Stereotype Threat in Black College Students Across Many Operationalizations

Patrick S. Forscher*[1], Valerie Jones Taylor*[2], Daniel R. Cavagnaro[3], Neil A. Lewis, Jr.[4], Erin Buchanan[5], Hannah Moshontz[6], Aimee Y. Mark[7], Sara C. Appleby[8], Carlota Batres[9], Brooke Bennett-Day[10], William J. Chopik[11], Rodica Ioana Damian[12], Claire E. Ellis[13], Caitlin Faas[14], Sarah E Gaither[15], Dorainne Green[16], Braeden F. Hall[17], Bianca Marie Hinojosa[18], Jennifer L. Howell[18], David C. Johnson[19], Franki Y. H. Kung[20], Angela R. Laird[21], Carmel A Levitan[22], Manyu Li[23], Keith B. Maddox[24], Mary C. Murphy[25], Erica D. Musser[21], Brianna Pankey[21], Laura Ruth Murry Parker[26], Sylvia P Perry[27], Jessica D. Remedios[24], Kathleen Schmidt[17], Surizaday Serrano[12], Crystal N. Steltenpohl[17], Daniel Storage[28], Brenda C. Straka[15], Heather L. Urry[24], Samuel C Wasmuth[12], Erin C. Westgate[29], John Paul Wilson[30], Shelby Wynn[31], David M. Zimmerman[31], Kim Peters[32], Christopher R. Chartier[33]

*Corresponding Authors: Patrick S. Forscher (schnarrd@gmail.com) and Valerie Jones Taylor (vjones.taylor@gmail.com). These two authors contributed equally to this project.

[1]Université Grenoble Alpes, [2]Lehigh University, [3]California State University, Fullerton, [4]Cornell University, [5]Missouri State University, [6]Duke University, [7]University of Southern Indiana, [8]Mercer University,[9]Franklin and Marshall University, [10]Wesleyan College, [11]Michigan State University,[12]University of Houston, [13]University of Southern Indiana, [14]Mount St. Mary's University, [15]Duke University, [16]Indiana University, [17]Southern Illinois University Carbondale, [18]University of California, Merced, [19]York College and Graduate Center, CUNY, [20]Purdue University Systems, [21]Florida International University, [22]Occidental College, [23]University of Louisiana at Lafayette, [24]Tufts University, [25]Indiana University, [26]University of Houston-Downtown, [27]Northwestern University, [28]University of Denver, [29]The Ohio State University, [30]Montclaire State University, [31]Missouri State University, [32]University of Queensland, [33]Ashland University

Abstract:

According to stereotype threat theory, the possibility of confirming a negative group stereotype evokes feelings of threat, leading people to underperform in domains where they are stereotyped as lacking ability. This theory has important theoretical and practical implications. However, many studies supporting it include small samples and varying operational definitions of "stereotype threat". We address the first challenge by leveraging a network of psychology labs to recruit a large Black student sample ($N_{\text{anticipated}}$ = 2700) from multiple US sites ($N_{\text{anticipated}}$ = 27). We address the second challenge by identifying three threat-increasing and three threat-reducing procedures that could plausibly affect performance and use an adaptive Bayesian design to determine which operationalization yields the strongest evidence for underperformance. This project should advance our knowledge of a scientifically and socially important topic: the conditions under which stereotype threat affects performance among current Black students in the United States.

Main Text:

In 1954, Earl Warren, Chief Justice of the United States Supreme Court, issued the majority opinion in the landmark Brown v. Board of Education case that ordered the racial integration of American schools. Brown was intended to equalize US educational opportunities, but its effects have fallen short of this aspiration [1]. Some schools integrated, but the experiences of students within those schools were, and still are, far from equal [2,3]. One source of these different experiences is the presence of stereotypes that some students are less intelligent than others. In US schools, stereotypes that Black students are unintelligent have been central in American education discourse since at least the mid-20th century [4]. These stereotypes create a challenge for Black students that many other students do not face: poor performance on tasks that are diagnostic of intelligence can be construed as confirming the Black unintelligence stereotype [5].

Stereotype threat theory posits that concerns arising from the possibility of confirming a negative stereotype are consequential because they can provoke feelings of threat [6,7]. To the extent that these feelings of threat divert people's attention away from task performance [8], the experience of stereotype threat can hinder the performance of group members on the very tasks on which they are stereotyped as lacking ability [6]. Although stereotype threat theory has enjoyed attention from both educators and policy-makers and has even been cited in briefs to the US Supreme Court (e.g., Fisher v. University of Texas [9]), the scientific community is conflicted about the conditions under which stereotype threat adversely impacts student performance. This project aims to provide evidence that will hopefully help resolve some of these questions, particularly with respect to the current population of Black students in the United States.

Stereotype threat theory is formulated broadly: any group that is negatively stereotyped on a particular task could potentially suffer stereotype threat's negative consequences, and any situational cue that makes a negative group stereotype salient could provoke feelings of threat [6]. However, the theory also predicts that not all performance tasks will give rise to stereotype threat, nor are all people equally vulnerable to its pernicious effects. Early formulations of the theory posited three factors that stand to influence the stereotype threat effect: stereotype relevance, task difficulty, and domain identification. Stereotype threat should impact performance if the task is **self-relevant**, that is, "the possibility of conforming to the stereotype or of being treated and judged in terms of it—becomes self-threatening", Steele [6], pg. 617. Furthermore, stereotype threat should only occur if a task is sufficiently **difficult** to bring about the possibility of poor performance [7,10]. In addition, people should experience stereotype threat most acutely when they **identify** with the domain in which they are being evaluated [7].

Since the early formulations of the theory, researchers have identified other potential exacerbating and limiting conditions. For example, people who are **chronically concerned** about the possibility of confirming negative stereotypes may be especially vulnerable to stereotype threat [11], while people who identify strongly with their racial or ethnic identity may be less vulnerable [12,13]. Black students in the US are themselves not monolithic, differing in ethnic background, family immigration history (forced or voluntary), and generation status, and any of these varying characteristics may impact the size of the stereotype threat effect. Finally, a broad array of other contextual factors could also make stereotype threat more or less likely, such as characteristics of the experimenter (e.g., Black students may experience more stereotype threat if the experimenter is White [14]) and the institution at which the experiment was conducted (e.g., Black students at minority-serving institutions may experience less stereotype threat [15]).

Owing to the theory's broad formulation, researchers have used a large array of procedures to increase and reduce feelings of threat. The threat-increasing procedures range from telling participants that the task they are about to complete measures the stereotyped ability (a diagnosticity prompt [10]), to informing participants that their group typically underperforms on the task they are about to complete (a group differences prompt [16]), to reminding participants of their negatively stereotyped group membership before they complete the task (a group-based prime [10,17]). Threat-reducing procedures also vary, ranging from telling participants that the task is not diagnostic of the stereotyped ability (a non-diagnostic prompt [10]) to telling participants that their group performs just as well as any other group on the upcoming task (a no group differences prompt [16]), to pairing participants with a high-status member of their group to whom they might identify (e.g., role models [18]). In a given study, stereotype threat is operationalized by comparing the performance of participants in a threat-increasing procedure to their performance in a threat-reducing procedure. Although any threat-increasing procedure can be compared to any threat-reducing procedure, in practice, researchers usually focus on procedures that manipulate the same conceptual variable (e.g., the diagnosticity variable by comparing diagnostic and non-diagnostic conditions).

Also owing to the theory's broad formulation, stereotype threat theory has been applied to many different populations, each of which faces its own set of negative stereotypes. These populations range from the elderly, whose performance on cognitive tasks might be impaired by the stereotype that older people are forgetful [19,20], to women, whose performance on math tests might be impaired by the stereotype that women are bad at math [16], and to students of low socioeconomic status (SES), whose performance on intelligence tasks might be impaired by the stereotype that low SES students are unintelligent [21]. However, the theory was originally formulated to help explain and address barriers that prevent members of historically disadvantaged US groups from fulfilling their potential, especially Black students on intelligence tasks. For this reason, it is somewhat surprising that, in a recent unpublished meta-analysis of stereotype threat research, only a small minority of stereotype threat studies focus on Black students (58/323 = 18%; Taylor, Forscher, and Walton). This oversight may be partly caused by pragmatic concerns: Black people constitute only 13% of students in US higher education [22], and an even smaller share of the student body at research-active universities [22,23] and are therefore harder for researchers to recruit than members of other groups, such as women in STEM fields.

Stereotype threat theory has many pragmatic implications. Due to its broad theoretical formulation, the theory could help explain ongoing and persistent gaps between a variety of social groups, ranging from the achievement gap between Black and White students in the United States [24] (or, alternatively, the "opportunity gap [25]) to the gap in the number of women and men who opt into STEM fields [26]. Insofar as stereotype threat contributes to these ongoing gaps, stereotype threat theory also offers a potential route to reducing them: implement strategies that reduce or eliminate the threat to group members of confirming negative stereotypes [27,28]. Consistent with this reasoning, stereotype threat research has inspired the development of a broad array of strategies intended to boost the performance of members of stereotyped groups [16,29,30]. Stereotype threat theory also has many theoretical implications, as its flexibility and broad formulation allows its application to a broad range of research domains, from education to social cognition, thereby building bridges between these disparate research areas [31,32].

The combination of theoretical and pragmatic importance has led to an avalanche of research examining the stereotype threat effect and the contexts and people among whom it is strongest. This work has generally supported the notion that the magnitude of the effects of threat on performance varies by characteristics of the methods used to induce it [33] and the sample under investigation [34,35]. Thus, until recently, the consensus was that stereotype threat is robust but sensitive to the populations and methods under study.

However, this consensus has recently been questioned. Because stereotype threat is a theory about how specific situations affect specific subgroups of people, many studies have used smaller samples (median n = 52; unpublished meta-analysis by Taylor, Forscher, and Walton) than research on topics without these restrictions. In small samples, effects are estimated imprecisely. By itself, imprecision is not a problem, as long as the literature contains multiple imprecise studies that can be synthesized into a more precise aggregate estimate. However, imprecision can lead to a misleading literature when combined with meta-scientific processes that lead to the selection of significant results at the expense of non-significant ones. In small samples, effects only reach significance when they are very large; in the presence of processes like publication bias that suppress non-significant results, overreliance on small samples can, therefore, result in a literature that gives a distorted view of the true population effect [36]. Meta-analytic tests for small-study bias suggest this problem may be true of subsets of the stereotype threat literature [35,37,38]. Moreover, two recent large-scale studies of the effects of stereotype threat on women taking math tests have found small to near-zero effects of threat on performance [26,39]. Taken together, recent meta-analytic and large-study evidence have given some scholars varying degrees of doubt about the size of a stereotype threat effect on performance [31,35].

The overreliance on small samples may also be a problem in combination with a feature that is in other ways a strength of stereotype threat research: the aforementioned variation in how stereotype threat is operationalized. Because any threat-increasing procedure can hypothetically be compared to any threat-reducing procedure to operationalize stereotype threat, the number of available operationalizations grows multiplicatively with the number of threat-increasing and threat-reducing procedures. For example, given four threat-increasing and four threat-reducing procedures, there are 16 possible ways to compare a threat-increasing procedure to a threat-reducing procedure, yielding 16 possible operationalizations of “stereotype threat”. Researchers have tested far more than four threat-increasing and four threat-reducing procedures [33]. The sheer variety of procedures yields a combinatorial explosion of potential ways to operationalize stereotype threat.

