Current Research in Ecological and Social Psychology 3 (2022) 100044

Contents lists available at ScienceDirect

Current Research in Ecological and Social Psychology

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

When a stereotype dumbfounds: Probing the nature of the surgeon = male belief☆

Kirsten N. Morehouse[a,*], Benedek Kurdi[a,b], Ece Hakim[a], Mahzarin R. Banaji[a,*]

[a] Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA
[b] Department of Psychology, Yale University, USA

ARTICLE INFO

Keywords:
Gender occupation stereotypes
Gender bias in medicine
Gender-neutral language
Surgeon riddle

ABSTRACT

"A father and his son are in a car accident. The father dies. The son is rushed to the ER. The attending surgeon looks at the boy and says, 'I can't operate on this boy. He's my son!' How can this be?" Fifty years after the riddle first received public attention, one likely answer proves elusive: the surgeon is the boy's mother. Seven studies ($N = 6{,}987$) were conducted to explore the vicissitudes of the surgeon = male stereotype. In Study 1, over 70% of participants failed to reach the mother solution. However, a reduction in bias was also observed: the percentage of mother inferences more than doubled when "son" was replaced with a gender-neutral kinship term ("child"), suggesting that even incidental exposure to gender-neutral language can loosen the grip of stereotypes. In fact, gender-neutral language was more effective in reducing bias than a condition ("daughter") with multiple mentions of the female gender. In Study 2, we replicated this finding in a nationally representative sample of the United States, and demonstrated that 82% of Americans failed to provide the mother inference in response to the classic riddle. Additionally, within this nationally representative sample, the demographic and psychological correlates of the surgeon = male stereotype were explored. In Studies 3–5, we interrogated the mechanisms of stereotype reduction in the child condition (Study 3), the degree to which this stereotype simply reflects base rates (Study 4), and eliminated an alternative explanation (Study 5). Finally, in Studies 6–7, the generalizability of the surgeon = male stereotype was tested and confirmed in a non-WEIRD country that supplies medical expertise to the world (India; Study 6), and the result was extended to an inverse gender–occupation stereotype (nurse = female; Study 7). Taken together, these data demonstrate the surprising strength of a gender occupational stereotype and its boundary conditions.

1. Introduction

> *"A father and his son are in a car accident. The father dies on the spot. The son is rushed to the ER. The attending surgeon looks at the boy and says, 'I can not operate on this boy. He's my son!' How can this be?"*

The surgeon riddle first entered the living rooms of Americans, to our knowledge, in an episode of the 1970s comedic TV series, *All in the Family* (Nicholl et al., 1972). Yet, even fifty years later, one likely answer proves elusive: *the surgeon is the boy's mother*. Indeed, in a pilot study that imposed no cognitive or time constraints, 80% of participants failed to report this solution (see Supplement 1). Instead, many respondents produced far-fetched answers such as "The father is a priest" or "The boy was kidnapped by a man and raised to believe that his kidnapper was his father [but the surgeon was the real father]."

In the present report, we explored the presumed source of this dumbfounding: the stereotype that connects surgery more with men than women. Through the lens of social role theory (Eagly, 1987, 2000; Eagly and Wood, 2012), which posits that occupational stereotypes emerge when one gender dominates in a specific role (Koenig and Eagly, 2014), the origin of the surgeon = male stereotype is perhaps unsurprising. After all, 80% of surgeons in the United States are men (AAMC, 2020, p. 29). However, what makes this particular manifestation of the stereotype interesting is its tenacity. Even with unlimited time to solve the riddle, the difficulty of imagining a woman as a surgeon appears to be so robust that it mentally precludes reaching the "mother"

☆ All data files and analysis scripts used in this project are available for download from the Open Science Framework (https://osf.io/2fmt5/). We thank the Harvard Department of Psychology, Time-Sharing Experiments for the Social Sciences, and Dean's Competitive Fund for Promising Research at Harvard University for supporting this research. Parts of these data were presented at the 21st Annual Meeting of the Society for Personality and Social Psychology, New Orleans, LA, in February 2020.

* Corresponding author.
E-mail addresses: kirsten_morehouse@g.harvard.edu (K.N. Morehouse), mahzarin_banaji@harvard.edu (M.R. Banaji).

https://doi.org/10.1016/j.cresp.2022.100044
Received 26 October 2021; Received in revised form 13 April 2022; Accepted 14 April 2022

response. Indeed, substantial anecdotal evidence, from our own teaching experience[1] and reports in news outlets (e.g., ABC, 2010; WBUR, 2013), suggests that respondents simply cannot seem to imagine the mother as the surgeon. Failure to produce the mother response is often accompanied by surprise bordering on shock, expressions of embarrassment, and a dumbfounding that, as the term suggests, precludes explanation.

