose: Studying mental processes underlying verbal problem solving ability	1
Sekaquaptewa & Thompson (2002)	Explicit Intervention: Test described as material impervious to gender stereotypes; women = men in performance	2
Smith & Hopkins (2004)	Task purpose: A laboratory task	1
Smith & White (2002)	Explicit Intervention: No racial differences	2
Spencer, Steele, & Quinn (1999)	Explicit Intervention: Test didn't yield gender differences	2
Spicer (1999)	Indirect Intervention: they score their own test & IAT after test	1
Steele & Aronson (1995)	Task purpose: Solving verbal problems	1
Steele & Aronson (1995)	Task purpose: (Problem solving task) No race inquiry before task.	1
Sternberg, et al. (unpublished)	Explicit Intervention: Girls = Boys in math performance	2
Tagler (2003)	Task purpose: Studying the psychology of decision making and problem solving. No gender inquiry	1
van Dijk, et al. (unpublished)	Task purpose: A reasoning task	1
van Dijk, et al. (unpublished)	Indirect Intervention: Diagnostic test + Individuality priming "I"	1
von Hippel, et al. (2005)	Explicit Intervention: Test is culturally fair; no racial differences; no race inquiry	2
Walsh, Hickey, & Duffy (1999)	Task purpose: Studying how well Canadian university students can solve American word problems	1
Walters (2000)	Task purpose: Non-diagnostic of intellectual ability	1
Wicherts, Dolan, & Hessen (2005)	Explicit intervention: Mean scores of males and females are equal	2
Wout, et al. (unpublished)	Test purpose: A test development project; test performance would not be assessed	1
Wout, et al. (unpublished)	Explicit intervention: There would not be group differences on the unbiased test.	2

Note. ^a Level: 1 = Subtle. 2 = Explicit

Appendix C

Excluded Studies

No.	CITATION	STUDY NUMBER	CRITERION NOT MET	NOTE
1	Aronson & Inzlicht (in press)	1 of 1	4	No cognitive ability test performance
2	Aronson, Fried, & Good (2002)	1 of 1	1	Not testing "performance interference"
3	Bell & Spencer (2002)	1 of 1	6	Not enough information to calculate effect sizes
4	Bell, Anderson-Cook & Spencer (2004)	1 of 1	6	Not enough information to calculate effect sizes
5	Bell, Spencer, Iserman, & Logel (2003)	1 of 1	6	Not enough information to calculate effect sizes
6	Ben-Zeev, Fein & Inzlicht (2005)	1, 2 of 2	2, 4	No cognitive ability test performance; Not relevant to ST paradigm (solo status paradigm)
7	Blascovich, Spencer, Quinn, & Steele (2001)	1 of 1	4	No cognitive ability test performance
8	Bosson, Haymovitz, & Pinel (2004)	1 of 1	4	No cognitive ability test performance
9	Brown & Josephs (1999)	3 of 3	2	Not relevant to ST paradigm
10	Brown & Lee (2005)	1 of 1	1	Not testing "performance interference"
11	Brown, Steele, & Atkins (unpublished)	1 of 1	6	Not enough information to calculate effect sizes
12	Chapell & Overton (2002)	1 of 1	2, 4	Correlation study. No cognitive ability test performance
13	Chatman (1999)	1 of 1	2, 4	Correlation study. No cognitive ability test performance
14	Chung-Herrera, Ehrhart, Ehrhart, Hattrup, & Solamon (2005-conference)	1 of 1	2	Not relevant to ST paradigm (Correlation study)
15	Cole, Michailidou, Jerome, & Sumnall (2005)	1 of 1	4	No cognitive ability test performance
16	Croizet & Claire (1998)	1 of 1	7	The negative stereotype is SES-based
17	Croizet, Depres, Gauzins, Huguet, Leyens, & Meot (2004)	1 of 1	7	The negative stereotype is related to college major
18	Croizet, Dutrevis, & Desert (2002)	1 of 1	7	The negative stereotype is related to college degree prestige
19	Cullen, Hardison, & Sackett (2004)	1 of 1	2	Not relevant to ST paradigm (Correlation study)
20	Davies & Spencer (2002)	1 of 2	4	No cognitive ability test performance
21	Davies, Spencer, Quinn, & Gerhardstein (2002)	pilot	2	Not relevant to ST paradigm
22	Davis & Silver (2003)	1 of 1	4	No cognitive ability test performance
23	DeRouin, Fritzsche, & Salas (2004)	1 of 1	4	No cognitive ability test performance

24	Dutrevis & Croizet (2005)	1 of 1	4	No cognitive ability test performance
25	Ford, Ferguson, Brooks, & Hagadone (2004)	1 of 1	1	Not testing "performance interference"
26	Forster, et al. (2004)	1 of 1	4	No cognitive ability test performance
27	Fournet (2005)	1 of 1	4	No cognitive ability test performance
28	Frantz, Cuddy, Burnett, Ray, & Hart (2004)	1-3 of 3	4	No cognitive ability test performance
29	Gonzales, Blanton, & Williams (2002)	1 of 1	6	Not enough information to calculate effect sizes
30	Good, Aronson, & Inzlicht (2003)	1 of 1	1	Not testing "performance interference"; Confounded with the effect of mentoring
31	Halpern & Tan (2001)	1 of 1	2	Not relevant to ST paradigm
32	Hess, Auman, Colcombe, & Rahhal (2003)	1 of 1	4	No cognitive ability test performance
33	Inzlicht & Ben-Zeev (2000)	1-2 of 2	1	Not relevant to ST paradigm (solo status paradigm)
34	Inzlicht & Ben-Zeev (2003)	1 of 1	1	Not relevant to ST paradigm (solo status paradigm)
35	Josephs & Newman (2003)	2 of 2	1	no ST theory, only status enhancement
36	Keller & Bless (unpublished)	1 of 3	4	No cognitive ability test performance
37	Kray, Galinsky, & Thompson (2002)	1-2 of 2	4	No cognitive ability test performance
38	Kray, Reb, Galinsky, & Thompson (2004)	1-2 of 2	4	No cognitive ability test performance
39	Leyens, Desert, Croizet, & Darcis (2000)	1 of 1	4	No cognitive ability test performance
40	Malhomes (2001)	1 of 1	2, 4	Not relevant to ST paradigm; No cognitive ability tests
41	Marx, Stapel & Muller (2005)	1-2 of 2	4	No cognitive ability test performance
42	Marx, Stapel (in press)	2 of 2	4	No cognitive ability test performance
43	Mayer & Hanges (2003)	1 of 1	6	Not enough information to calculate effect sizes
44	McKay, Doverspike, Bowen-Hilton, & Martin (2002)	1 of 1	3	No within-group data reported. From McKay's dissertation.
45	Mehranian, Lee & Binder (unpublished)	1 of 1	6	Not enough information to calculate effect sizes
46	Milner & Hoy (2003)	1 of 1	1, 2, 4	Not testing "performance interference" nor being an experimental study
47	Niemann, O'Connor, & McClorie (1998)	1 of 1	1, 2, 4	Not testing "performance interference"; Not ST paradigm (Cluster analysis)
48	Norris-Watts, & Lord (2003-conference)	1 of 1	4	No cognitive ability test performance
49	Osborne (2002)	1 of 1	2	Correlation study.
50	Prime (2000)	1, 2 of 2	3, 4	Missing standard deviations. No cognitive ability test performance.
51	Prime (2000)	2 of 2	4	No cognitive ability test performance
52	Pronin, Steele, & Ross	1-3 of 3	2, 4	Not relevant to ST paradigm; No

	(2003)			cognitive ability test performance
53	Quinn & Spencer (2001)	1 of 1	1	No ST condition
54	Rahhal (1998)	1-5 of 5	4	No cognitive ability test performance
55	Roberson, Deitch, Brief, & Block (2003)	1 of 1	2, 4	Correlation study. No cognitive ability test performance
56	Schimel, Arndt, Branko & Cook (2004)	1, 3 of 3	1	no ST theory
57	Schmader, Johns & Barquissan (2004)	1 of 1	4	No cognitive ability test performance
58	Schultz, Baker, Herrera, & Khazian (unpublished)	2 of 2	4	No cognitive ability test performance
59	Seagal (2001)	1-5 of 6	1	Not testing "performance interference"
60	Seibt & Forster (2004)	1-5 of 5	4, 7	No cognitive ability test performance; The negative stereotype is college major-based
61	Sekaquaptewa & Thompson (2002)	1-2 of 2	2	Not stereotype threat paradigm ("Solo status" paradigm)
62	Shih, Pittinsky, & Ambady (1999)	1 of 1	6	Not enough information to calculate effect sizes
63	Slot, de Roos, Sijperda, Morsman, & van de Graaf (unpublished)	1 of 1	4	No cognitive ability test performance
64	Smith & Johnson (in press)	1-3 of 3	1	Not testing "performance interference" (stereotype lift effect)
65	Smith (2002)	1 of 1	4	No cognitive ability test performance
66	Smith (2006)	1-2 of 2	4	No cognitive ability test performance
67	Smith, Sansone, & White (unpublished)	1-3 of 3	4	No cognitive ability test performance
68	Spencer, Steele & Quinn (1999)	1, 3 of 3	1, 6	Not testing "performance interference;" Not enough information to calculate effect sizes
69	Stangor, Carr, & Kiang (1998)	1-2 of 2	4	No cognitive ability test performance
70	Steele & Aronson (1995)	3 of 4	1	No cognitive ability test performance
71	Stone (2002)	1-2 of 2	4	No cognitive ability test performance
72	Stone, Lynch, Sjomeling, & Darley (1999)	1-2 of 2	4	No cognitive ability test performance
73	van Millingen (2000)	1 of 1	2, 4	Not relevant to ST paradigm; No cognitive ability tests
74	Walters (2000)	2 of 2	2	Not relevant to ST paradigm
75	Williams (2004)	1 of 1	6	Not enough information to calculate effect sizes