Variations in a construct’s operationalization benefit a scientific theory because they broaden the domains to which the theory applies [40,41]. However, when considering psychological theories, such variations can also introduce uncertainty: each new operationalization brings with it the possibility that the operationalization may not evoke the same psychological process as the previous ones [42,43]. Thus, some studies framed as investigating “stereotype threat” (and which therefore could be considered evidence for or against the theory) may not in fact be investigating the same “stereotype threat” as studies that use other operationalizations. The uncertainty is magnified when each operationalization is tested with a relatively small sample. The varying operationalizations of “stereotype threat” have therefore made it difficult to uniformly assess to what extent and in what populations “stereotype threat” produces a measurable and even robust impact on performance.

This diversity in operationalizations has had a second important consequence: some of the operationalizations may not validly capture “stereotype threat.” In many stereotype threat studies, the “threat-reducing” condition to which the “threat-increasing” condition is compared does not actually reduce or eliminate the threat of confirming the target negative stereotype. For example, in an unpublished meta-analysis by Taylor, Forscher, and Walton, 152/323 (47%) of samples compared a threat-increasing condition to an evaluative “threat-reducing” procedure in which participants were told their task measures a negatively stereotyped ability. This evaluative procedure, at times used as a “threat-reducing” operationalization, could itself increase feelings of threat: participants could reasonably infer that poor performance on an evaluative task confirms the negative stereotype [10,16]. Thus, the evaluative “threat-reducing” condition could have performance impacts that are similar to ones that most researchers believe are threat-increasing. A valid operationalization of stereotype threat requires a comparison between a threat-increasing procedure and a procedure that clearly decreases feelings of threat.

The current study has two primary aims. First, we will address past issues with sample size in the selection of the target population by recruiting a large sample from a US population that has experienced historical and social disadvantage and that was the focus of early stereotype threat research – Black college students. Second, we will address the methodological variation in this literature by simultaneously testing three procedures that ought to increase stereotype threat (i.e., diagnosticity prompt, group differences prompt, group-based prime) and three that ought to decrease it (i.e., non-diagnostic prompt, no group differences prompt, no group differences prompt communicated by a Black expert). We will also test a series of theoretically motivated moderators expected to impact performance for those who experience stereotype threat: domain identification (both general and task-specific) [7], chronic concern about stereotype threat [11], and racial/ethnic identification [12]. Finally, we hold two theoretically important variables, stereotype relevance and task difficulty, constant at high levels. We lay our focal study hypotheses in the Method, after we have described our detailed procedures to operationalize stereotype threat.

We will accomplish these aims by leveraging two major methodological innovations. First, we will gain access to a sufficient number of Black students to make our design informative by tapping into the network of labs provided through the Psychological Science Accelerator [44]. Second, we will use a so-called “adaptive design”, which optimizes how participants are allocated to conditions in order to efficiently seek condition pairs providing either positive or negative evidence for a stereotype threat effect in a current sample of Black undergraduates in the US. More specifically, we will prioritize seeking positive evidence for stereotype threat to address concerns about the weakness of past positive evidence and to minimize the possibility of false positives (see our simulation studies for details). Taken together, this work seeks to contribute to the extant stereotype threat literature by providing a robust test of the effect and its potential moderators among a large sample population of Black students, for whom such work is both important and urgent.

Method

Ethics information

All labs that are contributing to the data collection efforts have obtained ethics approval from their local IRBs at the time of writing of this writing. All participants will provide informed consent; some will receive course credit, while others will be paid for their time. Each site's IRB protocols with the relevant ethics details is at https://osf.io/myxuc/; documentation of the acceptance of each protocol is at https://osf.io/64g8n/.

Materials availability

All materials, ethics approvals, analysis code, simulation evidence, and our initial proposal to the Psychological Science Accelerator, are deposited at our project page at https://osf.io/7tgav/.

Participants and sites

To adequately test stereotype threat theory, we must recruit a population that could reasonably experience stereotype threat on a particular task. We have chosen self-identified Black college students in the United States for our population and intelligence tests as our task. Most Black undergraduate college students in the US should sufficiently identify with intelligence to be threatened by the stereotype that Black people are unintelligent, incompetent, or dumb [45]. Likewise, most Black college students should also identify with their racial group due to psychological processes such as optimal distinctiveness [46] and the shared experience of discrimination [47]. In part for these reasons, the first published stereotype threat studies tested whether the threat of confirming the Black unintelligence stereotype affects Black students [10]. However, the relative rarity of Black college students at research active universities does raise some feasibility concerns.

To address this issue, we have recruited 27 labs throughout the United States to participate in this project as collaborators through the standing network of psychology labs provided by the Psychological Science Accelerator (PSA) [44]. The PSA maintains a worldwide database of labs that have expressed the interest and ability to collaborate on multi-site projects and provides scientific and administrative support to accomplish such studies. Initial calls were sent for collaborators to the labs based in the United States in the Accelerator network, as well as solicitations through Twitter, the PsychMAP and PsychMAD Facebook groups, and personal networks in the Fall of 2018.

Each site has drafted a plan for recruiting a sample of Black college students. Each site will either rely on a local pool of Psychology students who will complete the study for course credit, a combination of flyers and other advertising to recruit students willing to complete the study for payment, or both (we will record site-specific recruitment details and conduct robustness checks to assess whether they influence results). Each site has provided an estimate of the number, based on their knowledge of local demographics and other conditions, of Black students they could feasibly recruit for this study over the course of a year. To be sure, some of our participants may come from institutions with a large proportion of Black students. This, as well as other institutional characteristics (noted below in descriptions of experimenter and site variables), may impact the size of the threat effect. Summing across sites, we estimate that we could feasibly obtain a sample of 2,700 Black students; see our Supplemental Method for details.

Measures

The measures described below are drawn from the broader literature on stereotype threat (specific citations are discussed with each measure). When available, we describe information about the reliability and validity of the measures. However, there are two important caveats for interpreting this information. First, consistent with other areas of social and personality psychology research [48], not all stereotype threat studies report reliability and validity information. Second, as noted above, only 18% of stereotype threat studies have focused on Black students [49]. Our knowledge about whether previously validated measures remain valid with the current sample is therefore limited.

Task performance measure. The primary outcome measure for assessing the stereotype threat effect is Raven's Advanced Progressive Matrices [50], a test of fluid intelligence intended for use with people with above average aptitude and designed to reliably differentiate among those in the top 25% of the population [51]. The Advanced Progressive Matrices are also sufficiently difficult to provoke anxiety among college students [52], and have been used in stereotype threat research with Black college students specifically [52–54]. The Advanced Progressive Matrices consist of a series of perceptual analytic reasoning problems, each in the form of a matrix. The problems involve both horizontal and vertical transformations: figures may increase or decrease in size, and elements may be added or subtracted, flipped, rotated, or show other progressive changes in the pattern. In each case, the lower right corner of the matrix is missing and the participant's task is to determine which of eight possible alternatives fits into the missing space such that row and column rules are satisfied [51].

Multiple versions of the Advanced Progressive Matrices exist. In this study, we will use the short form, which has been validated by Bors and Stokes [51] and predicts performance on the full set of Ravens items [55,56]. The Advanced Progressive Matrices has 48 items, including 12 items in Set I and 36 items in Set II. Participants will complete four items in Set I as practice and up to 36 items in Set II as our primary performance measure. Participants will have a time limit of 40 minutes to complete the matrices, consistent with Brown and Day. We will measure performance by summing the number of correct responses in Set II, yielding a performance index that ranges from 0 to 36, with higher scores indicating better performance.

Potential moderators of the threat effect. This group of measures consists of variables that, through exploratory analysis, could help us test whether certain subsets of our participants are particularly affected by stereotype threat. Four of these moderators, domain identification (both general and task-specific), racial identification, and chronic concern about stereotypes, are derived from stereotype threat theory. The others (e.g., experimenter variables, site characteristics) could plausibly identify participants who are vulnerable to stereotype threat but are less central to the core theory.

Domain identification-general. Our primary measure of identification with intelligence will capture the extent to which students identify with the performance domain. We will ask participants to answer four questions adapted from Lewis, Sekaquaptewa, and Meadows [57] and Schmader [58]: "Being intelligent is an important part of my self-image"; "Being intelligent is unimportant to my sense of what kind of person I am" (reverse-coded); "Being intelligent has very little to do with how I feel about myself" (reverse-coded); and "Being intelligent is an important reflection of who I am." Participants will rate their level of agreement with these items on scales from 1 (strongly disagree) to 7 (strongly agree). We will measure domain identification by reverse-coding the appropriate items and averaging item responses to form a 1-7 composite, with higher scores indicating higher identification. Previous research suggests that responses on these items predict the size of the stereotype threat effect among women in mathematics [58].

Domain identification-task specific. Our inferences and interpretations will focus on the primary measure of domain identification because of the previous validation evidence suggesting that it is important in stereotype threat processes. However, Black undergraduates may strongly identify with being an intelligent person, but may not be strongly identified with scoring high on a particular test designed to assess intelligence. That is, they may value intelligence, but lack faith in "intelligence tests", given the history of the construction and use of intelligence tests in the US and associated negative racial stereotypes. Thus, we will include a secondary measure of domain identification that is more specific to the Raven's Matrices task. These questions will only be asked after the participants take Raven's Matrices and will be identical to the primary measure of domain identification except that they will replace "Being intelligent" with "Being good on intelligence tasks like the one I am taking today". We will use this secondary measure as part of our exploratory analyses.

Racial identification. We will measure racial identification using the Centrality, Private Regard, and Public Regard subscales of the Multidimensional Inventory for Black Identity (MIBI; [59]). We will use Centrality as our primary racial identification indicator but will conduct exploratory analyses with the Private and Public Regard subscales as well.

The eight-item Centrality subscale assesses how central, defining, and important one's racial group membership is to the self. Sample items include, "In general, being Black/African American is an important part of my self-image" and "Overall, being Black/African American has very little to do with how I feel about myself" (reversed). The six-item Private Regard subscale assesses "the extent to which individuals feel positively or negatively towards Blacks/African Americans as well as how positively or negatively they feel about being Blacks/African American" (pg. 26, Sellers and colleagues [59]). Sample private regard items include, "I am proud to be Black/African American" and "I often regret that I am Black" (*reversed*). The six-item Public Regard subscale assesses "the extent to which individuals feel that others view Blacks/African Americans positively or negatively" (pg. 26, Sellers and colleagues [59]). Sample items include, "Overall, Blacks/African Americans are considered good by others" and "Blacks/African Americans are not respected in the broader society" (reversed). We will measure racial identification by reverse-coding the appropriate items and averaging item responses in a 1-7 composite for each subscale, with higher scores indicating higher racial centrality, private regard, and public regard, respectively.

Chronic concern about stereotypes. To capture the experience of stereotype threat more broadly, we will also ask participants about the pressure they feel when doing something that would cause them to be seen in terms of stereotypes about their race. We designed two items to measure general stereotype concern: "I worry that people will sometimes make assumptions about me based on what they think about my racial group" and "I worry that people will sometimes make assumptions about me based on stereotypes about people in my racial group."

Participants will rate their level of agreement with these items using 1 (strongly disagree) to 7 (strongly agree) scales. To measure stereotype concern more broadly, we will average responses across the two items, with higher scores indicating greater concern.

Experimenter variables. Given that the group membership of the experimenter has itself been used as an operationalization of stereotype threat [14,60], it is critical that experimenter characteristics are tracked systematically in the current study. We will ask participating sites to assign each experimenter an ID and report each experimenter's race and gender, and will allow sites to freely report other experimenter variables that could possibly affect participants' experiences during the study. We will also limit experimenter interaction with participants as much as possible, to reduce the effect that interaction might have on participants.

Site variables. Past stereotype threat studies have not tracked systematically whether characteristics of the data collection site are associated with the strength of the stereotype threat effect. There are some reasons to believe that they might: highly ranked schools may be especially likely to have a student body that is domain-identified, which could enhance the stereotype threat effect [6]; a similar dynamic could characterize private (vs. public) schools. Moreover, schools with a lower proportion of minority students may undermine minority students' feelings that they belong in the school, which may also enhance the stereotype threat effect – in fact, solo status has itself been used as a stereotype threat manipulation [61]. The Psychological Science Accelerator maintains a database of the characteristics of its sites. Upon the completion of data collection, we will merge this database with our collected data to access these site-level characteristics.