We brought this conundrum into the laboratory because the apparent mental logjam created by the surgeon riddle is indeed perplexing. If the father is dead, and the surgeon says, "I am the boy's parent," then the answer "the surgeon is the boy's mother" should emerge immediately and effortlessly. But our own failure tells us that this is not the case.

Of course, "mother" is not the only solution to the riddle; the surgeon could be the deceased father's husband or the boy's stepfather. However, "mother" is amongst the most statistically likely responses, and yet most participants fail to produce it (e.g., Belle et al., 2021), even among a list of possible solutions. As such, the surgeon riddle provides an effective experimental task to observe the role of stereotypes in hindering the ability to draw a rational and reasonable inference.

The present work posed variations of the surgeon riddle to a total of 6987 participants, across seven studies (including a nationally representative sample of the United States), to explore four specific questions: (1) What is the magnitude of the surgeon = male stereotype?; (2) What are the demographic correlates of this stereotype?; (3) Can interventions be designed to modulate stereotype expression?; and (4) What is the underlying representation of the stereotype? That is, to what extent does it rest on essentialist thinking about gender?

When the surgeon riddle made its debut in the 1970s, women contributed less than 10% of applicants to medical school (AAMC, 2005). In 2019, women represented half of all medical school applicants for the first time in U.S. history (AAMC, 2019). Despite considerable strides toward improving the talent pool of medicine, women remain *overrepresented* in many lower-status fields (e.g., pediatrics and family medicine) while *underrepresented* in many high-status fields of medicine (e.g., surgery and internal medicine; AAMC, 2020; p. 29 –31). Indeed, gender imbalances are particularly pronounced within surgery: Although 43% of surgical residents today are female, women comprise only 22% of general surgeons. Moreover, surgical specialties constitute 7 of the 10 medical specialties with the largest gender gaps, with women representing less than 10% of the physicians in thoracic surgery, orthopedic surgery, and neurological surgery (AAMC, 2020; p. 29–31). As discussed above, these gender disparities are the most likely cause of the surgeon = male stereotype. However, in the surgeon riddle, participants have unlimited time and access to all manner of conscious thought that could generate the mother solution – a solution that hardly requires mental gymnastics. As such, if the data demonstrate a failure to provide the mother response, then it would demonstrate the power of stereotypes to interfere with logical inference and egalitarian ideals.

Three previous reports have employed the surgeon riddle (Belle et al., 2021; Reynolds et al., 2006; Skorinko, 2018; for German adaptations, see Kollmayer et al. 2018, Stoeger et al. 2004) to study various aspects of this stereotype, relying on relatively small samples of undergraduate students. However, college samples tend to be younger and more liberal than the general population (e.g., Campbell and Horowitz, 2016; Pew Research Center, 2016), and less likely to be biased on explicit and implicit measures of stereotyping (e.g., Charlesworth and Banaji, 2019). Two of the questions we sought to address, regarding the prevalence of the surgeon = male stereotype, and its variability across different segments of society, required a large sample that better approximated the United States as a whole. Accordingly, we recruited a nationally representative sample ($N = 3010$) to answer these questions.

The nature and scope of the nationally representative sample also allowed us to explore variations of the bias across demographic groups. A robust literature suggests that demographic groups vary in the degree to which they demonstrate gender stereotypes (Charlesworth and Banaji, 2021; Nosek et al., 2007). For instance, women display *weaker* implicit and explicit male–science/female–arts associations than do men (Charlesworth and Banaji, 2021). On the other hand, women display slightly *stronger* implicit male–career/female–family stereotypes than men (Charlesworth and Banaji, 2021). Finally, on explicit measures, both women and men seem to favor male job candidates and to the same degree (e.g., Moss-Racusin et al., 2012).

Using the surgeon riddle, Belle et al. (2021) observed variability in bias across one demographic group (gender[2]), but stability across another (ideology). We expanded upon this investigation, and examined variation across six demographic variables – *gender, political ideology, age, race, education,* and *income* – in a nationally representative sample. Further, building from literature demonstrating that exposure to counterstereotypic exemplars (e.g., female scientists; Leblebicioglu et al., 2011; 2021) can reduce stereotyping (for a review, see Olsson and Martiny 2018), we examined whether participants with exposure to counterstereotypic exemplars (e.g., female physicians) would be less likely to exhibit stereotypic responding in adulthood. In doing so, this analysis further refines the question of prevalence by revealing whether this stereotype appears ubiquitously at the societal level (i.e., all groups exhibit this stereotype to a similar degree) or more selectively at the group level (i.e., this stereotype is limited to or exacerbated in specific groups in society).