Appendix D

Studies with a Stereotype Threat x Domain Identification Design ($k = 22$)

	CITATION	STUDY NO.	STEREO TYPED GROUP N	COMPARISON GROUP N	SELECTED DOMAIN ID. LEVEL ^a	DOMAIN IDENTIFICATION OPERATIONAL DEFINITION
1	Anderson (2001)	1 of 1	152	160	high	High DI: SAT math ≥ 620
2	Anderson (2001)	1 of 1	202	71	low	Low DI: SAT math < 550
3	Aronson, Lustina, Good Keough, Steele & Brown (1999)	1 of 2	23	-	high	Average and above on math identification scale (from neutral to strongly agree); SAT math ≥ 610
4	Aronson, Lustina, Good, Keough, Steele & Brown (1999)	2 of 2	26	-	high	Strongly endorsing the importance of math ability to self-concept
5	Aronson, Lustina, Good, Keough, Steele & Brown (1999)	2 of 2	23	-	medium high	Moderately endorsing the importance of math ability to self-concept
6	Brown & Pinel (2003)	1 of 1	46	-	medium high	Scored above the 20th percentile on the Mathematics Identification Questionnaire
7	Cadinu, Maass, Frigerio, et al. (2003)	1 of 2	25	-	high	Based on high ranking of logical-math abilities
8	Cadinu, Maass, Frigerio, et al. (2003)	1 of 2	38	-	low	Based on low ranking of logical-math abilities
9	Gresky, Eyck, Lord, & McIntyre (unpublished)	1 of 1	23	13	low	Not strongly endorsing the Math Identification Measure
10	Gresky, Eyck, Lord, & McIntyre (unpublished)	1 of 1	37	6	high	Strongly endorsing the Math Identification Measure
11	Harder (1999)	2 of 2	19	-	high	Took advanced math course; strongly identified with math
12	Josephs, Newman, Brown, & Beer (2003)	1 of 1	38	37	medium high	Scored above mid-point of a math identification scale
13	Keller (in press)	1 of 1	23	28	high	Strongly endorsing 2 items about math importance
14	Keller (in press)	1 of 1	32	25	low	Not strongly endorsing 2 items about math importance
15	Martens, Johns, Greenberg, & Schimel (2006)	1 of 2	22	-	high	SAT math scored at least 500

16	Quinn & Spencer (2001)	2 of 2	14	22	high	High SAT math score (650-670)
17	Schimmel, Arndt, Banko, & Cook (2004)	2 of 3	46	-	medium high	Scored above the midpoint of the academic self-esteem questionnaire
18	Schmader & Johns (2003)	1 of 3	28	31	medium high	Scored 500 or higher on SAT math
19	Schmader & Johns (2003)	3 of 3	28	-	medium high	Scored 500 or higher on SAT math
20	Spencer, Steele, & Quinn (1999)	2 of 3	30	24	high	Strong math background; High math identification
21	Spicer (1999) (Difficult test)	2 of 2	39	-	medium high	Academic Identity Scale: moderately endorsing that academic ability is important
22	Spicer (1999) (Easy test)	2 of 2	39	-	medium high	Academic identity scale: moderately endorsing academic ability is important

Note. The domain identification level for each sample was taken from individual research reports (i.e., assigned and defined by researchers themselves). Therefore, the operational definition of levels of domain identification may not be consistent across studies: some researchers used an objective measure of prior ability (e.g., SAT scores) to define their sub-samples, some used endorsement scores on a self-report measure of domain identification, and some others used both indicators. ^a Domain identification levels: $k_{high} = 10$; $k_{medium} = 8$; $k_{low} = 4$.

Appendix E

Studies with a Stereotype Threat x Test Difficulty Design ($k = 81$)

No.	STUDY		STEREOTYPED GROUP	TEST	TEST DIFFICULTY	LEVEL ^a
1	Anbady, Paik, Steele, Owen-Smith, & Mitchell (2004)	1 of 2	Female undergrads	Simple math test (addition & multiplication)	Easy	1
2	Anbady, Paik, Steele, Owen-Smith, & Mitchell (2004)	2 of 2	African American undergrads	Verbal aptitude test, 18 items, 20'	easy: answered correctly by 50% of sample	1
3	Aronson, Lustina, Good, & Keough (1999)	2 of 2	Female undergrads (British)	Math: Arithmetic test, 10 items	straightforward	1
4	Aronson, Lustina, Good, & Keough (1999)	2b of 2	Female undergrads (British)	Math: Arithmetic test, 10 items	straightforward	1
5	Bailey (2004)	1 of 1	Female undergrads (British)	Math: Arithmetic test, 10 items	straightforward	1
6	Brown & Day (in press)	1 of 1	Female undergrads	Multiplication, 20 items, 10'	easy	1
7	Brown & Josephs (1999)	1 of 3	Female secondary school students (German)	The third International Math & Science Study, 5 items	Easy	1
8	Brown & Josephs (1999)	2 of 3	Female secondary school students (German)	The third International Math & Science Study, 5 items	Easy	1
9	Brown, Steele, & Atkins (unpublished)	2 of 2	Female undergrads	GED math test (General Equivalency Diploma)	Easy	1
10	Cadinu, Maass, Frigerto, et al. (2003)	1 of 2	African American undergrads	GRE analytical, 20 items, 25'	Slightly easier (than a difficult test)	2
11	Cadinu, Maass, Frigerto, et al. (2003)	1b of 2	African American undergrads	GRE analytical, 20 items, 25'	Slightly easier (than a difficult test)	2
12	Cadinu, Maass, Rosabianca, & Kiesner (2005)	1 of 1	Female undergrads (Dutch)	4 math items	Easy-difficult	2
13	Davies, Spencer, Quinn, & Gerhardtstein (2002)	1 of 2	Female undergrads	GRE Math, 30 items, 20'	Moderate to challenging: Previous students answered 63.8% correctly	2

14	Davies, Spencer, Quinn, & Gerhardtstein (2002)	2 of 2	Minority high school students (Dutch)	composite of math, verbal, logical sequence	mixed difficult	2
15	Dodge, Williams, & Blanton (2001)	1 of 1	Female undergrads	GRE math, 10 items	Moderately difficult	2
16	Foels (1998)	1b of 1	Female undergrads	GRE math, 30 items	Moderately difficult (44% to 80% answered correctly)	2
17	Foels (1998)	1 of 1	White undergrads (male)	GRE math, 10 items	Moderately difficult	2
18	Foels (2000)	1 of 1	Latino American undergrads	Math equations, 72 items	Moderately difficult	2
19	Gresky, Eyck, Lord, & McIntyre (unpublished)	1 of 1	Female undergrads	GRE math, 30, 30 items	Mixed difficult	2
20	Gresky, Eyck, Lord, & McIntyre (unpublished)	1b of 1	Female undergrads	Math equations, 72 items	Moderately difficult	2
21	Harder (1999)	1 of 2 (pilot)	Female undergrads	composite of multiplication & SAT math	mixed difficult (easy + difficult)	2
22	Harder (1999)	2 of 2	Female undergrads	GRE Math, 30'	moderately difficult	2
23	Joins, Schneider, & Martens (2005)	1 of 1	Female undergrads	Math test, 26 items	mixed difficulty	2
24	Keller & Dauenhimer (2003)	1 of 1	African American undergrads	composite of math, verbal, anal	Mixed difficulty	2
25	Keller (in press)	1b of 1	Female secondary school students (German)	Short math test, 20'	mixed difficulty items	2
26	Keller (in press)	1 of 1	Female undergrads	SAT math, 15 items	medium difficulty	2
27	Keller (in press)	1c of 1	African American undergrads	Raven Advanced Progressive Matrices, 36 items, 45'	Moderately difficult	2
28	Lewis (1998)	1 of 1	African American undergrads	SAT verbal, 20 items	Mixed difficulty	2
29	Marx & Stapel (2005)	1 of 1	Female undergrads	GRE math, 24 items	mixed difficulty: correct response rates: 20-70%	2
30	Marx & Stapel (in press)	1 of 3	Female undergrads	GRE math, 20 items, 20'	A variety of difficulty levels and problem types	2
31	Marx & Stapel (in press)	3 of 3	Female undergrads (Italian)	GRE math, 7 items	Mean of correct responses: 4.4 (1.77)	2