Manipulation checks. Stereotype threat is theorized to occur when people are concerned about confirming a negative stereotype in a specific performance context. The performance task also needs to be sufficiently difficult to provide a real possibility that the stereotype will be confirmed. We are assuming that difficulty and task-evoked concern will be high among all participants in our study. We will validate this assumption using two manipulation checks. All manipulation checks will be administered at the end of the study session.

Task-evoked concern about stereotypes. To verify that participants are indeed experiencing task-evoked stereotype concern, we will ask participants to answer four questions adapted from Ramsey and colleagues [62]. Two of these questions are closely tied to perceptions of the test: "I am concerned that people will judge my race as a whole based on my performance on this test"; "I am concerned that people will think my race as a whole has less ability if I do not do well on this test". Two of these questions are tied to concerns about being judged in terms of group membership: "I am concerned that people will judge my performance based on negative stereotypes that exist about my racial group"; "I am concerned that people will think that I have less ability because of my racial group membership." Participants will rate their level of agreement with these items on a scale from 1 (strongly disagree) to 7 (strongly agree). We will average responses together, with higher scores indicating greater concern. We anticipate that, consistent with stereotype threat theory, the two task-evoked concern subscales will be strongly correlated, but to our knowledge this assumption has not been directly tested with Black students. We will therefore evaluate this correlation, and if the two subscales are modestly correlated ($r < .3$), we will test the effects of the threat manipulations on each subscale in our exploratory analyses.

Task difficulty. Raven's Advanced Progressive Matrices is designed to be difficult, producing a mean performance score of 22.17 (*SD* = 5.60) out of a theoretical maximum of 36 among 506 introductory psychology students at the University of Toronto at Scarborough [51]. Nevertheless, we will verify that the participants find the task difficult with a single item, "How difficult did you find the task that you completed today," on a scale from 1 (not at all difficult) to 5 (very difficult).

Potential exclusion criteria. Stereotype threat cannot occur unless participants are aware of the task-relevant stereotype and have paid close attention during the study. We will attempt to measure these variables and test whether excluding these people affects our results as part of a series of robustness checks, described in detail in our analysis plan.

Stereotype awareness. Participants will answer yes or no to a single item assessing awareness of the negative stereotypes about the intelligence of Blacks: "Before this study, had you ever heard of the stereotype that Blacks are less intelligent than other ethnicities?"

Memory checks. A series of questions will assess participants' memory for the details of the study. Items will include questions about the purpose of the study, the instructions provided prior to the performance task, and the type of task completed (i.e., puzzle, IQ test, etc.).

Funnel debriefing to probe for suspicion. At the end of the study, we will ask several questions capturing participants' suspicion about the aims of the study. Items assess whether participants 1) believed the rationale of the study, 2) had completed this type of task before, and if so, how many times, where was it taken, and their age when taken, 3) had heard of a study like this one, and 4) had ever heard of the phenomenon of stereotype threat prior to the study, and if so, when and where.

Demographics. Demographic items will include: age, biological sex, gender, class year (freshman through senior, other), academic major, academic minor, student status (full time, part time), ethnicity (all that apply; primary), citizenship, length of time in US, native language, state and country of birth, parents' places of birth, the number of grandparents born in the United States, generation status, city/state lived longest, socioeconomic status (parent's level of education and the MacArthur perceived SES ladder [63]), and employment status.

Procedure

Participants at between 18-21 sites for which it is locally feasible (not all labs have the necessary infrastructure to complete this process), will complete an online survey at least one week before their in-lab session. This pre-measure will include baseline measures of domain identification, racial identification, and an abbreviated demographics questionnaire. For many sites, these measures will be included in a battery of pre-measures administered to all students in qualifying psychology courses at the beginning of the semester.

For the main procedure, participants will come to their local lab site and complete an online survey in the lab. The survey will be completed in a quiet testing room to minimize distractions and standardize the amount of time spent on the task; at some sites, participants will complete the study in individual testing rooms, at other sites in larger testing rooms that have cubicles or computer dividers — this will be recorded as part of the site characteristics described earlier. The task (consent to debriefing) should take a maximum of 50 minutes. Each participant will be assigned to one of six conditions, three threat-increasing and three threat-reducing. The method of assigning participants will be an adaptive algorithm, which is described in more detail in the section entitled "Condition assignment through an adaptive design."

Following the threat-increasing or threat-reducing manipulation, participants will complete the focal task, Raven's Advanced Progressive Matrices. We plan for each participant to have a time limit of 40 minutes to complete the matrices. Following the focal task, participants will complete domain identification, racial identification, and stereotype threat concerns questionnaires, a series of memory and manipulation checks, demographic, stereotype awareness, and suspicion items. After the study is complete, the participants will be fully debriefed and asked to refrain from sharing the details of the study with others.

We determined the amount of session time through a feasibility pilot. We also used this feasibility pilot to ensure all study elements, including the adaptive algorithm, were properly implemented. We document this feasibility pilot in our Supplemental Method; proofs of concepts are at https://osf.io/tyasd/. Readers may view a mockup of the experiment implemented in the formr experiment platform [64] at https://psa005fullstudy.formr.org/?site=42

		Threat-increasing condition		Threat-reducing condition	
Conceptual variable	**Description**	*Name*	*Content*	*Name*	*Content*
Diagnosticity	Whether the participant believes the task measures the stereotyped ability	Diagnostic (*i1*)	Participants read that the task they're about to take is highly diagnostic of intelligence	Non-diagnostic (*r1*)	Participants read that the performance task is not diagnostic of intelligence
Priming	Whether the participant's group membership is made salient	Race primed (*i2*)	Participants are asked to indicate their race prior to taking the task		*Not applicable*
Group differences	Whether the participant believes there are group differences in task performance	Group differences (*i3*)	Participants read that White students outperform Black students on the task	No group differences (*r2*)	Participants read that White and Black students perform equally on the task
				No group differences, Black expert (*r3*)	A Black professor from a historically Black university delivers the no group differences prompt

Table 1. Table of threat-increasing and threat-reducing conditions for the current design. The threat-increasing conditions are labeled *i1-i3*, whereas the threat-reducing conditions are labeled *r1-r3*. To form an operationalization of stereotype threat, any threat-increasing condition can be compared to any threat-reducing condition, yielding nine possible operationalizations. We can pose a question about whether a threat effect is present for each operationalization; for example, question *i1_r2* asks whether a threat effect is present for the comparison between the diagnostic (*i1*) and the no group differences (*r2*) conditions.

Conceptual variables used to operationalize stereotype threat. Table 2 lists our three manipulated conceptual variables. We can use any pairing of one of the three threat-increasing conditions (diagnostic, race prime, group differences, or *i1, i2, i3*) and one of the three threat-reducing conditions (non-diagnostic, no group differences, no group differences-expert, or *r1*, *r2*, *r3*) to create an operationalization of stereotype threat, yielding nine possible operationalizations. Each operationalization can be designated by separating the code for the threat-increasing condition and the code for the threat-reducing condition with an underscore (e.g., *i1_r2* represents a comparison between the diagnostic condition and the no group differences condition).

Diagnosticity (conditions *i1* and *r1*). Diagnosticity refers to whether or not the target task is described as measuring the stereotyped characteristic (i.e., intelligence among Black students). Describing a task as diagnostic increases threat, as a task that is diagnostic of the stereotyped ability raises the specter of confirming the unintelligence stereotype by performing poorly on the task. The threat-increasing *diagnostic condition* (condition *i1*) therefore describes the task as evaluative of intellectual abilities:

> “The task that you will be working on today is an IQ test. The study is concerned with various personal factors involved in performance on problems requiring intellectual reasoning abilities. Like the SAT and the ACT, this test is frequently used to measure individuals' intellectual abilities. ...”

In contrast, a task that is persuasively described as non-diagnostic of the stereotyped ability decreases this threat [10]. In the threat-reducing non-diagnostic condition (condition *r1*), the task is described as non-evaluative of intellectual abilities:

> "In this research, we are studying a variety of puzzles for possible use in other research to understand how much people like them and find them interesting and involving. The items you'll complete today are just a series of puzzles. They don't, for example, have anything to do with intellectual ability or academic performance. …"

Priming (condition *i2*). "Priming" refers to whether the participant's stereotyped group membership is made salient prior to the performance task. The salience of this information should increase threat by increasing the likelihood that the participants think about their negatively stereotyped identity in the context of the performance task, thereby triggering stereotype threat (Steele and Aronson [10], Study 4). Thus, in the threat-increasing race primed condition (condition *i2*), participants will indicate their race prior to completing the performance task.

Group differences (conditions *i3*, *r2*, and *r3*). The "group differences" conceptual variable refers to whether the task is portrayed as producing or not producing group-based performance differences. If a participant is led to believe that group performance differences exist on a task, this raises the possibility that the participant's performance will recapitulate this pattern, thus confirming the unintelligence stereotype and increasing feelings of threat [16,65]. In the threat-increasing group differences condition (condition *i3*), the task is described as typically showing group differences:

> "As you may know, there has been some controversy about whether there are racial differences in intellectual and academic ability…The IQ test you will take today has been shown to produce racial differences, because such tests seem to be biased toward particular subcultural groups. Specifically, numerous studies have found that Blacks perform worse than Whites on such tests. ...".

By comparison, describing tasks as producing no group performance differences should alleviate the possibility that the participant confirms a negative stereotype, decreasing feelings of threat. In the current study, we include two no group differences conditions – one describing the task as producing no group differences, and another including a same-race expert describing the task as producing no group differences [65,66]. Thus, in the threat-reducing no group differences condition (condition *r2*), the task is described as showing no group differences in performance:

> "… Before starting the test, it is important to acknowledge that you may have heard that there are racial differences in test performance on certain types of tests. This is not the case for the test you will be taking today. The test you will be taking today shows no racial or group differences and such tests have been found to be culture fair and unbiased toward particular social groups. As we look towards understanding this test in today's study, it is important to note that numerous other studies have found that Black/African American students and White students always perform equally on such tests. ...".

In the threat-reducing no group differences-Black expert condition (condition *r3*), a professor with a name consistently perceived as Black (first name: DeAndre, Jamal, Jalen, Ebony, Jamila, or Amani; last name: Jackson, Johnson, Harris, Jones, Robinson, or Williams) at a university with a recognizably large number of Black college students (Howard University, University of Illinois at Chicago, University of Houston, University of Maryland, Florida A&M University, or Texas Southern University), who is quoted as describing the task as producing no group differences (as detailed above). The first names, last names, and institutions were all chosen on the basis of a pilot test with 101 Black participants recruited through MTurk. All items had a mean rating of at least 5 on a 1-7 scale of perceived Blackness; for more details see our Supplemental Method.

Confirmatory hypotheses. Each of our nine operationalizations (*i1_r1* through *i3_r3*) can be used to create a question about the effect of a particular operationalization, with a null hypothesis that the threat effect is not positive and an alternative hypothesis that it is positive. Our project nine questions, one per operationalization, which each correspond to a particular null and alternative hypothesis. We list these questions, the null and alternative hypotheses, our sampling and analytic plans, and our planned interpretations given different study outcomes in Table 3.