In addition to estimating the prevalence of the surgeon = male stereotype in the United States and its variability in the population, we examined potential *malleability* in the expression of this stereotype. Stereotypes in general have been shown to be resistant to change following short interventions (e.g., Lai et al., 2014; 2016), and a new analysis by Charlesworth et al. (2021) shows that gender stereotypes in a variety of sources (e.g., fiction and non-fiction books, TV shows, child-directed speech) have remained largely stable over time. And yet, implicit gender stereotypes at the *individual* level have decreased over the past decade (i.e., implicit male-science/female-arts and male-career/female-home stereotypes; Charlesworth and Banaji 2021). Moreover, previous research provides some confidence in the efficacy of interventions designed to break the "habit" of stereotyping (Carnes et al., 2015; Devine et al., 2012; Devine et al., 2017; Forscher et al., 2017), challenge stereotypic thought (Cheryan et al., 2011; Fuesting and Diekman, 2017), encourage individual-based (rather than category-based) thinking (e.g., Fiske and Neuberg, 1990), and increase perspective-taking (Catapano et al., 2019).

We examined whether a minimal intervention involving gender-fair language could modulate the expression of the surgeon = male stereotype. This intervention was selected because a burgeoning body of literature has examined the impact of gendered language on the expression of stereotypes (Banaji and Hardin, 1996; for a review, see Sczensy et al. 2016). Such work has converged to demonstrate that masculine generics (e.g., using "he" when gender is irrelevant) evoke a male bias in mental representation, which makes male exemplars of a person category (e.g., male scientists) more accessible than female exemplars of the same category (e.g., female scientists; Stahlberg et al. 2007). However, two gender-fair language (GFL) strategies have been employed to combat the detrimental effects of male-gendered language: (1) *language neutralization*, which involves replacing gender-specific words or pro-

[1] One author – Dr. Mahzarin Banaji – has posed the riddle to dozens of occupationally diverse U.S. and international audiences, consisting of thousands of employees of for-profit, non-profit, and governmental organizations. This anecdotal evidence produced a clear pattern: Only a small percentage of first-time listeners produced the "mother" response. Further, upon learning the mother solution, most listeners were dumbfounded. In fact, on several occasions, audience members reported being unable to report the mother solution, despite having a close female family member who is a surgeon.

[2] Women (36%; $n = 35$) were more likely to provide a mother response than men (18%; $n = 10$).

nouns (e.g., "policeman" and "he") with gender-neutral words or pronouns (e.g., "police officer" and "they"); and (2) techniques that make female exemplars more visible by using male and female generics symmetrically (e.g., "the surgeon…he *or she*…").

In the current investigation, we tested the efficacy of both strategies by making a small change to the surgeon riddle. Specifically, we devised three conditions that either referenced only male figures (father and his son), both male and female figures (father and his daughter), or a gender-indefinite figure (father and his child). Thus, by examining whether incidental exposure to gender-neutral or female-gendered language could modulate the expression of stereotyping, we (a) quantified the effect of using gender-neutral or specific female-gendered language, relative to traditional masculine language; and (b) compared the efficacy of two gender-fair strategies.

Finally, the present work investigated the underlying representation of the surgeon = male stereotype. Specifically, we measured the extent to which this stereotype is supported by statistical information (i.e., the prevalence of men versus women in surgery), essentialist beliefs (i.e., the belief that men are more naturally or innately suited for surgery, due to specific traits or physical abilities), or both. If the stereotype veridically tracks statistical patterns in the world, then increasing the representation of women in surgery should be sufficient to mitigate bias (Jussim, 2012). On the other hand, if essentialist beliefs support this stereotype, then interventions designed to dismantle these beliefs will likely prove more effective (Hammond and Cimpian, 2017). However, unlike beliefs that stem from observational learning, essentialist beliefs are not intrinsically tied to prevalence estimates (e.g., Gelman et al., 2004; Prasada et al., 2013), and persist in spite of countervailing evidence (Brandone et al., 2012; Leslie, 2008). As a result, if the surgeon = male stereotype is bolstered by essentialist beliefs, then it may prove harder to dislodge. We examine this malleability versus resistance to change in the present work.