32	Marx, Stapel, & Muller (2005)	3 of 4	African American undergrads	GRE analytical, 20 items, 25'	Challenging	3
33	Marx, Stapel, & Muller (2005)	3b of 4	African American undergrads	Raven's Advanced Progressive Matrices, 36 items	An ascending order of difficulty (= mixed difficulty)	2
34	Marx, Stapel, & Muller (2005)	4 of 4	Female undergrads	GRE math, 20 items, 20'	A variety of difficulty levels and problem types	2
35	McFarland, Kemp, Viern, & Odin (2003)	1 of 1	African American undergrads	GRE verbal, 17 items	Very difficult	3
36	McFarland, Lev-Arey, & Ziegert (2003)	1 of 1	Female undergrads (Canadian)	SAT math	Challenging	3
37	McIntyre, Lord, Gresky, Eyck, Frye, et al. (2005)	1 of 1	White undergrads (Australian)	Raven's Advanced Progressive Matrices	Difficult	3
38	McIntyre, Paulson, & Lord (2003)	1 of 2	Female high school students	GRE math, 40 items	Difficult	3
39	McIntyre, Paulson, & Lord (2003)	2 of 2	Female high school students	GRE math, 40 items	Difficult	3
40	McKay (1999)	1 of 1	African American undergrads	GRE verbal, 25 items, 25'	difficult (only 30% answered correctly)	3
41	Nguyen, O'Neal, & Ryan (2003)	1 of 1	African American undergrads	GRE verbal, 27 items, 25'	difficult (only 30% answered correctly)	3
42	Nguyen, Shivpuri, Ryan & Langset (2004)	1 of 1	African American undergrads	GRE verbal, 30 items, 30'	difficult (only 30% answered correctly)	3
43	O'Brien & Crandall (2003)	1b of 1	African American and Latino undergrads	GRE analytical, 20 items, 15'	Difficult	3
44	O'Brien & Crandall (2003)	1 of 1	Female undergrads	GRE math	Difficult	3
45	O'Brien & Crandall (2003)	1 of 1	African American undergrads	Verbal aptitude test, 12 items, 20'	very difficult: answered correctly by less than 50% of the sample	3
46	Oswald & Harvey (2000)	1 of 1	Hispanic American undergrads	GRE verbal, 20 items, 20'	Difficult	3
47	Ployhart, Ziegert & McFarland (2003)	1 of 1	Female undergrads	GRE math, 20 items, 20'	complex	3
48	Ployhart, Ziegert & McFarland (2003)	1b of 1	Hispanic American undergrads	GRE verbal, 20 items, 20'	Difficult	3

49	Prather (2005)	1 of 1	Female undergrads	GRE math, 20 items, 20'	Difficult	3
50	Rivadeneira (2001)	1 of 1	Latino high school students	PSAT & SAT math & Verbal (composite)	Challenging	3
51	Rosenthal & Crisp (2006)	2 of 3	Female undergrads	SAT math, 12 items, 11'	difficult	3
52	Rosenthal & Crisp (2006)	3 of 3	African American undergrads	GRE & GMAT math & analytical ability, 21 questions	difficult (only 37% test answered correctly)	3
53	Rosenthal & Crisp (2006)	3b of 3	African American undergrads	GRE & GMAT math & analytical ability, 21 questions	difficult (only 37% test answered correctly)	3
54	Schmader & Johns (2003)	1 of 3	Female undergrads	GRE math, 34 items, 20 minutes	difficult	3
55	Schmader & Johns (2003)	2 of 3	Female undergrads	GRE Math, 34 items	very difficult	3
56	Schmader & Johns (2003)	3 of 3	Female undergrads	GRE Math, 34 items	very difficult	3
57	Schmader (2002)	1 of 1	African American undergrads	GRE & GMAT math, verbal, and analytical	Difficult: answered correctly only 33% of items	3
58	Schmader, Johns & Barquissau (2004)	2 of 2	Female undergrads	GRE math, 20 items	Difficult	3
59	Schreoberger & Williams (2003)	1 of 1	Female undergrads	GRE math, 20 items	Difficult	3
60	Schultz, Baker, Herrera, & Khazian (unpublished)	1 of 2	Female undergrads	GRE Math test, 20 items, 30'	Challenging	3
61	Schultz, Baker, Herrera, & Khazian (unpublished)	2 of 2	Female undergrads	GRE math, 20 items	Difficult	3
62	Seagal (2001)	6 of 6	Female undergrads	GRE Math test, 20 items, 30'	Challenging	3
63	Smith & White (2002)	1 of 2	African American undergrads	GRE verbal, 27 items, 25'	Approximately 30% of students answered correctly.	3
64	Smith & White (2002)	2 of 2	Female undergrads	GRE Math test, 20 items, 30'	Challenging	3
65	Spencer, Steele, & Quinn (1999)	2 of 3	Female secondary school students	The third International Math & Science Study, 12 items	Difficult	3
66	Spicer (1999)	2 of 2	Female undergrads	Math test, 20'	challenging	3
67	Spicer (1999)	2b of 2	Female undergrads	GRE math, 30 items, 20'	Difficult	3
68	Steele & Aronson (1995)	1 of 4	Female undergrads	GRE math, 30 items, 20'	Difficult	3
69	Steele & Aronson (1995)	2 of 4	Female undergrads	GRE math, 30 items, 20'	Difficult	3
70	Steele & Aronson (1995)	4 of 4	Female undergrads	GRE math, 15 items, 20'	challenging	3

71	Sternberg, Jarvin, Leighton, Newman et al. (unpublished)	1 of 2	Female undergrads	GMAT math	Difficult	3
72	Sternberg, Jarvin, Leighton, Newman et al. (unpublished)	2 of 2	Female undergrads	GRE math	Difficult	3
73	Tagler (2003)	1 of 1	Female undergrads	GRE advanced math; 12 items	difficult	3
74	von Hippel, von Hippel, Conway, Preacher et al. (2005)	4 of 4	Female undergrads (Italian)	7 math items, 10'	difficult	3
75	Walsh, Hickey, & Duffy (1999)	2 of 2	Female undergrads (Italian)	7 math items, 10'	difficult	3
76	Walters (2000)	1 of 2	African American students	GRE verbal, 30 items, 25 minutes	very difficult	3
77	Wicherts, Dolan, & Hessen (2005)	1 of 3	White undergrads (men)	GRE calculus, 15 items, 20'	at the upper limit of students' ability	3
78	Wicherts, Dolan, & Hessen (2005)	3 of 3	Female undergrads	GRE math, 30 items, 20'	challenging	3
79	Wout, Shih, Jackson, & Sellers (unpublished)	2 of 4	Female undergrads	Canadian Math competition, 12 items, 20'	challenging	3
80	Wout, Shih, Jackson, & Sellers (unpublished)	3 of 4	Female undergrads	Canadian Math competition, 12 items, 20'	challenging	3
81	Wout, Shih, Jackson, & Sellers (unpublished)	4 of 4	White undergrads (men)	GRE calculus, 15 items, 20'	at the upper limit of students' ability	3

Note: The test difficulty level for each sample was taken from individual research reports (i.e., assigned and defined by researchers themselves). Therefore, the operational definition of levels of domain identification may not be consistent across studies. ^a Test difficulty levels: $k_{high} = 47$; $k_{medium} = 28$; $k_{low} = 12$.

Appendix F

Coding Manual

General Direction:

This is the Coding Manual for the meta-analysis on stereotype threat effects. (The Coding Form is identical to this manual minus the *direction* column.) Please use a copy of the Coding Form to code each independent study. The information on this form will be subsequently entered in a master Microsoft Excel file (Coding_Form.xls) which is similarly formatted.

NOTE: Please use the Inclusion Criteria Chart first to determine whether a study is eligible to be included in the sample.

Coder (Initial your name here): _____

Date: _____

A. Source Identification

	Direction
Citation: _____ _____	Cite the source paper in APA format. If there is a series of studies in one source paper, cite the full information for the first study on this form and cite only authors' name and year for subsequent forms.
Study number: _____	Code whether this study is part of a series. For example, if it is the first study of three studies in a source paper, code it as "1 of 3." If it is the only study, code it as "1 of 1." NOTE: Use a <i>separate coding form</i> for each study in a series.

B. Descriptive Variables

No	Variable Label	Code	Location*	Direction
B1	Source ID		n/a	Assign an identification number to the source paper. For example, H-001, E-001(2), or I-001(3). NOTE: The letter is coder's first initial; "001" is the arbitrarily assigned ID number; "(2)" is the study number if any (see Source Identification above)
B2	Year		n/a	List year when the source paper was published, presented, or conducted (unpublished manuscripts).
B3	Pub status?	1 0	n/a	Circle a number to indicate whether the source paper was (1) published in a peer-reviewed journal or (0) unpublished.
B4	Non-pub status?	1 2 3	n/a	If B3 is "0," circle a number to indicate if the source paper was (1) a dissertation or thesis, (2) a conference paper, or (3) a working manuscript.

Note. (*)Record the location where the information was found in the source paper (e.g., page 5; Table 2; Figure 1).

C. Sample Characteristics

No	Variable Label	Code	Location*	Direction
C1	How many ST sub-	1 2 3 4		Circle a number to indicate how many within-group sub-samples in this study (i.e., sub-samples of

	samples?					<p>stereotyped groups). For example, circle "1" when there are only either women or only Blacks (one stereotyped group); circle "2" when there are both Blacks and Hispanics (2 stereotyped ethnic groups) in this study, etc.</p> <p>NOTE: From this point forward, use a <i>separate coding form</i> for each sub-sample. For example, if there are both Blacks and Hispanics, code the information for "Blacks" on this form only, and use another coding form for "Hispanics." Repetitive descriptive information (above) can be skipped on the subsequent coding forms.</p>
C2	Comparison sample?	1	0			Circle a number to indicate whether (1) there is a comparison sample (e.g., men; Whites) of non-stereotyped groups in this study, or (0) there is no comparison group (i.e., only stereotyped groups).
C3	ST Sample: Description					Briefly describe the sample (or sub-sample) of the stereotyped group; for example: "Black college students;" "Female high school students." If a study is conducted outside of the U.S., note the nation/country here.
C4	Group-based ST coded	1	0	n/a		Circle a number to indicate whether (1) the stereotype is gender-based (e.g., male v. female) or (0) the stereotype is race-based (e.g., Blacks v. Whites). Note: The stereotype threat may be implied (not explicitly stated)
C5	Stereotyped sample size $N_{\text{Total-ST}}$					Record the total sample size (or sub-sample size) of the stereotyped group only (e.g., Blacks; women).
C6	Comparison sample size $N_{\text{Total-Compare}}$					Record the total sample size of the comparison group only (e.g., Whites; men) if any.