Conditions						**Interpretations**	
Threat-increasing	**Threat-reducing**	**Question**	**Hypothesis**	**Sampling plan**	**Analysis plan**	**log10(BF)**	**Verbal conclusion**
Diagnostic (*i1*)	Non-diagnostic (*r1*)	Does the **threat-increasing condition** (*i1*, *i2*, or *i3*) produce lower average scores on Raven's Progressive Matrices than the **threat-reducing condition** (*r1*, *r2*, or *r3*)? Labels: *i1_r1* through *i3_r3*	H_A: The threat-increasing condition will produce lower scores on Raven's Progressive Matrices than the threat-reducing condition H_0: The threat-increasing condition will not produce lower scores on Raven's Progressive Matrices than the threat-reducing condition	We will recruit at least 2,000 Black participants. Participants will be assigned in accordance with our adaptive algorithm, which prioritizes assignment to conditions that show evidence that people in the threat-increasing condition perform worse than people in the threat-reducing condition	The data will be analyzed concurrently with data collection. The analysis uses a Bayesian *t*-test, which we will use to compute a JZS Bayes factor measuring the relative evidence for H_A vs H_0	> 2.0	Extreme evidence in favor of H_A
Diagnostic (*i1*)	No group differences (*r2*)					1.5 to 2.0	Very strong evidence in favor of H_A
Diagnostic (*i1*)	No group differences, Black expert (*r3*)					1.0 to 1.5	Strong evidence in favor of H_A
Race primed (*i2*)	Non-diagnostic (*r1*)					0.5 to 1.0	Moderate evidence in favor of H_A
Race primed (*i2*)	No group differences (*r2*)					-0.5 to 0.5	Inconclusive evidence
Race primed (*i2*)	No group differences, Black expert (*r3*)					-1.0 to -0.5	Moderate evidence in favor of H_0
Group differences (*i3*)	Non-diagnostic (*r1*)					-1.5 to -1.0	Strong evidence in favor of H_0
Group differences (*i3*)	No group differences (*r2*)					-2.0 to -1.5	Very strong evidence in favor of H_0
Group differences (*i3*)	No group differences, Black expert (*r3*)					< -2.0	Extreme evidence in favor of H_0

Table 2. Design table.

Condition assignment through an adaptive design

Experiments involving a large number of conditions suffer a common problem: not all conditions are equally useful for testing the focal hypothesis, but we rarely know in advance which ones will be most informative. In between-subjects designs, the conventional way of coping with this problem is to allow an equal number of participants to experience each condition, an approach that quickly grows infeasible as the number of conditions increases. Adaptive designs solve this problem by evaluating the evidence at regular intervals and using the available evidence at a given interval to estimate the condition assignments that are likely to provide the most information for the next interval [67]. The result is that adaptive designs generally make more efficient use of experimental resources than do designs that are not adaptive [68]. Adaptive designs are therefore an ideal choice for experiments involving a large number of conditions, as they can render feasible designs that would require an unwieldy number of resources in a conventional design.

As applied to our study, we have a total of six conditions. In a conventional design, we would need an unfeasibly large number of Black students to precisely detect an effect between one of the threat-increasing and one of the threat-reducing procedures. However, adaptively allocating participants to these procedures should allow this design to yield greater evidence either in favor of or against a stereotype threat effect with a smaller number of participants.

Our adaptive design proceeds across a series of participant *cohorts*. After each cohort, we calculate the evidence that a stereotype threat effect exists within each of our nine possible comparisons between the three threat-increasing and three threat-reducing conditions. Our initial cohort will consist of 180 participants. In this first cohort, we assign an equal number of participants to each condition. Each subsequent cohort consists of a single participant, who will be assigned to a condition based on the current evidence, as calculated from all preceding cohorts. Critically, we weight the assignment probabilities such that the pairs of conditions where the evidence for a threat effect is strongest are the most likely to have participants assigned to them. Participants are therefore less likely to be assigned to conditions in which the evidence suggests that there is no threat effect. The experiment proceeds until all sites recruit their committed total number of participants, and the adaptive algorithm ensures that we make maximally efficient use of our participants' time and effort to find a threat effect.

Formally, the adaptive design experiment proceeds as follows. The first step is to collect data from the initial cohort of 180 participants, with 30 participants assigned to each of the six conditions. Based on these data, we compute the JZS Bayes factor from a Bayesian t-test [69] for each pair of threat-increasing and threat-reducing conditions. For any two conditions x and y, the Bayes factor is computed from the observed two-sample t value with degrees of freedom $\nu = N_x + N_y - 2$ and effective sample size $N = N_x N_y/(N_x + N_y)$, as

$$BF = \frac{\int_0^\infty \quad 1+Ng)^{-\frac{1}{2}}\left(1+\frac{t^2}{(1+Ng)\nu}\right)^{-\frac{\nu+1}{2}}(2\pi)^{1/2} g^{-3/2} e^{-1/(2g)}\, dg}{(1+\frac{t^2}{\nu})^{-(\nu-1)/2}} \tag{1}$$

A Bayes Factor is a ratio of the evidence against the null hypothesis relative to the evidence in favor of it. In our design, the null hypothesis is that, given two conditions, the mean difference between those conditions is zero, and the alternative hypothesis is that this difference is non-zero. A Bayes Factor greater than one therefore suggests that evidence favors the hypothesis that the condition difference is non-zero, whereas a Bayes Factor below one suggests that the evidence favors the hypothesis that the difference is 0. Due to the exponential increases in *BF* when evidence favors the alternative hypothesis, it is often convenient to report *BF* on a logarithmic scale, in which case values greater than zero indicate that the alternative is more likely, while values less than zero indicate that null is more likely. On a $\log_{10}$ scale, *BF* values greater indicate extreme evidence in favor of the alternative, values between -.5 and .5 indicate inconclusive evidence, and values less than -2 indicate extreme evidence in favor of the null. See https://osf.io/2zq7f/ for a table of all our intended evidence cutoffs, adapted from Lee and Wagenmakers [70].

The adaptive algorithm uses the Bayes factors from the initial cohort of participants to compute a probability distribution over pairs of conditions. This probability distribution will be used to determine the condition assignment for the next participant. For a given threat-increasing condition *x* and a given threat reducing condition *y*, we write *BF(x,y)* to denote the Bayes factor for the pair (*x,y*) and compute the following *pairwise assignment probability*:

$$p(x,y) = \frac{BF(x,y)}{\sum_{i=1}^{3} \sum_{j=1}^{3} BF(i,j)} \,. \qquad (2)$$

The assignment probability for a given pair of conditions is the ratio of the Bayes factor for that pair to the sum of the Bayes factors across all pairs. Participants are therefore most likely to be assigned to pairs of conditions with the highest likelihood of containing a non-zero effect.

In each subsequent cohort, the participant is assigned uniformly at random to one of the two conditions from a pair drawn from the assignment distribution computed from Equation (2). Once data has been collected from a particular cohort, they are combined with all of the previously collected data and used to compute updated Bayes factors for each pair of conditions using Equation (1). These updated Bayes factors are then used to update the assignment probabilities for the next cohort using Equation (2), and the cycle repeats. The process continues until all sites have exhausted their committed total number of participants.

Simulation evidence of the efficiency of the adaptive design. We tested the proposition that an adaptive design would yield more evidence with fewer participants in a series of three simulation studies. All simulations were run using MATLAB R2019a [71]. Our first two simulation studies assessed the relative efficiency of adaptive versus non-adaptive designs when only *one* of our threat-increasing conditions affects performance. In the first study, this one condition produced a moderate effect on performance (i.e., d = .4 when its effect is compared to the other conditions using the standardized mean difference); in our second, the condition produced a small effect (d = .2).

There are nine possible comparisons between threat-increasing and threat-reducing conditions (3 * 3 = 9). Thus, in a situation where one threat-increasing condition produces a small effect, three of the nine possible comparisons between threat-increasing and threat-reducing conditions are truly non-zero (i.e., all of the comparisons between the performance-affecting threat-increasing condition and the three threat-reducing conditions). Given this situation as ground truth, each of the two simulation studies consisted of 1,000 adaptive experiments and 1,000 fixed experiments. Both types of experiments started with an initial cohort of 180 simulated participants, split evenly across the six experimental conditions. The difference is in how simulated participants were assigned in subsequent cohorts. In the fixed experiments, simulated participants were divided equally among conditions, whereas in the adaptive experiments, simulated participants were assigned using the adaptive algorithm described above. In either case, simulated data were generated from normal distributions with equal variances, with an effect of the corresponding size in one condition. For each experiment, we computed the Bayes Factor for each comparison after each cohort of participants and recorded the maximum Bayes Factor across the three comparisons. We also recorded the proportion of the total N that had been allocated to the condition with a true effect on performance.

Figure 1 shows the results of the two simulation studies. When one condition has a medium effect (top left panel), both designs accumulate evidence against the alternative hypothesis, but the adaptive design does so especially fast. When one condition has a small effect (top right panel), the adaptive design usually reaches the threshold for strong evidence ($\log_{10}(BF) = 1.0$) after about 1,440 participants, whereas the fixed design usually fails to reach that cutoff even after 2,004 participants have been recruited. The bottom panels reveal how the adaptive design is able to achieve this efficiency gain: it preferentially allocates participants to the condition with a true effect on performance.

Although the adaptive design makes decisive results more likely than a fixed design, it does not guarantee them. The grey bands in Figure 1 represent the 25% and 75% percentiles across the 1,000 simulations. There is wide variation in the obtained evidence ratios across simulations. Reassuringly, even when the single non-null condition has a small effect (d = .2 in comparison with the other conditions), at least 75% of the simulated experiments yielded strong evidence for the alternative ($\log_{10}(BF) > 1.0$) after about 1,800 participants had been recruited.