1.1. Summary of the present work

Seven studies ($N = 6987$) were conducted to probe the nature of the surgeon = male stereotype. In Studies 1–2, we interrogated the *magnitude* of the surgeon = male stereotype in the United States. Additionally, we introduced a minimal intervention to explore whether incidental exposure to female-gendered or gender-neutral language can *reduce the expression* of this stereotype. In Studies 3–5, we examined the mechanisms that may explain the results of Studies 1–2, and tested two alternative explanations. Finally, in Studies 6–7, we probed the generalizability of the effects in two additional ways: (a) by testing the effect in a non-US sample (India, Study 6); and (b) by examining an additional gendered occupational stereotype, nurse = female (Study 7).

2. Study 1

In Study 1, the classic surgeon riddle, involving a father and his son, was presented alongside two variants of the riddle. In these variants, the gender of the child (son, daughter, or child) was manipulated. This manipulation allowed us to examine whether incidental exposure to a female-gendered (daughter) or gender-neutral (child) figure could *modulate* the strength of the surgeon = male stereotype that has been observed in our pilot study (see Supplement 1) and past work (e.g., Belle et al., 2021).

2.1. Method

2.1.1. Participants

606 participants ($M_{age} = 37.83$, 53% female, 76% White) residing in the United States were recruited from Amazon Mechanical Turk (MTurk), and paid $0.25 in exchange for 3–5 min of their time. The only requirement for inclusion was that participants had no previous exposure to the riddle or its possible solutions. 380 participants (63%) met this criterion.[3]

2.1.2. Open practices statement

All materials and measures used in this project are reported in Supplement 2. Statistical analyses were performed using the R statistical computing environment (R Core Team, 2021), and the data objects and R code required to reproduce all analyses reported in the article are available from the Open Science Framework (OSF; https://osf.io/2fmt5/).

2.1.3. Measures and materials

Riddle. In the classic riddle, a father and his *son* are in a car accident. To examine the impact of incidental exposure to gendered language, we created two additional variations of the riddle. Specifically, these variations described a father and his *daughter* or *child* (rather than *son*). The full text of all three riddles is included in Fig. 1.

Questions about demographic membership and counter-stereotypic exposure. Standard questions about the gender, age, race/ethnicity, political orientation, level of education of the participant were administered in the study. Participants were also asked to report whether they were born in the United States, the number of years that they have lived in the United States, and the zip code or county of their residence during childhood. Additionally, participants provided the (a) gender of their childhood physician(s); (b) gender of their sibling(s); and (c) whether their "[mother][father] work[ed] outside of the home while [they] were growing up." If participants answered that either parent worked outside of the home, then they were asked to report whether their parent(s) worked full-time or part-time, and, in a free-text entry, their occupation.[4]

2.1.4. Procedure

Participants were randomly assigned to read one of three versions of the surgeon riddle in which the gender of the child (son/daughter/child) was manipulated. After reading and providing an answer to the riddle, participants reported whether they had heard the riddle before and remembered its answer. Finally, participants responded to demographic items and an item about the gender of their childhood physician(s).

2.1.5. Coding of text responses

The primary dependent variable of interest was whether or not participants reported that the operating surgeon could be the mother (i.e., a woman). Two independent coders assigned participants' responses to one of fifteen categories (see Supplement 3). These categories were created via an iterative process. First, open-text responses from the pilot study (see Supplement 1) were submitted to an algorithm that searched for keywords commonly used in riddle responses, and categorized responses according to these keywords. For instance, the algorithm looked for strings such as "mother," "biological father," "priest," and "mistaken identity." Riddle responses that were not captured by the keywords were manually reviewed. If five or more of the remaining responses related to a new category (e.g., grandfather), then the algorithm was updated to include keywords related to these new categories. This process was repeated until the remaining responses could not be further combined in any meaningful way.

If a response included multiple answers *including* mother (e.g., "mother or stepfather"), then the response was coded as "mother." Otherwise, responses were categorized on the basis of the first possibility

[3] This inclusion criterion is standard for reports that employ the surgeon riddle. Further, the inclusion rate observed in the present study (63%) aligns with the rate observed in Belle et al. (2021), which was also 63%.

[4] The sample was not sufficiently large to examine the impact of specific parental professions (e.g., surgeon) on the type of solutions offered to the riddle. As such, we refrain from discussing these results further. However, the data are available for reanalysis from OSF (https://osf.io/2fmt5/).