Note. (*)Record the location where the information was found in the source paper (e.g., page 5; Table 2; Figure 1).

D. Moderators: Domain Identification

No	Variable Label	Code	Location*	Direction
D1	Domain identification included?	1 0		Circle a number to indicate whether (1) domain identification was included or mentioned in the study design or (0) not mentioned. If "0," skip to the next section.
D2	Sample pre-selected based on domain identification?	1 0		Circle a number to indicate whether (1) the stereotyped sample was pre-selected based on levels of identification with an intellectual or cognitive ability domain, or (0) there was no pre-selection.
D3	Pre-selected domain identification level			Describe whether the pre-selected criterion in terms of domain identification was high, medium (average), low, or any other combination.
D4	Domain			Describe how domain identification was operationalized in

	identification: Description			this study. For example, to pre-select individuals with high math domain identification, did the researcher use high SAT scores, endorsement on a math domain identification scale, or both?
D5	Domain identification: Categorical data			If the authors categorized their sample (of the stereotyped group) into sub-groups based on levels of domain identification, such as high, medium high, medium, and/or low, describe what sub-groups are used (e.g., high and low only).
D6	Domain identification: Continuous data			Record descriptive statistics (i.e., mean, standard deviation, etc.) regarding to the levels of domain identification, if any. Also record how levels of domain identification were operationalized (e.g., "high" is above a cutoff score).

Note. (*)Record the location where the information was found in the source paper (e.g., page 5; Table 2; Figure 1).

E. Moderators: Research Design Characteristics

No	Variable Label	Code	Location*	Direction
E0	Experimental design: Description			Describe how the authors designed this study. For example: "2 (gender) x 2 (ST)"

(1) For Stereotyped group(s) only

No	Variable Label	Code			Location*	Direction
E1	ST condition: How many ST cues?					Record how many stereotype threat cues were used in the Stereotype threat condition (treatment condition).
E2	ST cue 1: Description					Describe how the 1st stereotype threat cue was operationalized.
E3	ST cue 2: Description					Describe how the 2nd stereotype threat cue (if any) was operationalized.
E4	ST cue 3: Description					Describe how the 3rd stereotype threat cue (if any) was operationalized.
E5	ST cue 1: Presentation mode	1	2	3		Circle a number to indicate whether the 1st stereotype threat cue was presented (1) implicitly, (2) moderately explicitly, or (3) blatantly explicitly.
E6	ST cue 2: Presentation mode	1	2	3		Circle a number to indicate whether the 2nd stereotype threat cue was presented (1) implicitly, (2) moderately explicitly, or (3) blatantly explicitly.
E7	ST cue 3: Presentation mode	1	2	3		Circle a number to indicate whether the 3rd stereotype threat cue was presented (1) implicitly, (2) moderately explicitly, or (3) blatantly explicitly.
E8	ST-removal condition?	1	0			Circle a number to indicate whether (1) there was a treatment condition where stereotype threat was removed or refuted explicitly or (0) there was no such condition.

E9	ST-removal: Description				Describe how the stereotype threat removal strategy was operationalized.
E10	ST removal : Presentation mode	1	2		Circle a number to indicate whether the stereotype threat removal cue was presented (1) implicitly or (2) explicitly. NOTE.
E11	Control condition?	1	0		Circle a number to indicate whether (1) there was a control condition where there were no special directions regarding stereotype threat or (0) there was no such condition.
E12	Control: Description				Describe how the control condition was operationalized.
E13	Cell <i>n</i> : ST condition				ST condition sample size (for stereotyped group only)
E14	Cell <i>n</i> : ST removal				ST-removal condition sample size (for stereotyped group only)
E15	Cell <i>n</i> : Control				Control condition sample size (for stereotyped group only)

Note. (*)Record the location where the information was found in the source paper (e.g., page 5; Table 2; Figure 1).

(2) For comparison group(s) only (if any)

No	Variable Label	Code	Location*	Direction
E16	Cell <i>n</i> : ST condition			ST condition sample size (for comparison group only)
E17	Cell <i>n</i> : ST removal			ST-removal condition sample size (for comparison group only)
E18	Cell <i>n</i> : Control			Control condition sample size (for comparison group only)

Note. (*) Record the location where the information was found in the source paper (e.g., page 5; Table 2; Figure 1).

F. Moderators: Characteristics of Cognitive Ability Tests

No	Variable Label	Code	Location*	Direction
F1	How many ability domains were assessed?			Record how many cognitive ability domains were assessed in the study. For example, if there is only math, write "1;" if there were math, verbal and analytical abilities assessed, write "3."
F2	Math: Description; alpha			Describe how the math domain of cognitive ability was tested. For example, 30 questions from the GRE-Math. What is the internal consistency alpha?
F3	Verbal: Description; alpha			Describe how the verbal domain of cognitive ability was tested. For example, 30 questions from the GRE-Verbal. What is the internal consistency alpha?
F4	Analytical: Description; alpha			Describe how the analytical domain of cognitive ability was tested. For example, 30 questions from the GRE-Analytical. What is the internal consistency alpha?
F5	Spatial:			Describe how the spatial domain of cognitive

	Description; alpha				ability was tested. For example, 30 questions of spatial ability. What is the internal consistency alpha?
F6	General: Description; alpha				If overall cognitive ability test performance or overall intelligence test performance is reported, describe how it was tested. What is the internal consistency alpha?
F7	Test difficulty reported?	1	0		Circle a number to indicate if test difficulty level was reported (1 = Yes, 0 = No)
F8	Math test difficulty: Description				Describe how math "test difficulty" was operationalized. For example: very difficult (only 30% completed).
F9	Verbal test difficulty: Description				Describe how Verbal "test difficulty" was operationalized.
F10	Analytical test difficulty: Description				Describe how Analytical "test difficulty" was operationalized.
F11	Spatial test difficulty: Description				Describe how Spatial "test difficulty" was operationalized.
F12	General test difficulty: Description				Describe how overall "test difficulty" was operationalized.
F13	Math test difficulty level	1	2	3	Circle a number to indicate whether the math test was (1) easy/very easy; (2) moderately/mixed difficult; or (3) difficult/very difficult
F14	Verbal test difficulty level	1	2	3	Circle a number to indicate whether the Verbal test was (1) easy/very easy; (2) moderately/mixed difficult; or (3) difficult/very difficult
F15	Analytical test difficulty level	1	2	3	Circle a number to indicate whether the Analytical test was (1) easy/very easy; (2) moderately/mixed difficult; or (3) difficult/very difficult
F16	Spatial test difficulty level	1	2	3	Circle a number to indicate whether the Spatial test was (1) easy/very easy; (2) moderately/mixed difficult; or (3) difficult/very difficult
F17	General Intelligence test difficulty level	1	2	3	Circle a number to indicate whether the overall test was (1) easy/very easy; (2) moderately/mixed difficult; or (3) difficult/very difficult

Note. (*) Record the location where the information was found in the source paper (e.g., page 5; Table 2; Figure 1).

G. Dependent Variables: Descriptive and/or Inferential Statistics

(1) For stereotyped group only

No	Variable Label	Code	Location*	Direction
G1	ST Mean: Math			Mean of math performance for ST condition.
G2	ST SD: Math			Standard deviation of math performance for ST condition.
G3	ST-removed Mean: Math			Mean of math performance for ST-removed condition.

G4	ST-removed SD: Math			Standard deviation of math performance for ST-removed condition.
G5	Control Mean: Math			Mean of math performance for control condition.
G6	Control SD: Math			Standard deviation of math performance for control condition.
G7	t (df): Math			Independent sample t-test estimate for math: between ST & ST-removed
G8	F (1, #): Math			F test estimate for math
G9	MSE			Mean Squared Error; Extracted from F-value table
G10	ST Mean: Verbal			Mean of verbal performance for ST condition.
G11	ST SD: Verbal			Standard deviation of verbal performance for ST condition.
G12	ST-removed Mean: Verbal			Mean of verbal performance for ST-removed condition.
G13	ST-removed SD: Verbal			Standard deviation of verbal performance for ST-removed condition.
G14	Control Mean: Verbal			Mean of verbal performance for control condition.
G15	Control SD: Verbal			Standard deviation of verbal performance for control condition.
G16	t (df): Verbal			Independent sample t-test estimate for verbal: between ST & ST-removed
G17	F (1, #): Verbal			F test estimate for verbal
G18	MSE			Mean Squared Error; Extracted from F-value table
G19	ST Mean: Analytical			Mean of analytical performance for ST condition.
G20	ST SD: Analytical			Standard deviation of analytical performance for ST condition.
G21	ST-removed Mean: Analytical			Mean of analytical performance for ST-removed condition.
G22	ST-removed SD: Analytical			Standard deviation of analytical performance for ST-removed condition.
G23	Control Mean: Analytical			Mean of analytical performance for control condition.
G24	Control SD: Analytical			Standard deviation of analytical performance for control condition.
G25	t (df): Analytical			Independent sample t-test estimate for analytical: between ST & ST-removed
G26	F (1, #): Analytical			F test estimate for analytical
G27	MSE			Mean Squared Error; Extracted from F-value table
G28	ST Mean: Spatial			Mean of spatial performance for ST condition.
G29	ST SD: Spatial			Standard deviation of spatial performance for ST condition.
G30	ST-removed Mean: Spatial			Mean of spatial performance for ST-removed condition.
G31	ST-removed			Standard deviation of spatial performance for ST-removed