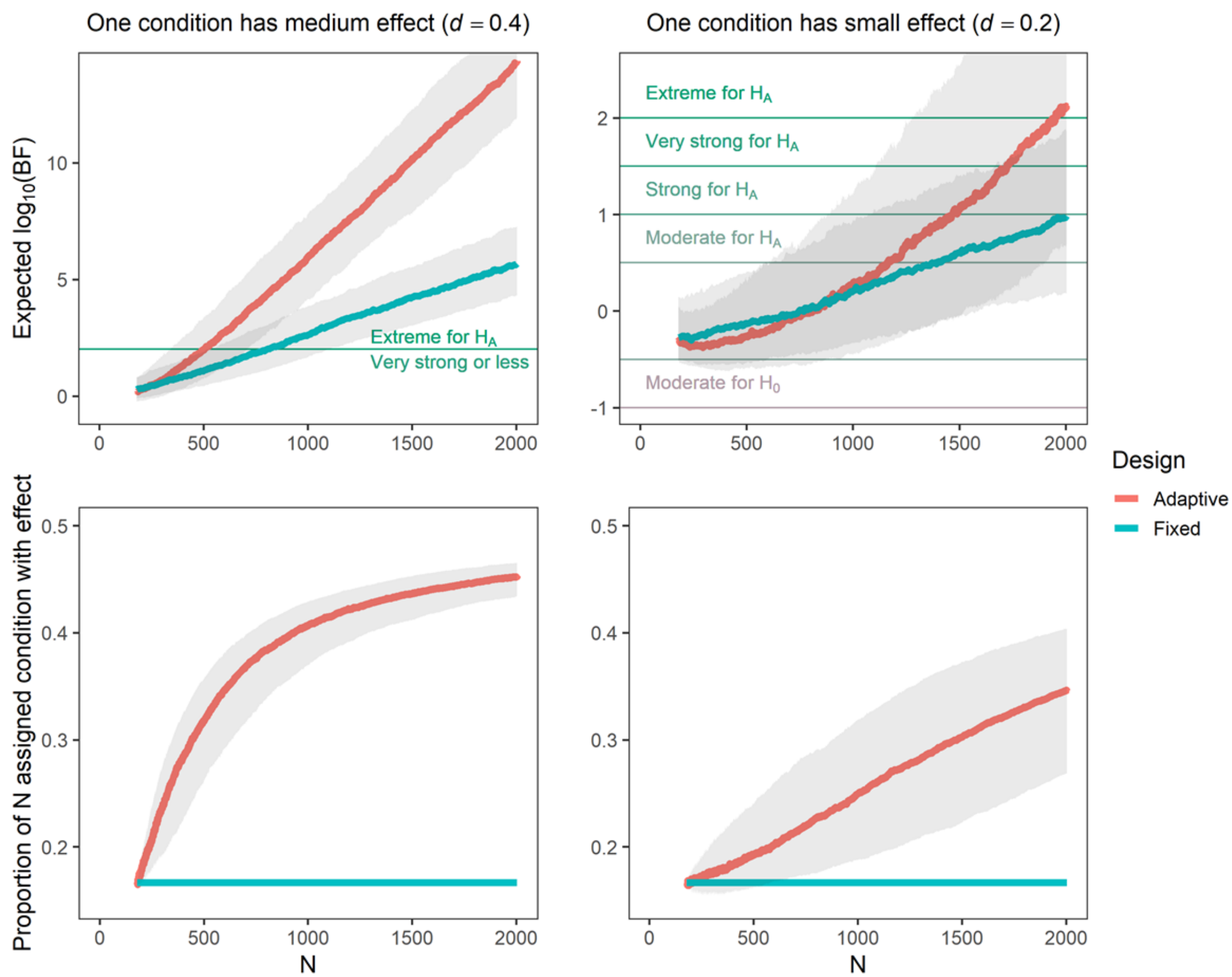


Figure 1. Results from two simulation studies with 2,000 runs each (1,000 for the adaptive version, 1,000 for the fixed). The top two panels use different scales in the y-axis for clarity. In one study (left two panels) one of our six conditions produces a small effect (d = .2 in comparison with the other conditions); in the other (right two panels) it produces a medium effect (d = .4). At each step of a given simulation run, we recorded, of the three truly non-null comparisons, the Bayes Factor of the comparisons that yielded the maximum evidence against the null, as well as the proportion of the total study N assigned to the condition that does have an effect. *N* refers to the number of participants recruited at a particular point in the design. Lines represent the medians across the 1,000 simulations of the quantity in question; envelopes represent the 25% and 75% quantiles. When either a small or medium effect is present, the adaptive design accumulates evidence against the null more efficiently than does a fixed design. It does so by preferentially allocating participants to the condition that provides the best evidence of an effect.

Finally, we investigated how the adaptive and fixed designs performed in the presence of no stereotype threat effects – in other words, in a situation where the mean difference in all comparisons between threat-increasing and threat-reducing conditions was equal to 0. This is a situation where one might expect the adaptive design to be at a disadvantage, since, on average, assigning people to the six conditions with equal probability is already “optimal”.

Because the null hypothesis is true for each of the nine comparisons, the experiment is successful if it yields negative values of the $\log_{10}$(BF) for every hypothesis test, indicating strong evidence in favor of the null hypothesis. In contrast, any positive Bayes Factor indicates false evidence against one of the null hypotheses. To assess each design’s performance in each simulated experiment, we computed, across all comparisons, the minimum Bayes Factor (i.e., the evidence ratio that is most in favor of the null and therefore “correct”), the maximum Bayes factor (the evidence ratio most in favor of the alternative and therefore “incorrect”), and the average. As shown in our Supplemental Method, both designs yielded moderate to strong evidence in favor of the null across all possible comparisons of procedures. Moreover, the adaptive design performed no worse than the fixed design, and, one particular dimension, even had a slight advantage –they were somewhat less likely to produce a comparison that yielded false evidence in favor of the alternative hypothesis. This may occur because, if a particular comparison does provide (incorrect) evidence in favor of the alternative, the adaptive design preferentially allocates people to that comparison until the evidence ratio begins to favor the null. In essence, the adaptive design performs a small “replication study” for a comparison favoring the alternative, which provides some protection from drawing false positive conclusions (see our Supplemental Method for additional simulation evidence on this point).

Across our simulations, we note that the Bayesian test we used is somewhat conservative – that is, it is calibrated such that the null hypothesis is favored unless a relatively large effect is observed. The conservatism likely reflects the fact that we used the JZS Bayes factor with a scale parameter $r = 1$, which anticipates effects between -1 and 1. Additional simulation results (described in our Supplemental Method) show that adjusting this prior to $r = 0.5$, or $r = 0.2$, does not have much effect on the efficiency of the adaptive design relative to the fixed design. However, it can affect the Bayes Factor’s absolute magnitude. To address this conservatism, we will survey the participating sites to estimate an expected effect size. We will use this expected effect size to adjust the scaling parameter prior to our final analysis.

Site balancing, data flow, and by-site variance. The adaptive algorithm does not explicitly account for the possibility of site-specific differences in participant performance. The appropriate statistical approach to account for this kind of by-site variance would be to use a random effects model with site-specific random parameters. However, adding site-specific random parameters to the adaptive algorithm would create a computational bottleneck in the calculation of the Bayes factors for updating condition assignments. Moreover, since condition assignments from the algorithm are based on the magnitudes of the fixed effects (i.e., the average differences between conditions), the estimates of the random effects would have minimal effect on condition assignments. In other words, even if the model in the adaptive algorithm were misspecified due to the absence of random effects parameters, the algorithm can still achieve its goal of increasing power in conditions where the average effects are largest. Therefore, the adaptive algorithm will not include random effects by site, but we will examine the degree to which adding random effects for site affects our results after our data are collected as part of our robustness checks.

The possibility of by-site variance also necessitates additional controls on data flow to ensure that sites are balanced throughout the experiment. An especially dangerous scenario occurs if a given site dominates sampling at a particular point in time during the experiment, a phenomenon we refer to as clumping. For example, suppose there is no true effect between conditions 1 and 2 at site A, but a moderate true effect at every other site. Suppose also that there is a moderate effect between conditions 3 and 4 at site A, but no true effect at any other site. Thus, on average across sites, there is a larger difference between conditions 1 and 2 than between 3 and 4. If all 180 of the participants in the initial block came from Site A, the initial data would suggest an effect between conditions 3 and 4 but not between 1 and 2. This would bias the sampling in subsequent blocks toward conditions 3 and 4, when it would be more fruitful to test conditions 1 and 2. On the other hand, if Site A was not represented at all until the final blocks of the experiment, then most of the participants from Site A would be assigned to conditions 1 and 2, since that is where the largest effect would appear to be at the point where participants from Site A enter the experiment. But since the effect at Site A is between conditions 3 and 4, not between 1 and 2, and since the effect between conditions 3 and 4 does not exist at any of the other sites, the algorithm may never learn about the presence of the true effect between conditions 3 and 4.

The above example is extreme, but it illustrates the potential risks of clumping for statistical inference and algorithmic efficiency. We will therefore take the following three concrete measures to mitigate these risks. First, we will require that at least 15 sites are represented in the initial block of 180 participants, with each site contributing at least five participants. Second, we will not allow any single site to contribute more than 10 participants in a given week. Third, we will set minimum targets for the number of participants each site should aim to contribute each week. For instance, if a site plans to contribute 20 participants over the course of 5 weeks, we will ask them to contribute at least four participants each week. These measures should help guard against the possible risks of clumping described above.

Analysis plan

Manipulation checks. Experiencing a task as difficult is a theoretically necessary condition for producing stereotype threat [10]. We have selected a performance task, Raven's Advanced Progressive Matrices, that should be experienced as difficult by most college students [51]. Nevertheless, we will check that our participants did indeed experience the task as difficult by examining the percentage of participants who reported that the level of difficulty was above the midpoint on the perceived difficulty measure and by testing whether the average rated difficulty across all students is significantly above the scale midpoint. We will also test whether our manipulations did indeed evoke feelings of concern about confirming the negative Black-unintelligent stereotype by testing whether reported task-evoked concern in the threat-increasing conditions is significantly greater than reported task-evoked concern in the threat-reducing conditions. We will use the same Bayesian model for these manipulation checks that we use for the main analysis.

Confirmatory analyses. Because we are using an adaptive design, our main analysis will proceed with and guide the data collection process. After gathering an initial cohort of 180 participants, we will follow the adaptive design outlined in the previous section: we will calculate $\log_{10}(BF)$ values approximating the posterior odds that a stereotype threat effect exists within each of the nine possible comparisons between our three threat-increasing and three threat-reducing conditions. For each subsequent cohort of six participants, these $\log_{10}(BF)$ values are used to determine the probability that any given participant is assigned to each of the six conditions. If we obtain extreme evidence that a particular comparison does or does not produce a stereotype threat effect ($\log_{10}(BF) > 2.0$ or $\log_{10}(BF) < -2.0$), we will cease sampling that comparison to ensure that this comparison does not dominate future sampling and thereby prevent us from gathering evidence about other comparisons. Data collection proceeds until we either obtain strong evidence about the presence or absence of stereotype threat across all comparisons or the labs who have committed to collecting data for this project all reach their committed recruitment totals. After all data has been collected, we will use the final $\log_{10}$(BF) values to assess the likelihood that a stereotype threat effect exists within each of the nine comparisons. Thus, the set of nine final $\log_{10}$(BF) represents our focal tests of our nine questions about the presence of a stereotype threat effect (questions *i1_r1* through *i3_r3*) in a given operationalization.

Robustness checks. In addition to our main analysis, we will also conduct a series of robustness checks in the form of a multiverse analysis [72], which we will use to assess the degree to which our results change across alternative strategies for analyzing our data. We identify five points of flexibility in our analysis where different choices or assumptions could have been made. These include the statistical framework, priors, two types of random effects, and rules for excluding observations. For each point of flexibility, we identify several alternatives to be considered. We will rerun the analysis under various combinations of those alternatives, as shown in Table 4. Taken together, including only the combinations of alternatives that are theoretically compatible as well as computationally tractable, our planned robustness analyses span 160 separate analyses: 32 in a Bayesian statistical framework and 128 in a Frequentist statistical framework. We consider random effects in a Frequentist framework only due to the additional complexities that arise when formulating and estimating these models in a Bayesian framework.

Exploratory analyses. Although we expect most of our participants to be identified with intelligence and their race, we will test whether those who are less identified are more (or less) affected by stereotype threat. Similarly, we will test whether people who are chronically concerned about stereotype threat are more affected. More specifically, if we find a stereotype threat effect for one of our comparisons, we will test three interactions, one between the comparison and racial identification, one between the comparison and general domain identification, and one between chronic concern and the comparison. In addition, given sufficient data and sufficient variation in the applicable variables, we will test other potential moderators of the threat effect such as generation status.

We also plan to assess the degree to which there is substantive variation between experimenters and sites in the magnitude of the stereotype threat effect by measuring the size of the by-experimenter and by-site random slopes for threat in mixed effects models. If there is substantive variation in the size of the stereotype threat effect, we will explore possible sources of this variation by testing interactions between our threat manipulations and either our experimenter variables (if there is by-experimenter variation) or our by-site variables (if there is by-site variation).