Condition	Riddle Text[a]
Son (*Classic Riddle*)	A father and his **son** are in a car accident. The father dies on the spot. The **son**, badly injured, is taken to the hospital. In the ER, the attending surgeon looks at the **boy** and says, "I can't operate on this **boy**. **He's my son**!" How is this possible?
Daughter	A father and his **daughter** are in a car accident. The father dies on the spot. The **daughter**, badly injured, is taken to the hospital. In the ER, the attending surgeon looks at the **girl** and says, "I can't operate on this **girl**. **She's my daughter**!" How is this possible?
Child"	A father and his **child** are in a car accident. The father dies on the spot. The **child**, badly injured, is taken to the hospital. In the ER, the attending surgeon looks at the **child** and says, "I can't operate on this **patient**. **This is my child!**" How is this possible?

Fig. 1. Variations of the surgeon riddle
[a]Bolded words were manipulated across conditions.

mentioned. For instance, if a participant reported that the surgeon could be a "priest or ghost," then the response was coded as "priest." The categorizations made by each independent human coder were compared, and any disagreement between the two human coders was resolved by a third independent human coder, and additionally by discussion, when necessary.

After disagreements were resolved, responses that included "mother" were assigned a value of 1, and all other responses were assigned a value of 0. For more information about the coding of these responses, see Supplement 3.

2.1.6. Analytic approach

Bayesian generalized linear models were fitted to the data. Unlike frequentist approaches, Bayesian models yield a so-called posterior distribution, allowing researchers to (a) quantify evidence for or against the null hypothesis, and (b) make meaningful inferences from null findings (for a further discussion of the advantages, see Etz et al. 2018). Posterior distributions were used to conduct Bayesian equivalence testing (Kruschke, 2015). This testing requires researchers to set a region of practical equivalence (ROPE), which contains values consistent with the null hypothesis.

In this paper, the coefficients of interest are odds ratios. If two groups have an odds ratio of 1, then the odds of success (i.e., providing a mother response vs. a different response) are equal in both groups. As such, we defined the ROPE as an interval ranging from 0.9 to 1.1 (see, for example, Atkin et al. 2005). Data is said to conclusively support the null hypothesis (H_0) if >95% of the posterior distribution fell *within* the ROPE. Conversely, data is said to conclusively support the alternative hypothesis (H_1) if >95% of the posterior distribution fell *outside* of the ROPE. When the data supported the alternative hypothesis, we examined the *direction* of the effect. Specifically, we quantified the proportion of posterior distribution that fell to the right of the ROPE upper bound (PH_{1+}), suggesting a positive effect, or to the left of the ROPE lower bound (PH_{1-}), suggesting a negative effect.

Additional analyses were conducted in a frequentist framework. Specifically, the effect of gendered language was estimated by fitting a generalized linear model to the data. The effects of any experiential items (e.g., having a female doctor in childhood) were estimated by fitting a generalized linear model with the dichotomized mother response (1 = Mother response; 0 = All other answers) as the outcome variable, and riddle condition (i.e., son, daughter, child) and the variable of interest (e.g., gender of childhood doctor) as the two predictor variables. These variables were entered as main effects because in Study 1, and subsequent studies, we observed no interaction between any variables of interest and the riddle condition. Additionally, the riddle condition variable was simple coded such that the intercept reflected the grand mean of all conditions (rather than the mean of the intercept). The effect of each individual variable, or lack thereof, on mother responses was further analyzed by using likelihood ratio tests and pairwise comparisons. These metrics were used to compute estimated marginal means from specified factors in a linear model and comparisons among them, when necessary. These analytic strategies were also used in all subsequent studies, unless otherwise noted.

Table 1
Frequency of Riddle Solutions by Category.

Response Category	Frequency	Response Subcategory
Father	47.4%	Stepfather (19.5%) Adoptive Father (12.0%) Gay Fathers (9.8%) Two Fathers (3.0%) Biological Fathers (2.3%) Second Father* (<1%) (**Relation not specified*)
Mother	27.1%	*n/a*
Nonsense	21.1%	Other[a] (17.3%) Priest (2.3%) Grandfather (1.5%)
Unsure	4.5%	*n/a*

[a] "Other" refers to responses that were nonsensical but were not sufficiently populous.