	SD: Spatial			condition.
G32	Control Mean: Spatial			Mean of spatial performance for control condition.
G33	Control SD: Spatial			Standard deviation of spatial performance for control condition.
G34	t (df): Spatial			Independent sample t-test estimate for spatial: between ST & ST-removed
G35	F (1, #): Spatial			F test estimate for spatial
G36	MSE			Mean Squared Error; Extracted from F-value table
G37	ST Mean: General			Mean of cognitive ability (total) performance for ST condition.
G38	ST SD: General			Standard deviation of cognitive ability (total) performance for ST condition.
G39	ST-removed Mean: General			Mean of cognitive ability (total) performance for ST-removed condition.
G40	ST-removed SD: General			Standard deviation of cognitive ability (total) performance for ST-removed condition.
G41	Control Mean: General			Mean of cognitive ability (total) performance for control condition.
G42	Control SD: General			Standard deviation of cognitive ability (total) performance for control condition.
G43	t (df): General			Independent sample t-test estimate for cognitive ability (total): between ST & ST-removed
G44	F (1, #): General			F test estimate for cognitive ability (total)
G45	MSE			Mean Squared Error; Extracted from F-value table

(2) For comparison group only (if any)

No	Variable Label	Code	Location*	Direction
G46	ST Mean: Math			Mean of math performance for ST condition.
G47	ST SD: Math			Standard deviation of math performance for ST condition.
G48	ST-removed Mean: Math			Mean of math performance for ST-removed condition.
G49	ST-removed SD: Math			Standard deviation of math performance for ST-removed condition.
G50	Control Mean: Math			Mean of math performance for control condition.
G51	Control SD: Math			Standard deviation of math performance for control condition.
G52	t (df): Math			Independent sample t-test estimate for math: between ST & ST-removed
G53	F (1, #): Math			F test estimate for math
G54	MSE			Mean Squared Error; Extracted from F-value table
G55	ST Mean: Verbal			Mean of verbal performance for ST condition.
G56	ST SD: Verbal			Standard deviation of verbal performance for ST condition.

G57	ST-removed Mean: Verbal			Mean of verbal performance for ST-removed condition.
G58	ST-removed SD: Verbal			Standard deviation of verbal performance for ST-removed condition.
G59	Control Mean: Verbal			Mean of verbal performance for control condition.
G60	Control SD: Verbal			Standard deviation of verbal performance for control condition.
G61	t (df): Verbal			Independent sample t-test estimate for verbal: between ST & ST-removed
G62	F (1, #): Verbal			F test estimate for verbal
G63	MSE			Mean Squared Error; Extracted from F-value table
G64	ST Mean: Analytical			Mean of analytical performance for ST condition.
G65	ST SD: Analytical			Standard deviation of analytical performance for ST condition.
G66	ST-removed Mean: Analytical			Mean of analytical performance for ST-removed condition.
G67	ST-removed SD: Analytical			Standard deviation of analytical performance for ST-removed condition.
G68	Control Mean: Analytical			Mean of analytical performance for control condition.
G69	Control SD: Analytical			Standard deviation of analytical performance for control condition.
G70	t (df): Analytical			Independent sample t-test estimate for analytical: between ST & ST-removed
G71	F (1, #): Analytical			F test estimate for analytical
G72	MSE			Mean Squared Error; Extracted from F-value table
G73	ST Mean: Spatial			Mean of spatial performance for ST condition.
G75	ST SD: Spatial			Standard deviation of spatial performance for ST condition.
G75	ST-removed Mean: Spatial			Mean of spatial performance for ST-removed condition.
G76	ST-removed SD: Spatial			Standard deviation of spatial performance for ST-removed condition.
G77	Control Mean: Spatial			Mean of spatial performance for control condition.
G78	Control SD: Spatial			Standard deviation of spatial performance for control condition.
G79	t (df): Spatial			Independent sample t-test estimate for spatial: between ST & ST-removed
G80	F (1, #): Spatial			F test estimate for spatial
G81	MSE			Mean Squared Error; Extracted from F-value table
G82	ST Mean: General			Mean of cognitive ability (total) performance for ST condition.
G83	ST SD: General			Standard deviation of cognitive ability (total) performance for ST condition.

G84	ST-removed Mean: General			Mean of cognitive ability (total) performance for ST-removed condition.
G85	ST-removed SD: General			Standard deviation of cognitive ability (total) performance for ST-removed condition.
G86	Control Mean: General			Mean of cognitive ability (total) performance for control condition.
G87	Control SD: General			Standard deviation of cognitive ability (total) performance for control condition.
G88	t (df): General			Independent sample t-test estimate for cognitive ability (total): between ST & ST-removed
G89	F (1, #): General			F test estimate for cognitive ability (total)
G90	MSE			Mean Squared Error; Extracted from F-value table

(3) ***Between-group inferential statistics (t estimates; F estimates)***

DIRECTION: Record the inferential statistics obtained for between-group statistical procedures. For example, the independent sample *t* estimate of mean performance differences between women and men on a math test.

No	Variable Label	Code	Location*	Direction
G91				
G92				
G93				
G94				

H. Miscellaneous Coding:

No	Variable Label	Code		Location*	Direction
G95	Non-independent data points	1	0		Circle "1" for Yes when the same test takers took more than one cognitive ability test (e.g., taking both math and verbal), resulting in more than one dependent variable observation (e.g., both math and verbal scores). Circle "0" if it is not the case.

I. NOTES

DIRECTION: On this page, write down coding information of which you may be unsure, want to revisit or discuss with other coders.

Appendix G

Coding Form

General Direction:

This is the Coding Form for the meta-analysis on stereotype threat effects. The information on this form will be subsequently entered in a Microsoft Excel file (Coding_Form.xls) which is similarly formatted. The direction for each coding item can be found in the accompanied Coding Manual.

NOTE: Please use the Inclusion Criteria Chart first to determine whether a study is eligible to be included in the sample.

Coder: _____

Date: _____

A. Source Identification

Citation:
Study number:

B. Descriptive Variables

No	Variable Label	Code		
B1	Source ID			
B2	Year			
B3	Pub status?	1	0	
B4	Non-pub status?	1	2	3

C. Sample Characteristics

No	Variable Label	Code				Location
C1	How many ST sub-samples?	1	2	3	4	
C2	Comparison sample?	1		0		
C3	ST Sample: Description					
C4	Group-based ST coded	1		0		n/a
C5	Stereotyped sample size $N_{Total-ST}$					
C6	Comparison sample size $N_{Total-Compare}$					

D. Moderator Variables: Domain Identification

No	Variable Label	Code		Location
D1	Domain identification included?	1	0	
D2	Sample pre-selected based on domain identification?	1	0	
D3	Pre-selected domain identification level			
D4	Domain identification: Description			
D5	Domain identification: Categorical data			
D6	Domain identification: Continuous data			

E. Moderators: Research Design Characteristics

No	Variable Label	Code	Location
E0	Experimental design: Description		

(1) For Stereotyped group(s) only

No	Variable Label	Code			Location
E1	ST condition: How many ST cues?				
E2	ST cue 1: Description				
E3	ST cue 2: Description				
E4	ST cue 3: Description				
E5	ST cue 1: Presentation mode	1	2	3	
E6	ST cue 2: Presentation mode	1	2	3	
E7	ST cue 3: Presentation mode	1	2	3	
E8	ST-removal condition?	1	0		
E9	ST-removal: Description				
E10	ST removal : Presentation mode	1	2		
E11	Control condition?	1	0		
E12	Control: Description				
E13	Cell <i>n</i> : ST condition				
E14	Cell <i>n</i> : ST removal				
E15	Cell <i>n</i> : Control				

(2) For comparison group(s) only (if any)

No	Variable Label	Code	Location
E16	Cell <i>n</i> : ST condition		
E17	Cell <i>n</i> : ST removal		
E18	Cell <i>n</i> : Control		

F. Moderators: Characteristics of Cognitive Ability Tests

No	Variable Label	Code			Location
F1	How many ability domains were assessed?				
F2	Math: Description; alpha				
F3	Verbal: Description; alpha				
F4	Analytical: Description; alpha				
F5	Spatial: Description; alpha				
F6	Overall/Intelligence: Description; alpha				
F7	Test difficulty reported?	1	0		
F8	Math test difficulty: Description				
F9	Verbal test difficulty: Description				
F10	Analytical test difficulty: Description				
F11	Spatial test difficulty: Description				
F12	Overall/Intelligence test difficulty: Description				
F13	Math test difficulty level	1	2	3	
F14	Verbal test difficulty level	1	2	3	
F15	Analytical test difficulty level	1	2	3	
F16	Spatial test difficulty level	1	2	3	
F17	Overall/Intelligence test difficulty level	1	2	3	