Point of flexibility		Alternatives considered	Justification
Statistical framework	(1)	Bayesian*	It may be useful to quantify evidence in favor of the null hypothesis
	(2)	Frequentist	It may be reasonable to want an analysis that does not rely on priors
Prior	(1)	Informed scale factor*	Priors should align with the expectations of experts in the field
	(2)	Small scale prior (r=0.2)	It may be a priori reasonable to expect small effects
	(3)	Medium scale prior (r=0.5)	It may be a priori reasonable to expect medium effects
	(4)	Unit scale prior (r=1.0)	It may be a priori reasonable to expect large effects
Random site effects	(1)	None*	There may not be much site clustering in participant performance
	(2)	Random by-site intercepts	Different sites may have different average levels of participant performance
	(3)	Random by-site slopes for threat effects	Sites may differ substantively in the size of a given threat effect
	(4)	Combine (2) and (3)	---
Random experimenter effects	(1)	None*	There may not be much experimenter clustering in participant performance
	(2)	Random by-experimenter intercepts	Different experimenters may produce different average levels of participant performance
	(3)	Random by-experimenter slopes the threat effects	Experimenters may vary in the degree to which they produce threat effects
	(4)	Combine (2) and (3)	---
Observations	(1)	All*	All participants may provide useful information about the presence of a threat effect
	(2)	Exclude people who do not have good memory of the study's details	These people may not have been properly exposed to the manipulation
	(3)	Exclude people who are unaware of the Black-intelligence stereotype	These people may not have the proper cultural awareness for stereotype threat to affect their behavior
	(4)	Exclude suspicious participants	These people may not have been affected by the manipulation because they didn't believe it
	(5)	Combine (2) and (3)	---
	(6)	Combine (2) and (4)	---
	(7)	Combine (3) and (4)	---
	(8)	Combine (2), (3), and (4)	---

Table 3. Potential points of flexibility in our analysis plan. Robustness with respect to priors will be explored within a Bayesian statistical framework. Robustness with respect to random effects will be explored within a frequentist statistical framework. Robustness with respect to observations will be explored within both statistical frameworks. Together, these points of flexibility yield 160 possible statistical models. We will assess the degree to which our results change across these models. Alternatives marked by * are those used in the main analysis.

References:

1. Sunstein, C. Black on Brown: 50 Years of Brown v. Board of Education. *Va. Law Rev.* **90**, 1649–1655 (2004).
2. Lewis, N. A. & Yates, J. F. Preparing Disadvantaged Students for Success in College: Lessons Learned From the Preparation Initiative. *Perspect. Psychol. Sci.* **14**, 54–59 (2019).
3. Oyserman, D. & Lewis, N. A. Seeing the Destination AND the Path: Using Identity-Based Motivation to Understand and Reduce Racial Disparities in Academic Achievement: Seeing the Destination and the Path. *Soc. Issues Policy Rev.* **11**, 159–194 (2017).
4. Warren, E. *Brown v. Board of Education of Topeka*. *United States Reports* vol. 347 (1954).
5. Steele, C. M. *Whistling Vivaldi: and other clues to how stereotypes affect us*. (W.W. Norton & Company, 2010).
6. Steele, C. M. A threat in the air: How stereotypes shape intellectual identity and performance. *Am. Psychol.* **52**, 613–629 (1997).
7. Steele, C. M., Spencer, S. J. & Aronson, J. Contending with group image: The psychology of stereotype and social identity threat. in *Advances in Experimental Social Psychology* vol. 34 379–440 (Elsevier, 2002).
8. Schmader, T., Johns, M. & Forbes, C. An integrated process model of stereotype threat effects on performance. *Psychol. Rev.* **115**, 336–356 (2008).
9. Kennedy, A. *Fisher v. University of Texas*. *United States Reports* vol. 570 (2013).
10. Steele, C. M. & Aronson, J. Stereotype threat and the intellectual test performance of African Americans. *J. Pers. Soc. Psychol.* **69**, 797–811 (1995).
11. Aronson, J. Stereotype Threat. in *Improving Academic Achievement* 279–301 (Elsevier, 2002). doi:10.1016/B978-012064455-1/50017-8.

12. Davis, C., Aronson, J. & Salinas, M. Shades of Threat: Racial Identity as a Moderator of Stereotype Threat. *J. Black Psychol.* **32**, 399–417 (2006).
13. Oyserman, D., Harrison, K. & Bybee, D. Can racial identity be promotive of academic efficacy? *Int. J. Behav. Dev.* **25**, 379–385 (2001).
14. Marx, D. M. & Goff, P. A. Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *Br. J. Soc. Psychol.* **44**, 645–657 (2005).
15. Woodcock, A., Hernandez, P. R., Estrada, M. & Schultz, P. W. The consequences of chronic stereotype threat: Domain disidentification and abandonment. *J. Pers. Soc. Psychol.* **103**, 635–646 (2012).
16. Spencer, S. J., Steele, C. M. & Quinn, D. M. Stereotype Threat and Women's Math Performance. *J. Exp. Soc. Psychol.* **35**, 4–28 (1999).
17. Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A. & Mitchell, J. P. Deflecting negative self-relevant stereotype activation: The effects of individuation. *J. Exp. Soc. Psychol.* **40**, 401–408 (2004).
18. McIntyre, R. B., Paulson, R. M. & Lord, C. G. Alleviating women's mathematics stereotype threat through salience of group achievements. *J. Exp. Soc. Psychol.* **39**, 83–90 (2003).
19. Abrams, D., Eller, A. & Bryant, J. An age apart: The effects of intergenerational contact and stereotype threat on performance and intergroup bias. *Psychol. Aging* **21**, 691–702 (2006).
20. Hess, T. M. & Hinson, J. T. Age-related variation in the influences of aging stereotypes on memory in adulthood. *Psychol. Aging* **21**, 621–625 (2006).
21. Croizet, J.-C. & Claire, T. Extending the Concept of Stereotype Threat to Social Class: The Intellectual Underperformance of Students from Low Socioeconomic Backgrounds. *Pers. Soc. Psychol. Bull.* **24**, 588–594 (1998).

22. McFarland, J. *et al. The Condition of Education 2019*. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019144 (2019).
23. Cottom, T. M. *Lower Ed: The troubling rise of for-profit colleges in the new economy*. (2018).
24. Vanneman, A., Hamilton, L., Anderson, J. B. & Rahman, T. *Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress*. (2009).
25. Carter, P. L. & Welner, K. G. *Closing the Opportunity Gap: What America Must Do to Give Every Child an Even Chance*. (Oxford University Press, 2013).
26. Stoet, G. & Geary, D. C. Can Stereotype Threat Explain the Gender Gap in Mathematics Performance and Achievement? *Rev. Gen. Psychol.* **16**, 93–102 (2012).
27. Murphy, M. C., Steele, C. M. & Gross, J. J. Signaling Threat: How Situational Cues Affect Women in Math, Science, and Engineering Settings. *Psychol. Sci.* **18**, 879–885 (2007).
28. Murphy, M. C. & Taylor, V. J. The role of situational cues in signaling and maintaining stereotype threat. in *Stereotype threat: Theory, process, and application* (eds. Inzlicht, M. & Schmader, T.) 17–33 (Oxford University Press, 2012).
29. Arbuthnot, K. The Effects of Stereotype Threat on Standardized Mathematics Test Performance and Cognitive Processing. *Harv. Educ. Rev.* **79**, 448–473 (2009).
30. Brown, R. P. & Pinel, E. C. Stigma on my mind: Individual differences in the experience of stereotype threat. *J. Exp. Soc. Psychol.* **39**, 626–633 (2003).
31. Lewis, N. A. & Michalak, N. M. *Has Stereotype Threat Dissipated Over Time? A Cross-Temporal Meta-Analysis*. https://osf.io/w4ta2 (2019) doi:10.31234/osf.io/w4ta2.

32. Lewis, N. A. & Sekaquaptewa, D. Beyond test performance: a broader view of stereotype threat. *Curr. Opin. Psychol.* **11**, 40–43 (2016).

33. Nguyen, H. D. & Ryan, A. M. Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *J. Appl. Psychol.* **93**, 1314–1334 (2008).

34. Nadler, J. T. & Clark, M. H. Stereotype Threat: A Meta-Analysis Comparing African Americans to Hispanic Americans. *J. Appl. Soc. Psychol.* **41**, 872–890 (2011).

35. Shewach, O. R., Sackett, P. R. & Quint, S. Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *J. Appl. Psychol.* **104**, 1514–1534 (2019).

36. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).

37. Flore, P. C. & Wicherts, J. M. Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *J. Sch. Psychol.* **53**, 25–44 (2015).

38. Zigerell, L. J. Potential publication bias in the stereotype threat literature: Comment on Nguyen and Ryan (2008). *J. Appl. Psychol.* **102**, 1159–1168 (2017).

39. Finnigan, K. M. & Corker, K. S. Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance? *J. Res. Personal.* **63**, 36–43 (2016).

40. Finkel, E. J., Eastwick, P. W. & Reis, H. T. Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *J. Pers. Soc. Psychol.* **113**, 244–253 (2017).

41. Lykken, D. T. Statistical significance in psychological research. *Psychol. Bull.* **70**, 151–159 (1968).

42. LeBel, E. P., Berger, D., Campbell, L. & Loving, T. J. Falsifiability is not optional. *J. Pers. Soc. Psychol.* **113**, 254–261 (2017).

43. Devezer, B., Nardin, L. G., Baumgaertner, B. & Buzbas, E. O. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE* **14**, e0216125 (2019).

44. Moshontz, H. *et al.* The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Adv. Methods Pract. Psychol. Sci.* **1**, 501–515 (2018).

45. Pinel, E. C. Stigma consciousness: The psychological legacy of social stereotypes. *J. Pers. Soc. Psychol.* **76**, 114–128 (1999).

46. Brewer, M. B. The Social Self: On Being the Same and Different at the Same Time. *Pers. Soc. Psychol. Bull.* **17**, 475–482 (1991).

47. Branscombe, N. R., Schmitt, M. T. & Harvey, R. D. Perceiving pervasive discrimination among African Americans: Implications for group identification and well-being. *J. Pers. Soc. Psychol.* **77**, 135–149 (1999).

48. Flake, J. K., Pek, J. & Hehman, E. Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Soc. Psychol. Personal. Sci.* **8**, 370–378 (2017).

50. Raven, J. C., Court, J. H. & Raven, J. Manual for Raven’s Progressive Matrices and vocabulary scales. (1988).

51. Bors, D. A. & Stokes, T. L. Raven’s Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. *Educ. Psychol. Meas.* **58**, 382–398 (1998).

52. Mayer, D. M. & Hanges, P. J. Understanding the Stereotype Threat Effect With ‘Culture-Free’ Tests: An Examination of its Mediators and Measurement. *Hum. Perform.* **16**, 207–230 (2003).
53. Brown, R. P. & Day, E. A. The difference isn’t black and white: Stereotype threat and the race gap on raven’s advanced progressive matrices. *J. Appl. Psychol.* **91**, 979–985 (2006).
54. McKay, P. F., Doverspike, D., Bowen-Hilton, D. & McKay, Q. D. The Effects of Demographic Variables and Stereotype Threat on Black/White Differences in Cognitive Ability Test Performance. *J. Bus. Psychol.* **18**, 1–14 (2003).
55. Conway, A. R. A., Kane, M. J. & Engle, R. W. Working memory capacity and its relation to general intelligence. *Trends Cogn. Sci.* **7**, 547–552 (2003).
56. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* **6**, 316–322 (2003).
57. Lewis, N. A., Sekaquaptewa, D. & Meadows, L. A. Modeling gender counter-stereotypic group behavior: a brief video intervention reduces participation gender gaps on STEM teams. *Soc. Psychol. Educ.* **22**, 557–577 (2019).
58. Schmader, T. Gender Identification Moderates Stereotype Threat Effects on Women’s Math Performance. *J. Exp. Soc. Psychol.* **38**, 194–201 (2002).
59. Sellers, R. M., Smith, M. A., Shelton, J. N., Rowley, S. A. J. & Chavous, T. M. Multidimensional Model of Racial Identity: A Reconceptualization of African American Racial Identity. *Personal. Soc. Psychol. Rev.* **2**, 18–39 (1998).
60. Marx, D. M. & Roman, J. S. Female Role Models: Protecting Women’s Math Test Performance. *Pers. Soc. Psychol. Bull.* **28**, 1183–1193 (2002).