2.2. Results and discussion

2.2.1. Classic son riddle: magnitude of bias

The magnitude of the stereotype was estimated by quantifying the proportion of participants who failed to report that the surgeon was the *boy's mother*. Strikingly, 72.9% of participants failed to provide this solution. That is, only 27.1% of participants provided a mother response. By contrast, 47.4% of participants offered a second father solution (e.g., gay fathers, biological father). An additional 21.1% of participants provided nonsensical answers such as "the doctor does not realize he has died," "[the surgeon] is a priest," and "the doctor is the ghost of his father"; and 4.5% of participants were unable to provide any solution (e.g., "I have no idea...I thought his dad was dead?"). For a breakdown of all responses by frequency, see Table 1.

Notably, this result was not confined to men. 80.6% of women ($n = 72$) and 63.9% of men ($n = 61$) failed to provide the mother response ($PH_{1+} = 0.971$, $p = .034$), further emphasizing the strength and robustness of the surgeon = male stereotype, and converging with prior evidence suggesting that even women internalize this stereotype (Belle et al., 2021; Salles et al., 2019). In fact, this initial data suggests that women may exhibit the surgeon = male stereotype to an even

stronger degree than men. In Study 2, we tested whether this finding persisted in a large nationally representative sample of the United States.

2.2.2. The effects of gendered language

Despite providing strong evidence for the robustness of the surgeon = male stereotype, the data also revealed interesting evidence of malleability. Specifically, manipulating the gender of the surgeon's offspring – "son," "daughter," "child" – significantly modulated the strength of this stereotype, $\chi^2(2) = 29.34$, $p < .0001$. As elaborated below, the "daughter" and "child" versions of the riddle produced higher proportions of mother responses, relative to the classic "son" version, suggesting that introducing female or gender-neutral descriptors can loosen the grip of the gender stereotype.

Female-Gendered ("Daughter") Condition. In contrast to the "son" version of the riddle, where 72.9% of participants failed to provide the mother response, 59.5% of participants in the daughter condition failed to provide the mother response ($PH_{1+} = 0.973$, $p = .060$). In other words, introducing female linguistic markers (e.g., "daughter," "she") increased the proportion of mother responses, suggesting this technique may successfully reduce reliance on the surgeon = male stereotype.

Gender-Neutral ("Child") Condition. A third condition represented the surgeon's offspring in gender-neutral terms ("child") to explore the efficacy of *language neutralization*. Strikingly, only 39.7% of participants in this condition failed to provide the mother response. Although it is noteworthy that a substantive proportion of participants still failed to report the mother response, this condition demonstrates the effectiveness of gender-neutral language on weakening the surgeon = male stereotype. Compared to the male-gendered "son" condition, introducing gender-neutral terms ("child," "they") doubled the percentage of mother responses, representing a significant reduction in stereotyping, $PH_{1+} = 1.0$; $p < .0001$.

Unexpectedly, the "child" condition produced almost 1.5 times more mother responses than the female-gendered "daughter" condition, which included multiple mentions of the female gender, $PH_{1+} = 0.997$; $p = .006$. In other words, replacing male-gendered descriptors with gender-indefinite descriptors was the most effective debiasing strategy. While unexpected, this finding converges with research demonstrating that replacing gender-specific words or pronouns ("policeman") with gender-neutral words or pronouns ("police officer") can reduce gender bias (for a review, see Sczesny et al. 2016). Further, it indicates that *language neutralization* may be better suited to reduce the stereotype expression than even explicit references to the female gender. In Study 3, the mechanisms underlying the reduction in the "child" condition were explored.

2.2.3. Impact of early counterstereotypic exemplars

Exposure to early counterstereotypic exemplars did not reliably reduce the surgeon = male stereotype in adulthood. Specifically, participants with early exposure to female physicians or a mother who worked outside of the home during childhood were no more likely to offer a "mother" response than participants with only exposure to a male physician in childhood ($PH_0 = 0.255$, $p = .994$) or a stay-at-home mother ($PH_{1+} = 0.857$, $p = .148$), respectively.

2.2.2. Additional tests: responses among participants with prior exposure

The main analyses reported here and in subsequent studies relied on data from participants with no prior exposure to the riddle.[5] Nevertheless, we additionally examined the subset of participants who had previously heard the riddle to demonstrate the role of prior exposure on responses. 226 participants (37% of the total sample) reported hearing the surgeon riddle before, and 77% of these participants reported remembering the riddle's solution. Collapsing across all three conditions,[6] only 18.6% of participants with prior exposure, and 5.2% of participants with prior exposure *and* memory of the riddle's answer failed to provide the mother inference. That is, participants with prior exposure overwhelmingly reported that the surgeon could be the boy's mother. Among participants who failed to produce the mother inference ($n = 42$), 73.8% reported that the surgeon was another father figure, and 26.2% provided nonsensical answers such as "the father who died was not the father of the child." Analyses regarding the effect of prior exposure on riddle responses for Studies 2–6 are reported in Supplement 6.