G. Dependent Variables: Descriptive and/or Inferential Statistics

(1) For stereotyped group only

No	Variable Label	Code	Location
G1	ST Mean: Math		
G2	ST SD: Math		
G3	ST-removed Mean: Math		
G4	ST-removed SD: Math		
G5	Control Mean: Math		
G6	Control SD: Math		
G7	t (df): Math		
G8	F (1, #): Math		
G9	MSE		
G10	ST Mean: Verbal		
G11	ST SD: Verbal		
G12	ST-removed Mean: Verbal		
G13	ST-removed SD: Verbal		
G14	Control Mean: Verbal		
G15	Control SD: Verbal		
G16	t (df): Verbal		
G17	F (1, #): Verbal		
G18	MSE		
G19	ST Mean: Analytical		
G20	ST SD: Analytical		
G21	ST-removed Mean: Analytical		
G22	ST-removed SD: Analytical		
G23	Control Mean: Analytical		
G24	Control SD: Analytical		
G25	t (df): Analytical 1		
G26	F (1, #): Analytical		
G27	MSE		
G28	ST Mean: Spatial		
G29	ST SD: Spatial		
G30	ST-removed Mean: Spatial		
G31	ST-removed SD: Spatial		
G32	Control Mean: Spatial		
G33	Control SD: Spatial		
G34	t (df): Spatial		
G35	F (1, #): Spatial		
G36	MSE		
G37	ST Mean: Total		
G38	ST SD: Total		
G39	ST-removed Mean: General		
G40	ST-removed SD: General		
G41	Control Mean: General		
G42	Control SD: General		
G43	t (df): General		
G44	F (1, #): General		
G45	MSE		

(2) *For comparison group only (if any)*

No	Variable Label	Code	Location
G46	ST Mean: Math		
G47	ST SD: Math		
G48	ST-removed Mean: Math		
G49	ST-removed SD: Math		
G50	Control Mean: Math		
G51	Control SD: Math		
G52	t (df): Math		
G53	F (1, #): Math		
G54	MSE		
G55	ST Mean: Verbal		
G56	ST SD: Verbal		
G57	ST-removed Mean: Verbal		
G58	ST-removed SD: Verbal		
G59	Control Mean: Verbal		
G60	Control SD: Verbal		
G61	t (df): Verbal		
G62	F (1, #): Verbal		
G63	MSE		
G64	ST Mean: Analytical		
G65	ST SD: Analytical		
G66	ST-removed Mean: Analytical		
G67	ST-removed SD: Analytical		
G68	Control Mean: Analytical		
G69	Control SD: Analytical		
G70	t (df): Analytical		
G71	F (1, #): Analytical		
G72	MSE		
G73	ST Mean: Spatial		
G75	ST SD: Spatial		
G75	ST-removed Mean: Spatial		
G76	ST-removed SD: Spatial		
G77	Control Mean: Spatial		
G78	Control SD: Spatial		
G79	t (df): Spatial		
G80	F (1, #): Spatial		
G81	MSE		
G82	ST Mean: General		
G83	ST SD: General		
G84	ST-removed Mean: General		
G85	ST-removed SD: General		
G86	Control Mean: General		
G87	Control SD: General		
G88	t (df): General		
G89	F (1, #): General		
G90	MSE		

(3) Between-group inferential statistics (*t* estimates; *F* estimates)

DIRECTION: Record the inferential statistics obtained for between-group statistical procedures. For example, the independent sample *t* estimate of mean performance differences between women and men on a math test.

No	Variable Label	Code	Location
G91			
G92			
G93			
G94			

H. Miscellaneous Coding:

No	Variable Label	Code	Location
G95	Non-independent data points	1	0

I. NOTES

Appendix H

Hypothesized Within-Group Meta-Analytic Findings from Sensitive Subsets

<u>Overall Findings</u>			
<i>K</i>			112
<i>N</i>			6040
Mean <i>d</i>			-0.31
Var <i>d</i>			0.287
Var <i>e</i>			0.076
Mean δ			-0.337
Var δ			0.25
% var SE			26.52
% var acc. for (V%)			26.59
90% CI			(-0.98) - (0.30)
Fail safe <i>N</i>			459
<u>Race/ethnicity Subset Findings</u>		<u>Female Subset Findings</u>	
<i>K</i>	43	<i>K</i>	70
<i>N</i>	2520	<i>N</i>	3601
Mean <i>d</i>	-0.378	Mean <i>d</i>	-0.245
Var <i>d</i>	0.202	Var <i>d</i>	0.325
Var <i>e</i>	0.07	Var <i>e</i>	0.079
Mean δ	-0.412	Mean δ	-0.268
Var δ	0.156	Var δ	0.292
% var SE	34.82	% var SE	24.43
% var acc. for (V%)	34.99	% var acc. for (V%)	24.47
90% CI	(-0.92) - (0.09)	90% CI	(-0.96) - (0.42)
Fail safe <i>N</i>	206	Fail safe <i>N</i>	242

Stereotype threat-activating cues

Minority test takers

	<u>STA Blatant</u> ^a	<u>STA Explicit</u>	<u>STA Subtle</u> ^b
<i>K</i>	4	n/a	24
<i>N</i>	175		1543
Mean <i>d</i>	-0.70		-0.28
Var <i>d</i>	0.20		0.25
Var <i>e</i>	0.10		0.06
Mean δ	-0.76		-0.31
Var δ	0		0.22
% var SE	0.31		0.26
% var acc. for (V%)	100		26
90% CI	n/a		(-0.91) – (0.29)
Fail safe <i>N</i>	32		91

Women test takers

	<u>STA Blatant</u>	<u>STA Explicit</u> ^c	<u>STA Subtle</u> ^d
<i>K</i>	n/a	18	29
<i>N</i>		770	1085
Mean <i>d</i>		-0.37	-0.38
Var <i>d</i>		0.10	0.41
Var <i>e</i>		0.10	0.11
Mean δ		-0.41	-0.42
Var δ		0.01	0.36
% var SE		94.53	26.94
% var acc. for (V%)		94.85	27.02
90% CI		(-0.51) – (-0.31)	(-1.18) – (0.34)
Fail safe <i>N</i>		85	139

Note.

^a Excluding Seagal (2001), *N* = 101 and Smith & Hopkins (2004), *N* = 160.

^b Excluding Anderson (2001), *N* = 468.

^c Excluding Dinella (2004), *N* = 232, and Tagler (2003), *N* = 136.

^d Excluding Anderson (2001), *N* = 604, Stricker & Ward (2004), *N* = 730, and Elizaga & Markman (unpublished), *N* = 145.

Stereotype threat-removing strategies

Minority test takers

	<u>STR Explicit</u>	<u>STR Subtle</u>
<i>K</i>	n/a	n/a
<i>N</i>		
Mean <i>d</i>		
Var <i>d</i>		
Var <i>e</i>		
Mean δ		
Var δ		
% var SE		
% var acc. for (V%)		
90% CI		
Fail safe <i>N</i>		

Women test takers

	<u>STR Explicit^a</u>	<u>STR Subtle</u>
<i>K</i>	30	n/a
<i>N</i>	1394	
Mean <i>d</i>	-0.24	
Var <i>d</i>	0.26	
Var <i>e</i>	0.88	
Mean δ	-0.26	
Var δ	0.21	
% var SE	33.55	
% var acc. for (V%)	33.60	
90% CI	(-0.84) – (0.33)	
Fail safe <i>N</i>	102	

Note.

^a Excluding Dinella (2004), *N* = 232.

<u>Domain identification</u>			
<u>Women test takers</u>			
	<u>High ^a</u>	<u>Medium</u>	<u>Low ^b</u>
<i>K</i>	8	<i>n/a</i>	3
<i>N</i>	228		105
Mean <i>d</i>	-0.41		-0.42
Var <i>d</i>	0.30		0.42
Var <i>e</i>	0.15		0.12
Mean δ	-0.44		-0.46
Var δ	0.18		0.36
% var SE	48.67		28.07
% var acc. for (V%)	48.79		28.17
90% CI	(-0.99) – (0.11)		(-1.23) – (0.31)
Fail safe <i>N</i>	41		16

Note.

^a Excluding Anderson (2001), *N* = 152.

^b Excluding Anderson (2001), *N* = 202.

Test Difficulty

Minority test takers

	<u>Difficult</u>	<u>Moderate</u>	<u>Easy</u>
K	n/a	n/a	n/a
N			
Mean d			
Var d			
Var e			
Mean δ			
Var δ			
% var SE			
% var acc. for (V%)			
90% CI			
Fail safe N			

Women test takers

	<u>Difficult</u>	<u>Moderate</u>	<u>Easy</u>
K	n/a	n/a	n/a
N			
Mean d			
Var d			
Var e			
Mean δ			
Var δ			
% var SE			
% var acc. for (V%)			
90% CI			
Fail safe N			

REFERENCES

- Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94-107.
- *Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., & Mitchell, J. P. (2004). Detecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology, 40*, 401-408.
- *Anderson, R. D. (2001). *Stereotype threat: The effects of gender identification on standardized test performance*. Unpublished dissertation. James Madison University, Harrisonburg, VA.
- Aronson, J. (2002). Stereotype threat: Contending and coping with unnerving expectations. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 279-301). San Diego, CA: Academic Press.
- Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38*, 113-125.
- *Aronson, J., Lusting, M.J., Good, C., Keough, K., Steele, C.M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35*, 29-46.
- *Bailey, A. A. (2004). *Effects of stereotype threat on females in Math and Science fields: An investigation of possible mediators and moderators of the threat-performance relationship*. Unpublished dissertation. Georgia Institute of Technology, GA.
- Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer, Jr. (Ed.), *Advances in social cognition* (Vol. 10, pp. 1-61). Mahwah, NJ: Erlbaum.
- Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41*, 174-181.
- Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science, 12*, 225-229.
- Bobko, P., Roth, P., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561-589.
- Brehm, J. W. (1966). *A theory of psychological reactance*. New York: Academic Press.