61. Sekaquaptewa, D. & Thompson, M. The Differential Effects of Solo Status on Members of High- and Low-Status Groups. *Pers. Soc. Psychol. Bull.* **28**, 694–707 (2002).
62. Ramsey, L. R., Betz, D. E. & Sekaquaptewa, D. The effects of an academic environment intervention on science identification among women in STEM. *Soc. Psychol. Educ.* **16**, 377–397 (2013).
63. Adler, N. E., Epel, E. S., Castellazzo, G. & Ickovics, J. R. Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy white women. *Health Psychol.* **19**, 586–592.
64. Arslan, R. C., Walther, M. P. & Tata, C. S. formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behav. Res. Methods* (2019) doi:10.3758/s13428-019-01236-y.
65. Blascovich, J., Spencer, S. J., Quinn, D. & Steele, C. African Americans and High Blood Pressure: The Role of Stereotype Threat. *Psychol. Sci.* **12**, 225–229 (2001).
66. Wout, D. A., Shih, M. J., Jackson, J. S. & Sellers, R. M. Targets as perceivers: How people determine when they will be negatively stereotyped. *J. Pers. Soc. Psychol.* **96**, 349–362 (2009).
67. Cavagnaro, D. R., Myung, J. I., Pitt, M. A. & Kujala, J. V. Adaptive Design Optimization: A Mutual Information-Based Approach to Model Discrimination in Cognitive Science. *Neural Comput.* **22**, 887–905 (2010).
68. Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M. & Perugini, M. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychol. Methods* **22**, 322–339 (2017).

69. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).

70. Lee, M. D. & Wagenmakers, E.-J. Bayesian data analysis for cognitive science: A practical course. (2013).

71. *MATLAB and Statistics Toolbox Release R2019a*. (The Mathworks, Inc, 2019).

72. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).

Acknowledgements:
C. N. S. & A. Y. M. were supported by a USI College of Liberal Arts Faculty Development Award. N. L., Jr. was supported by a Faculty Fellowship from the Cornell Center for the Social Sciences. B. S. and S. G. were supported by the Charles Lafitte Foundation. A.R.L. was supported by NSF 1631325 and NIH R01 DA041353. P. S. F. and M. L. were supported by an SPSP Inside the Grant Panel Award. M.C.M's effort is supported by NSF grants DRL-1450755 and HRD-1661004.

The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript

Author Contributions:
(as determined by the collaboration agreement)

Tier 1: *Contributions to conceptualization, methodology, formal analysis, software, or resources, and writing - original draft, review and editing.*

Patrick S. Forscher*, Valerie Jones Taylor*, Daniel R. Cavagnaro, Neil A. Lewis, Jr., Erin Buchanan, Hannah Moshontz

Authors contributed equally. Order was determined with the following R code:

```
set.seed(1941)
authors <- c("Valerie", "Patrick")
sentence <- paste("The first-listed author is", sample(authors, size=1))
print(sentence)
```

Tier 2: Major contributions to validation, project administration, and/or writing - original draft, review and editing. Ordered alphabetically unless otherwise determined by discussion.

Aimee Y. Mark

Tier 3: Investigation and writing - review and editing. Ordering alphabetical.

Sara C. Appleby, Carlota Batres, Brooke Bennett-Day, William J. Chopik, Rodica Ioana Damian, Claire E. Ellis, Caitlin Faas, Sarah E Gaither, Dorainne Green, Braeden F. Hall, Bianca Marie Hinojosa, Jennifer L. Howell, David C. Johnson, Franki Y. H. Kung, Angela R. Laird, Carmel A Levitan, Manyu Li, Keith B. Maddox, Mary C. Murphy, Erica D. Musser, Brianna Pankey, Laura Ruth Murry Parker, Sylvia P Perry, Jessica D. Remedios, Kathleen Schmidt, Surizaday Serrano, Crystal N. Steltenpohl, Daniel Storage, Brenda C. Straka, Heather L. Urry, Samuel C Wasmuth, Erin C. Westgate, John Paul Wilson, Shelby Wynn, David M. Zimmerman

Tier 4: Supervision and writing - review and editing. Ordered alphabetically with Chartier last.

Kim Peters, Christopher R. Chartier

Competing Interests:
The authors declare no competing interests.

Supplementary Information:

Supplemental Methods

Here we give additional detail on the following methodological issues: (1) our selection of names and institutions for the “no group differences – Black expert” condition; (2) the performance of the adaptive design (relative to a fixed design) in the presence of null effects; (3) the sensitivity of the adaptive design to priors; (4) evidence of the feasibility of our project. The data and materials for our names and institutions pilot are at https://osf.io/726qn/; the code required to run the simulations described in this supplement is at https://osf.io/vxd5y/; the proofs of concepts described in our feasibility section are at https://osf.io/tyasd/.

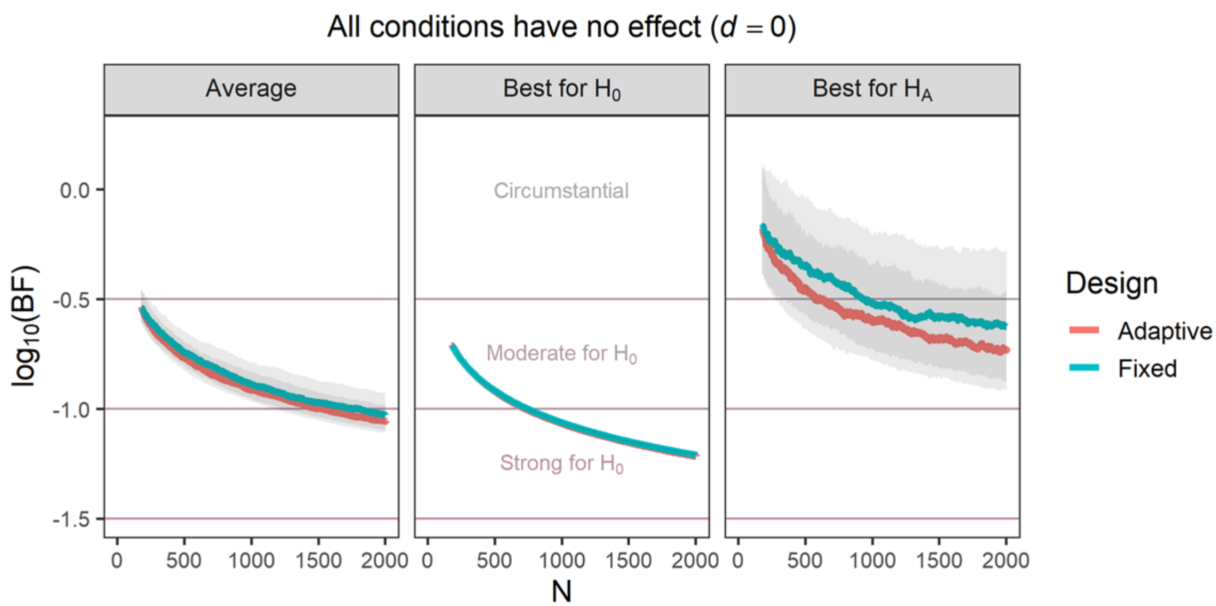
Piloting names and institutions. We conducted a pilot to test whether the names and institutions we chose for our “no group differences – Black expert” condition did indeed imply that the expert who delivers the no group differences prompt is Black. We recruited 101 Black participants (three additional participants made it to the consent form but gave no responses) using TurkPrime and asked them to rate, using 7-point Likert scales (“Extremely unlikely” to “Extremely likely”), the likelihood that each of 12 last names is Black/African American and the likelihood that they are White. We also asked the participants to rate the likelihood that 10 female first names come from a Black woman and a White woman, and conducted a similar process to assess the perceived likelihood that 10 male first names come from a Black man and a White man. Finally, we asked the participants to rate the likelihood that each of 12 institutions are associated with Blacks/African Americans.

Our results are displayed in Supplemental Table 1. On the basis of these results, our selected male names are DeAndre, Jamal, and Jalen, and our selected female names are Ebony, Jamila, and Amani. Our selected last names are Jackson, Johnson, Harris, Jones, Robinson, and Williams. Finally, our selected universities are Howard University, University Illinois at Chicago, University of Houston, University of Maryland, Florida A&M University, and Texas Southern University.

		Black perception		White perception		Difference	
		M	*SD*	*M*	*SD*	*M*	*SD*
Male first names	**DeAndre**	6.46	0.95	1.88	1.54	4.57	2.21
	Jamal	6.46	0.85	2.01	1.51	4.45	2.09
	Jalen	5.98	1.27	2.37	1.71	3.61	2.43
	Terrance	6.10	1.14	3.20	1.99	2.90	2.49
	Darryl	5.94	1.43	3.16	2.08	2.78	2.95
	Reginald	4.54	2.10	3.98	2.20	0.56	3.68
	James	5.32	1.36	5.40	1.72	-0.08	2.02
	Michael	5.45	1.45	5.79	1.46	-0.35	1.84
	Chris	5.54	1.32	5.93	1.35	-0.39	1.59
	Kevin	4.81	1.82	5.95	1.34	-1.14	2.26
Female first names	**Ebony**	6.11	1.41	1.91	1.52	4.20	2.45
	Jamila	6.06	1.09	2.25	1.60	3.81	2.18
	Amani	5.85	1.25	2.49	1.76	3.37	2.48
	Desiree	6.06	1.16	3.00	1.85	3.06	2.49
	Jada	6.03	1.07	3.01	1.85	3.02	2.28
	Renee	5.56	1.40	3.84	2.01	1.72	2.81
	Jasmine	5.76	1.34	4.52	1.90	1.24	2.43
	Laila	4.64	1.76	4.15	1.99	0.50	3.08
	Crystal	4.75	1.66	4.81	1.90	-0.06	2.87
	Amanda	3.34	1.76	6.23	1.42	-2.89	2.24
Last names	**Brown**	6.03	1.14	3.55	1.87	2.48	2.36
	Jackson	6.08	1.19	3.92	1.98	2.16	2.50
	Johnson	5.95	1.31	4.50	1.97	1.46	2.61
	Harris	5.60	1.43	4.21	1.81	1.40	2.57
	Jones	5.78	1.49	4.41	1.85	1.38	2.48
	Robinson	5.63	1.55	4.31	1.89	1.33	2.61
	Williams	5.99	1.27	4.69	1.82	1.30	2.17
	Davis	5.40	1.56	4.55	1.83	0.84	2.69
	Washington	5.32	1.73	4.51	1.97	0.80	2.95
	Coleman	4.80	1.70	4.47	1.93	0.34	2.85
	Thomas	4.67	1.73	4.88	1.70	-0.21	2.56
	Banks			4.14	1.89		
	Dixon	4.35	1.84				
Universities	**Howard University**	5.74	1.59				
	University of Illinois at Chicago	5.44	1.59				
	University of Houston	5.26	1.40				
	University of Maryland	5.25	1.56				
	Florida A&M	5.17	1.59				
	Texas Southern University	5.06	1.61				
	North Carolina A&T University	5.01	1.63				
	Hampton University	4.74	1.81				
	Florida International	4.72	1.58				
	UCLA	4.30	1.77				
	Harvard University	3.75	1.72				

Supplemental Table 1. Descriptive statistics of ratings from 101 Black raters from Turkprime of different names and institutions on perceived blackness and whiteness from 101 Black raters recruited through Turkprime. The names and institutions that we selected for the "No group differences – Black expert" condition are bolded.