3. Study 2

In Study 1, we demonstrated both the robustness and malleability of the surgeon = male stereotype. However, the sample obtained in Study 1 was significantly more liberal and educated than the U.S. population (see Supplement 5), which raised two possible issues. First, the effect obtained in Study 1, and in previous reports, may have underestimated the true magnitude of bias, given that these demographic groups tend to be less biased on both explicit and implicit measures (Charlesworth and Banaji, 2019). Second, if there was heterogeneity in treatment effects, then the efficacy of the minimal intervention observed in Study 1 may not generalize to the population as a whole. For example, if the incidental exposure manipulation was sufficient to achieve malleability in liberal but not in conservative participants, then the reduction in stereotyping that was observed in Study 1 may not generalize. In Study 2, we directly addressed both possibilities by recruiting a nationally representative sample of the United States.

The size and nature of the sample also allowed us to explore variations of the bias across demographic groups, which was not possible in Study 1. Specifically, we expanded upon Belle et al. (2021) investigation of the effects of gender and political ideology, and examined variation across six demographic variables: *gender, age, race, political ideology, education,* and *income.* Additionally, we explored whether the stereotype stems from statistical information, essentialist beliefs, or both. If the stereotype simply reflects statistical knowledge about the actual gender distribution within the field of surgery (*frequency-based hypothesis*), then participants who provide higher estimates of male (relative to female) surgeons should be less likely to offer a mother response than participants who provide lower estimates of male (relative to female) surgeons. Alternatively, or in addition, participants who believe that men are more naturally suited to be surgeons may be less likely to report the mother response than participants who believe that women are more suited to be surgeons (*essentialist hypothesis*).

3.1. Method

3.1.1. Participants

3010 participants ($M_{age} = 47.88$, 52% female, 64% White) were recruited by Time-sharing Experiments for the Social Sciences (Freese and Visser, 2010) through the internet-based platform NORC AmeriSpeak Panel, which provides a nationally representative panel of adults living in the United States. The sample included respondents between the ages of 18 and 93 residing in all 50 states. As in Study 1, participants were retained for analysis if they reported no prior exposure to the riddle. 1657 participants met this inclusion criterion. For a comparison of the sample demographics between the full and final samples, see Supplement 5.[7]

[5] This constraint was imposed because "mother" is typically offered as the answer to the surgeon riddle, and the present data cannot arbitrate between participants who independently reached the mother inference, and participants who learned and later reported that the "mother" was the solution to the riddle.

[6] 22.2% of participants in the son condition, 17.1% of participants in the son condition, and 16.0% of participants in the son condition failed to provide the mother inference. Bayesian evidence in favor of a null result was only weakly suggestive for all contrasts (all $PH_{0s} < .150$; all $ps > .05$).

[7] In general, the final sample more closely resembled the demographic composition of the United States than the full sample. The only exception was that

3.1.2. Measures and materials

Riddle. Participants received one of the three riddle variations used in Study 1 (son, daughter, and child; see Fig. 1).

Questions about Demographic Membership and Counterstereotypic Exposure. Demographic data (e.g., age, race, and gender) were collected by NORC before and independently of the current study.[8] To examine the impact of exposure to close counterstereotypic exemplars on mother responses, participants were asked, "Are you or any of your close family members doctors or surgeons?" Participants responded by either reporting that none of their family members are doctors or surgeons or by specifying their relation(s) to male and female physicians. Specifically, participants were shown the following list of relations, and asked to select all of the following that are surgeons: (a) the participant themselves; (b) the participant's father; (c) the participant's mother; (d) the participant's spouse or significant other ([male][female]); (e) the participant's son; (f) the participant's daughter; (g) the participant's sister; (h) the participant's brother; or (i) an extended family member ([male][female]).