- *Brown, J. L., Steele, C. M., & Atkins, D. (Unpublished). *Performance expectations are not a necessary mediator of stereotype threat in African American verbal test performance.*
- *Brown, R. P., & Day, E. A. (in press). The difference isn't Black and White: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology.*
- *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.
- *Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology, 39*, 626-633.
- *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.
- *Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16*, 572-578.
- Cheryan, S., & Bodenhausen, G. V. (2000). When positive stereotypes threaten intellectual performance: The psychological hazards of "model minority" status. *Psychological Science, 11*, 399-402.
- *Cohen, G. L., & Garcia, J. (2005). "I Am Us": Negative Stereotypes as Collective Threats. *Journal of Personality and Social Psychology, 89*, 566-582.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (2002, June). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101.
- *Cotting, D. I. (2003). *Shedding light in the black box of stereotype threat: The role of emotion.* Unpublished dissertation. City University of New York, US.
- Croizet, J. -C., Desert, M., Dutrevis, M., & Leyens, J.-P. (2001). Stereotype threat, social class, gender, and academic under-achievement: when our reputation catches up to us and take over. *Social Psychology of Education, 4*, 295-310.
- Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*, 588-594.

- Croizet, J.-C., Despres, G., Gauzins, M.-E, Huguet, P., Leyens, J.-P, & Meot. A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin*, 30, 721-731.
- Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology*, 89, 220-230.
- *Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*, 28, 1615-1628.
- DeRouin, R. E., Fritzsche, B. A., & Salas, E. (2004, April). *Age, Stereotype threat, and training performance*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Chicago.
- Devos, T., & Banaji, M. R. (2003). Implicit self and identity. In M. R. Leary, & J. P. Tangney (Eds.). *Handbook of self and identity* (pp. 153-175). New York: The Guildford Press.
- *Dinella, L. M. (2004). *A developmental perspective on stereotype threat and high school mathematics*. Unpublished dissertation. Arizona State University, US
- *Dodge, T. L., Williams, K. J., & Blanton, H. (2001, April). *Motivational mediators of the stereotype threat effect*. Paper presented at the Society for Industrial and Organizational Psychology, San Diego, CA.
- *Edwards, B. D. (2004). *An examination of factors contributing to a reduction in race-based subgroup differences on a constructed response paper-and-pencil test of achievement*. Unpublished dissertation. Texas A&M University, TX.
- Egger, M., & Smith, G. D. (1998, January 3). Meta-analysis: Bias in location and selection of studies. *Biomedical Journal* (online). Retrieved in April 2006 from <http://bmj.bmjjournals.com/archive/7124/7124ed2.htm>.
- *Elizaga, R. A., & Markman, K. D. (Unpublished). *Peers and performance: How in-group and out-group comparisons moderate stereotype threat effects*.
- Essed, P. J. M. (1991). *Understanding everyday racism*. Newbury Park, CA: Sage.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- *Foels, R. (1998, June). *Women's math ability: An investigation of stereotype threat*. Poster presented at the Society for the Psychological Study of Social Issues conference: Ann Arbor, MI.

- *Foels, R. (2000, February). *Disidentification in the face of stereotype threat*. Paper presented at the Society for Personality and Social Psychology conference: Nashville, TN.
- *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.
- Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H. & Hart, A. (2004). A threat in the computer: The race Implicit Association Test as a stereotype threat experience. *Personality and Social Psychology Bulletin*, 30, 1611-1624.
- *Gamet, M. M. (Unpublished) *Stereotype threat and the effects of women in mathematical tasks*.
- Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3-8.
- Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on test performance of Latino women. *Personality and Social Psychology Bulletin*, 28, 659-670.
- Good, C., Aronson, J. Inzlicht, M. (2003). Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, 24, 645-662.
- *Gresky, D. M., Eyck, L. L. T., & Lord, C. G. (Unpublished). *Effects of salient multiple identities on women's performance under mathematics stereotype threat*.
- *Guajardo, G. A. (2005). *Modifying stereotype relevance and altering affect attributions to reduce performance suppression on cognitive ability selection tests*. Unpublished thesis. Northern Illinois University, IL.
- *Harder, J. A. (2000). *The effect of private versus public evaluation on stereotype threat for women in mathematics*. Unpublished dissertation. University of Texas at Austin, US.
- Hess, T. M.; Hinson, J. T. & Statham, J. A. (2004). Explicit and implicit stereotype activation effects on memory: Do age and awareness moderate the impact of priming? *Psychology and Aging*, 19, 495-505.
- Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 136-168). New York: Guilford Press.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18, 117-144.

- Hunter, J.E. & Schmidt, F.L. (1990). *Methods of meta-analysis: correcting error and bias in research findings*. Newbury Park (CA): SAGE Publications.
- Hyde, J. S., & Kling, K. C. (2001). Women, motivation, and achievement. *Psychology of Women Quarterly*, 25, 364-378.
- 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 implication of threatening environments on intellectual processing. *Journal of Educational Psychology*, 95, 796-805.
- *Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: teaching stereotype threat as a means of improving women's math performance. *Psychological Science*, 16, 175-179.
- *Josephs, R. A., Newman, M. L., Brown, R. P., & Beer, J. M. (2003). Status, testosterone, and human intellectual performance: Stereotype threat as status concern. *Psychological Science*, 14, 158-163.
- Judge, T. A., Colbert, A., & Ilies, R. (2004). Intelligence and leadership: a quantitative review and test of theoretical propositions. *Journal of Applied Psychology*, 89, 542-552.
- *Keller, J. (2002). Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles*, 47, 193-198.
- *Keller, J. (In press). 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*.
- *Keller, J., & Bless, H. (Unpublished). *When positive and negative expectancies disrupt performance: Regulatory focus as a catalyst*.
- *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.
- Koslowsky, M., & Sagie, A. (1993). On the efficacy of credibility intervals as indicators of moderator effects in metaanalytic research. *Journal of Organizational Behavior*, 14, 695-699.
- Kray, L. J., Galinsky, A. D., & Thompson, L. (2002). Reversing the gender gap in negotiations: An exploration of stereotype regeneration. *Organizational Behavior and Human Decision Processes*, 87, 386-409.

- Lalonde, R. N., & Cameron, J. E. (1994). Behavioral responses to discrimination: A focus on action. In M. P. Zanna & J. M. Olson (Eds.), *The psychology of prejudice: The Ontario Symposium, Volume 7* (pp. 257-288). Hillsdale, NJ: Lawrence Erlbaum.
- Landis, J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Levy, B. (1996). Improving memory in old age by implicit self-stereotyping. *Journal of Personality & Social Psychology*, 71, 1092-1107.
- *Lewis, P. B. (1998). *Stereotype threat, implicit theories of intelligence, and racial differences in standardized test performance*. Unpublished dissertation. Kent State University, OH.
- Leyens, J.-P., Desert, M., Croizet J.-C., & Darcis, C. (2000). Stereotype threat: Are lower status and history of stigmatization preconditions of stereotype threat? *Personality and Social Psychology Bulletin*, 26, 1189-1199.
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist*, 48, 1181-1201.
- *Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, 42, 236-243.
- *Martin, D. E. (2004). *Stereotype threat, cognitive aptitude measures, and social identity*. Unpublished dissertation. Howard University, DC.
- *Marx, D. M., & Stapel, D. A. (2005). It's all in the timing: Measuring emotional reactions to stereotype threat before and after taking a test. *European Journal of Social Psychology*, 35, 1-12.
- *Marx, D. M., & Stapel, D. A. (In press). Distinguishing stereotype threat from priming effects: On role of the social self and threat-based concerns. *Journal of Personality and Social Psychology*.
- *Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: the interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology*, 88, 432-446.
- Mayer, D. M., & Hanges, P. J. (2003). Understanding the stereotype threat effect with "culture-free" tests: An examination of its mediators and measurement. *Human Performance*, 16, 207-230.

- *McFarland, L. A., Kemp, C. F., Viera, Jr. L., & Odin, E. P. (2003, April). *Stereotype threat and male-female differences in test performance*. Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
- *McFarland, L. A., Lev-Arey, D. M., & Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance*, 16, 181-205.
- *McIntyre, R. B., Lord, C. G., Gresky, D. M., Eyck, L. L. T., Frye, G. D. J., & Bond, C. F., Jr. (2005). A social impact trend in the effects of role models on alleviating women's mathematics stereotype threat. *Current Research in Psychology*, 10, 116-136.
- *McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology*, 39, 83-90.
- *McKay, P. F. (1999). *Stereotype threat and its effect on the cognitive ability test performance of African-Americans: The development of a theoretical model*. Unpublished dissertation. University of Akron, US
- McKay, P. F., Doverspike, D., Bowen-Hilton, D., & Martin, Q. D. (2002). Stereotype threat effects on the Raven Advanced Progressive Matrices Scores of African Americans. *Journal of Applied Social Psychology*, 32, 767-787.
- Mendoza-Denton, R., Downey, G., Purdie, V., Davis, A., & Pietrzak, J. (2002). Sensitivity to status-based rejection: Implications for African American students' college experience. *Journal of Personality and Social Psychology*, 83, 896-918.
- Mendoza-Denton, R., Page-Gould, E., & Pietrzak, J. (2005). Mechanisms for coping with race-based rejection expectations. In S. Levin & C. van Laar (Eds.), *Stigma and group inequality: Social psychological approaches*. New York: Erlbaum.
- Miller, C. T., & Kaiser, C. R. (2001). A theoretical perspective on coping with stigma. *Journal of Social Issues*, 57, 73-92.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246-268.
- *Nguyen, H.-H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance*, 16, 261-294.
- *Nguyen, H. -H. D., Shivpuri, S., Ryan, A. M., & Langset, K. (2004, April). *Relations of stereotype threat effects to assessment domains and self-identity*. Paper presented at the annual meeting of the Society for Industrial-Organizational Psychology.