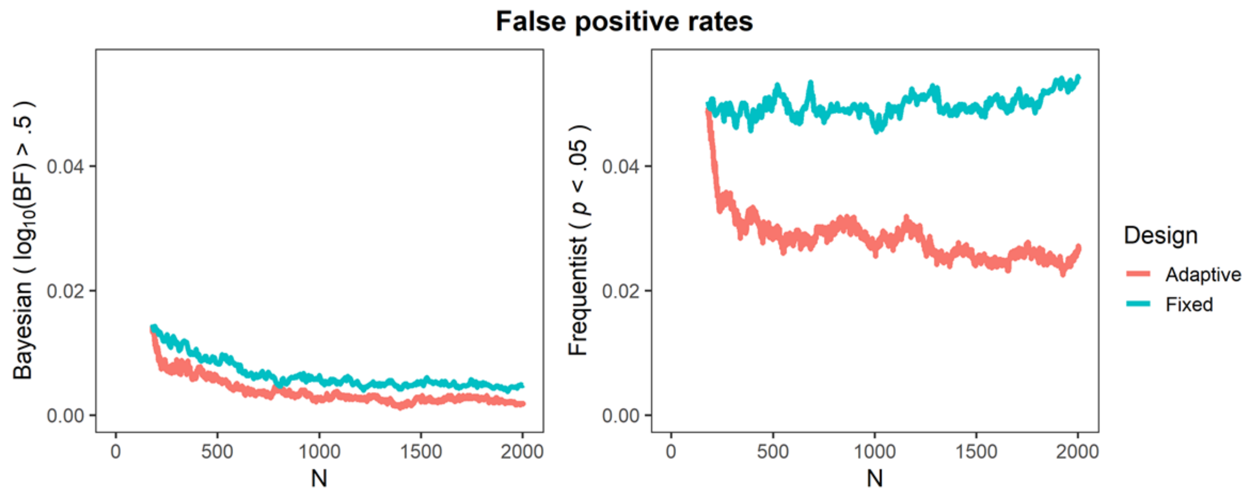
Null effects and the adaptive design. We conducted a 1000-run simulation study to assess the performance of the adaptive design when all comparisons between the threat-increasing and threat-reducing conditions yield null effects ($d = 0$). We simulated 1000 experiments using the adaptive algorithm, and 1000 experiments using a fixed design allotting equal numbers of participants to each condition. In each simulated experiment, the mean difference in all comparisons between threat-increasing and threat-reducing conditions was equal to 0 (i.e., no effect). Specifically, data in each condition were generated from a normal distribution with $\mu = 100$ and $\sigma = 10$. In both the fixed and adaptive-designed simulations, data were generated with an initial block of $N = 180$, with 30 assigned to each condition, and then subsequently in blocks of $N = 6$, up to a total of $N = 2004$ observations.



Supplemental Figure 1. Results from a 1,000-run simulation study in which all six conditions have the same group mean. N refers to the number of participants recruited at a particular point in the design. Lines represent the evidence ratio, across all six conditions, that either most favors the null (i.e., that is most correct), most favors the alternative (i.e., that is most incorrect), or the average across the six conditions. Envelopes represent the 25% and 75% quantiles. The adaptive design performs no worse than the fixed at accumulating evidence in favor of the null, and even provides some slight protection from providing (false) evidence in favor of the alternative.

As shown in Supplemental Figure 1, the adaptive design (correctly) accumulated evidence in favor of the null at a rate that was no worse than the fixed design. The design even provides a slight advantage over the fixed in that comparison that provides the most decisive evidence in favor of the alternative (and thus that draws an incorrect conclusion) tends to favor the alternative less strongly than in the fixed design. This may be because the adaptive algorithm detects the possibility of a threat effect in this comparison and thus preferentially allocates participants there. This speeds up the rate at which the algorithm correctly adjusts the evidence ratio back toward favoring the null. In a sense, the algorithm performs a small “replication study” to see whether the past evidence that favors the alternative holds up when new participants are allocated to that condition.

We investigated this latter error-preventing feature of the adaptive design further by tracking the number of times a comparison yielded $\log_{10}(\mathrm{BF}) > 0.5$, as well as the number of times the frequentist version of our test yielded $p < 0.05$. Dividing these values by the number of comparisons (i.e., dividing by nine) yields the Bayesian and frequentist false positive rates, respectively. For example, if at a given point in a simulated experiment the log10(BF) for one of the nine comparisons was greater than 0.5, while the other eight were all less than 0.5, then the Bayesian false-positive rate would be 1/9. Taking the average of the false positive rates across simulated experiments yields the overall false positive rates for the entire batch of simulations. For instance, the overall Bayesian false-positive rate at N = 180 is the average across simulated experiments of the Bayesian false-positive rates at N = 180.



Supplemental Figure 2. Results from a 1,000-run simulation study in which all six conditions have the same group mean. N refers to the number of participants recruited at a particular point in the design. At each stage in a given run of the simulation study, we selected the comparison that most favored the alternative hypothesis; lines represent either the rate of Bayesian false-positives for that comparison (moderate evidence in favor of the alternative, or $\log_{10}(BF) > .5$) or the rate of frequentist false-positives for that comparison ($p < .05$ in a test of the hypothesis that the comparison is 0).

As shown in Supplemental Figure 2, the adaptive design provides an advantage over the fixed design in protecting against false positives. Overall false positive rates in the fixed and adaptive designs are identical at N = 180, the lowest value on the x-axis, because both designs assign participants evenly across conditions in the initial block. However, both the Bayesian and frequentist false-positive rates are lower under the adaptive design than under fixed design after every subsequent block. In both cases, the rates under the adaptive design are about half that under the fixed design.

In the Bayesian case, shown in the left panel, the rates are very low under both the fixed adaptive and fixed designs. Both remain near or below .01 over the course of an experiment, so the absolute magnitude of the difference between the two designs is small. For instance, at N = 2004, the average false positive rates were .0049 and .0027 under the fixed and adaptive designs, respectively. This indicates that the Bayesian analysis is virtually immune to false positive conclusions, regardless of the statistical framework.

In the frequentist case, shown in the right panel, the false-positive rate under a fixed design hovers around the nominal rate of .05. However, the positive rate under the adaptive design starts at .05 after the initial block and then drops quickly before appearing to asymptote around .025. Speculatively, this may have occurred because when, due to random fluctuations, a particular comparison shows some signs of being non-zero, the adaptive design preferentially allocates future cohorts of participants to that comparison. The greater numbers of participants allocated to that condition lets the comparison regress to the true mean of zero faster than would happen under a fixed design.

The adaptive design's sensitivity to priors. We conducted computer simulations to assess the sensitivity of the analysis to the scale parameter of the prior for the JZS Bayes factor. We simulated experiments under three different scenarios regarding the underlying means of the six conditions. In the "no effect" scenario, the means were identical in all six conditions. In the "small" scenario, the mean in one threat condition was 0.2 standard deviations lower than the means in the other conditions, which were identical to each other (d = 0.2). In the "medium" scenario, one condition produced a medium effect (d = 0.4). Within these scenarios, we simulated experiments with three different scale parameters for the prior v (r = 1, r = 0.5, and r = 0.2). For each combination of scale parameter and effect size, we simulated 1000 experiments using the adaptive algorithm, and 1000 experiments using a fixed design allotting equal numbers of participants to each condition. In both the fixed and adaptive-designed simulations, data were generated with an initial block of N = 180, with 30 assigned to each condition, and then subsequently in blocks of N = 6, up to a total of N = 2004 observations. For each simulated experiment with each combination of scale parameter and true effect size, we record the maximum $\log_{10}$(BF) value (i.e., the strongest evidence in favor of an effect) at the halfway point of the experiment (N = 1002) and at the conclusion of the experiment (N = 2004).

		True effect size		
Sample size	**Scale parameter**	*0.0*	*0.2*	*0.4*
1002	0.2	0.02	0.88	5.55
	0.5	-0.24	0.81	6.01
	1.0	-0.50	0.55	6.15
2004	0.2	-0.07	2.42	13.46
	0.5	-0.38	2.41	14.15
	1.0	-0.65	2.22	14.28

Supplemental Table 2. Average $\log_{10}$(BF) values at different true effect sizes, scale parameters, and sample sizes.

As shown in Supplemental Table 2, the scale parameter has little to no effect on the maximum $\log_{10}$(BF) value in the scenarios where there is a small (d = .2) and medium (d = .4) effect. In all cases, the experiment produces extreme evidence in favor of the (true) alternative hypothesis ($\log_{10}$(BF) > 2.0) by the conclusion of the experiment. In the scenario with a small effect, at the halfway point in the experiment, using r=1.0 results in a somewhat smaller $\log_{10}$(BF) value than using r=0.5 or 0.2, but all results in this column are in the category of "Moderate evidence in favor of the null hypothesis" (0.5 < $\log_{10}$(BF) < 1.0).

The scale parameter seems to have largest effect on the maximum $\log_{10}$(BF) value when there is no true effect. In that scenario, only r=1.0 results in a maximum $\log_{10}$(BF) value less than -0.5, on average. A maximum $\log_{10}$(BF) value less than -0.5 means that there was at least moderate evidence in favor of the null hypothesis (no effect) in all 9 comparisons. When the maximum $\log_{10}$(BF) value is not less than -0.5, it means that there was at least one comparison for which the experiment failed to produce at least moderate evidence in favor of the null hypothesis.

To assess the effect of the prior on the adaptive algorithm’s assignment of participants to conditions, we also recorded the number of participants that had been assigned to the condition where there was a true effect at the halfway point of each simulated experiment (N = 1002), and again at the conclusion of each simulated experiment (N = 2004). For the scenario where there was no true effect, we recorded the number of participants that had been assigned to an arbitrarily selected condition.

		True effect size		
Sample size	**Scale parameter**	*0.0*	*0.2*	*0.4*
1002	0.2	168	242	373
	0.5	167	253	386
	1.0	166	254	391
2004	0.2	335	636	873
	0.5	333	650	885
	1.0	331	656	891

Supplemental Table 3. Average number of participants assigned to the condition with the target effect at different true effect sizes, scale parameters, and sample sizes.

As shown in Supplemental Table 3, the algorithm distributes participants approximately one-out-of-six participants to each condition, regardless of the scale parameter. In the scenarios with a small or medium effect, the algorithm preferentially assigns participants to the condition with the effect. The values are very similar within each column, suggesting that the scale parameter has minimal influence on the degree to which participants are preferentially assigned to conditions.

Feasibility. We examined the feasibility of our proposal in two ways. First, we surveyed all our collaborating labs with IRB approval as to the number of Black participants they could expect to recruit if financial considerations were not a constraint. We also asked the amount of money they would need to meet this recruitment goal and compared the sum of these financial resources to our project budget.

The sum of these participants as of September, 2020, along with the characteristics of the sites that plan to recruit these participants, is shown in Supplemental Table 4. We estimate that our sites could recruit 2,700 participants. This recruitment goal exceeds what is needed according to our adaptive design simulations and is within our project budget.

		Sites		Expected participants	
		N	*%*	*N*	*%*
Institution	Public	15	56%	2,080	77%
	Private	12	44%	620	23%
% Black students	0% - 5%	8	30%	490	18%
	5% - 10%	8	30%	620	23%
	>10%	11	41%	1,590	59%
US region	East	8	30%	440	16%
	Midwest	9	33%	1,000	37%
	West	3	11%	200	7%
	South	7	26%	1,040	39%
Total		**27**		**2,700**	

Supplemental Table 4. Characteristics of the 27 sites with IRB approval that are involved in this study as of September, 2020. According to our estimates, the sites should be able to recruit 2,700 Black participants.

Second, we implemented the adaptive design in the formr online survey platform [64] and conducted an extensive series of tests to ensure that our implementation worked as expected. This testing verified whether three goals were possible using formr: that we could use previously collected data to inform successive waves of data collection, that we could accurately and rapidly compute the Bayes Factors necessary to update the condition assignment probabilities, and that the previous two steps could be combined, as required by our adaptive algorithm. Our testing revealed that all three goals could be achieved in formr, even during live testing.