- Nosek, B., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math \neq me. *Journal of Personality and Social Psychology*, 83, 44-59.
- *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.
- Osborne, J. W. (2001a). Unraveling underachievement among African American boys from an identification with academic perspective. *The Journal of Negro Education*, 68, 555-565.
- Osborne, J. W. (2001b). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology*, 26, 291-310.
- *Oswald, D. L., & Harvey, R. D. (2000/2001). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology: Developmental, Learning, Personality, Social*, 19, 338-356.
- *Pellegrini, A. V. (2005). *The impact of stereotype threat on intelligence testing in Hispanic females*. Unpublished dissertation. Carlos Albizu University, FL.
- *Philipp, M. C., & Harton, H. C. (2004). *The role of social dominance in stereotype threat effects*. Paper presented at the meeting of the Society for Personality and Social Psychology.
- Pietrzak, J., Downey, G., & Ayduk, O. (2005). Rejection sensitivity as an interpersonal vulnerability. In M. W. Baldwin (Ed.), *Interpersonal Cognition*. New York: Guilford Publications.
- Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology*, 76, 114-128.
- *Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance*, 16, 231-259.
- *Prather, H. M. (2005). *Controlling the threat of stereotypes: The effectiveness of mental control strategies in increasing female math ability test performance*. Unpublished dissertation. George Washington University, Washington, DC.
- Prime, J. L. (2000). *An exploratory look at the relationship between racial identification and stereotype threat effects*. Unpublished dissertation.
- Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues*, 57, 55-71.

- *Rivadeneyra, R. (2002). *The influence of television on stereotype threat among adolescents of Mexican descent*. Unpublished dissertation. University of Michigan, MI.
- Roberson, L., Deitch, E. A., Brief, A. P., & Block, C. J. (2003). Stereotype threat and feedback seeking in the workplace. *Journal of Vocational Behavior*, 62, 176-188.
- *Rosenthal, H. E. S., & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Journal of Personality and Social Psychology*, 32, 501-511.
- Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, 1, 331-340.
- Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, 11, 446-453.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S. III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, 54, 297-330.
- Rowley, S., Sellers, R. M., Chavous, T. M., & Smith, M. A. (1997). The relationship between racial identity and self-esteem in African American college and high school students. *Journal of Personality and Social Psychology*, 74, 715-724.
- Ryan, A. M. (2001). Explaining the Black-White test score gap: The role of test perceptions. *Human Performance*, 14, 45-76.
- Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance*, 16, 295-309.
- Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2005). On interpreting research on stereotype threat and test performance. *American Psychologist*, 60, 271-272.
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56, 302-318.
- *Salinas, M. F. (1998). *Stereotype threat: The role of effort withdrawal and apprehension on the intellectual underperformance of Mexican-Americans*. Unpublished dissertation. University of Texas at Austin, TX

- *Sawyer, T. P., Jr., & Hollis-Sawyer, L. A. (2005). Predicting stereotype threat, test anxiety, and cognitive ability test performance: An examination of three models. *International Journal of Testing*, 5, 225-246.
- *Schimel, J., Arndt, J., Banko, K. M., & Cook, A. (2004). Not all self-affirmations were created equal: The cognitive and social benefits of affirming the intrinsic (vs. extrinsic) self. *Social Cognition*, 22, 75-99.
- *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.
- *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.
- Schmidt, F. L., & Le, H. A. (2004). *The Hunter-Schmidt meta-analysis programs package*, version 1.1 (revised Oct. 2005).
- *Schneeberger, N. A., & Williams, K. (2003, April). *Why women "can't" do math: The role of cognitive load in stereotype threat research*. Paper presented at the 18th meeting of the Society for Industrial-Organizational Psychology. Orlando, FL.
- *Schultz, P. W., Baker, N., Herrera, E., & Khazian, A. (Unpublished). *Stereotype threat among Hispanic-Americans and the moderating role of ethnic identity*.
- *Seagal, J. D. (2001). *Identity among members of stigmatized groups: A double-edged sword*. Unpublished dissertation, University of Texas at Austin, US
- Seibt, B., & Forster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory foci. *Journal of Personality and Social Psychology*, 87, 38-56.
- *Sekaquaptewa, D., & Thompson, M. (2002). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*, 39, 68-74.
- Sellers, R. M., Rowley, S. A., Chavous, T. M., Shelton, J. N., & Smith, M. A. (1997). Multidimensional Inventory of Black Identity: A preliminary investigation of reliability and construct validity. *Journal of Personality and Social Psychology*, 73, 805-815.
- Shelton, J. N., & Sellers, R. M. (2000). Situational stability and variability in African American racial identity. *Journal of Black Psychology*, 26, 27-50.

- Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80-83.
- *Smith, C. E., & Hopkins, R. (2004). Mitigating the Impact of Stereotypes on Academic Performance: The Effects of Cultural Identity and Attributions for Success among African American College Students. *Western Journal of Black Studies*, 28, 312-321.
- Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review*, 16, 177-206.
- *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., Fein, S., Strahan, E. J., & Zanna, M. P. (2005). The role of motivation in the unconscious: How our motives control the activation of our thoughts and shape our actions. In K. D. Williams & J. P. Forgas (Eds.), *Social motivation: Conscious and unconscious processes*. (pp. 113-129). New York: Cambridge University Press.
- *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.
- *Spencer, S. L. (2005). *Stereotype threat and women's math performance: The possible mediating factors of test anxiety, test motivation and self-efficacy*. Unpublished dissertation, Rutgers The State University of New Jersey, New Brunswick, US.
- *Spicer, C. V. (1999). *Effects of self-stereotyping and stereotype threat on intellectual performance*. Unpublished dissertation. University of Kentucky, US.
- Stangor, C., Carr, C., & Kiang, L. (1998). Activating stereotypes undermine task performance expectations. *Journal of Personality and Social Psychology*, 75, 1191-1197.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613-629.
- Steele, C. M. (1999, August). Thin ice: Stereotype threat and black college students. *The Atlantic Monthly*, 284(2), 44-54.
- *Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797-811.
- Steele, C.M., & Aronson, J. (1998). Stereotype threat and the test performance of academically successful African Americans. In C. Jencks & M. Phillips (Eds.),

- The Black-White test score gap* (pp. 401-427). Washington, DC: Brookings Institution.
- Steele, C.M., & Davies, P.G. (2003). Stereotype threat and employment testing: A commentary. *Human Performance*, 16, 311-326.
- Steele, C.M., Spencer, S.J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 379-440). San Diego, CA: Academic Press.
- Stein, R., Blanchard-Fields, F., & Hertzog, C. (2002). The effects of age-stereotype priming on memory performance in older adults. *Experimental Aging Research*, 28, 169-181.
- *Sternberg, R. J., Jarvin, L., Leighton, J., Newman, T., Moon, T., Callahan, C., & Grigorenko, E. L. (Unpublished). *Girls can't do math?: The disidentification effect and gifted high school students' math performance*.
- Stone, J., Lynch, C., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White Athletic Performance. *Journal of Personality and Social Psychology*, 77, 1213-1227.
- Stricker, L. J., & Bejar, I. I. (2004). Test difficulty and stereotype threat on the GRE General test. *Journal of Applied Social Psychology*, 34, 563-597.
- *Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and sex, and standardized test performance. *Journal of Applied Social Psychology*, 34, 665-693.
- *Tagler, M. J. (2003). *Stereotype threat: Prevalence and individual differences*. Unpublished dissertation. Kansas State University, US
- Thalheimer, W., & Cook, S. (2002, August). How to calculate effect sizes from published research articles: A simplified methodology. Retrieved May 11, 2004, from http://work-learning.com/effect_sizes.htm.
- Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.
- *van Dijk, A., Koenders, H., Korenhof, I. H., Mulder, H. R., & de Vries, H. (Unpublished). *The moderating role of group membership activation on stereotype lift and threat*.
- Vandenberg, R. B., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-69.

- *von Hippel, W., von Hippel, C., Conway, L., Preacher, K. J., Schooler, J. W., & Radvansky, G. A. (2005). Coping With Stereotype Threat: Denial as an Impression Management Strategy. *Journal of Personality and Social Psychology*, 89, 22-35.
- *Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles*, 41, 219-240.
- *Walters, A. M. (2000). *Stereotype threat: An examination of process*. Unpublished dissertation. University of Florida.
- Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology*, 39, 456-467.
- Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, 132, 249-268.
- Wheeler, S., C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin*, 127, 797-826.
- Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, 75, 315-321.
- *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.
- Williams, J. G. (2004). *Forewarning: A tool to disrupt stereotype threat effects*. Unpublished dissertation. University of Texas at Austin, TX.
- *Wout, D. A., Shih, M. J., Jackson, J. S., & Sellers, R. M. (Unpublished). *Targets as perceivers: How Blacks determine if they will be stereotyped*.

MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 02845 4589