Psychological Correlates. Participants completed two self-report items designed to assess the frequency-based and essentialist hypotheses. To assess the *frequency-based* hypothesis, a gender estimation item asked participants to estimate "the gender breakdown of practicing surgeons in the United States" on a scale of "100% Male/0% Female" to "0% Male/100% Female" in increments of 10%. The left-most anchor of the scale ("100% Male/0% Female" vs. "0% Male/100% Female") was counterbalanced. To assess the *essentialist* hypothesis, participants were asked to report which of the following statements they most agreed with: (a) "Men are naturally more suited to be surgeons than women."; (b) "Men and women are naturally equally suited to be surgeons."; (c) "Women are naturally more suited to be surgeons than men."

3.1.3. Procedure

As in Study 1, participants were randomly assigned to read one of three versions (son/daughter/child) of the riddle. After providing an answer to the riddle, participants reported whether they had heard the riddle, or a similar version, before. Then, participants completed three self-report items – two items designed to assess the frequency-based and essentialist hypotheses, and one item designed to measure the impact of counterstereotypic exposure – in a counterbalanced order.

3.2. Results and discussion

3.2.1. Magnitude of bias (Classic "Son" riddle)

Remarkably, the vast majority – 82% – of participants failed to report that the surgeon could be the boy's mother. That is, only 18.1% of Americans provided the mother solution, suggesting that the surgeon = male stereotype is sufficiently robust, even in a sample that more closely resembles the United States population. In fact, given that 27.1% of participants reported the mother response in the convenience sample recruited in Study 1, these data suggest that the true estimate of the stereotype may be even stronger than previously reported.

Further, as in Study 1, roughly half (51.0%) of all participants offered a stereotypic response: the surgeon was the father. An additional 27.2% of participants provided a nonsensical responses such as, "maybe the father was not that son's father and just a father," "the son is dreaming," "the father who dies is a priest," and "the surgeon is a ghost." Although some participants seemed to recognize the implausibility of their responses – "[the] dad's soul could have reincarnated, but I'm not positive on my answer" – others were confident that the situation was simply "impossible". Indeed, a small portion (3.7%) of participants were not able to find any solution (e.g., "I have no logical answer as to how this could happen."; "Well the father passed so how can the surgeon be there to operate on his son if he has passed?"). In some cases, this dumbfounding persisted even after reading the riddle multiple times (e.g., "I am going to be honest, I read the riddle about 5 or 6 times slowly and still cannot figure it out"). Taken together, these data reveal the pervasiveness of the surgeon = male stereotype in the United States: Over 80% of Americans failed to report that the surgeon could be the boy's mother.

3.2.2. The effects of gendered language

Female-Gendered ("Daughter") Condition. Descriptively, using female kinship terms reduced stereotype expression, relative to the son condition. Specifically, in contrast to the "son" version of the riddle, where 81.9% of Americans failed to provide the mother response, 76.8% of Americans in the daughter condition failed to provide the mother response. However, this decrease was suggestive but not statistically significant ($PH_{1+} = 0.925$, $p = .092$).

Gender-Neutral ("Child") Condition. As in Study 1, the "child" condition produced the lowest levels of bias: 63.0% of Americans in the gender-neutral child condition failed to provide a mother response. Although only a minority of participants offered the counterstereotypic solution, further emphasizing the robustness of the surgeon = male stereotype, introducing a gender-neutral term ("child") significantly reduced stereotypic responding relative to the male-gendered son condition ($PH_{1+} = 1.0$, $p < .0001$). In fact, further replicating the pattern of results observed in Study 1, the "child" condition produced significantly lower levels of bias than the female-gendered "daughter" condition ($PH_{1+} = 1.0$, $p < .0001$), suggesting that gender-neutral language may be a more effective debiasing strategy than even multiple mentions of the female gender.

Generalizability of Effects. As described above, the nationally representative sample allowed us to examine two possible issues that arose in Study 1: (a) the effect obtained in Study 1 and in previous reports, may underestimate the true magnitude of bias, given their reliance on convenience samples; (b) if there is heterogeneity in treatment effects, then the efficacy of the minimal intervention observed in Study 1 may not generalize to the broader population. The latter concern was laid to rest by the fact that there was no heterogeneity in treatment effects; that is, we did not observe any statistically meaningful interactions between any of the demographic variables (e.g., gender, age) and the riddle conditions. However, the bias observed in the nationally representative sample (Study 2) was significantly greater than in Study 1 (all PH_{1}-s > 0.95). In fact, in the classic son condition, the convenience sample (Study 1) underestimated the magnitude of bias by almost 50 percent. This finding reveals the usefulness of relying on convenience samples in understanding basic mechanisms but highlights their limitations in producing accurate estimates of population parameters in an absolute sense.

3.2.3. Demographic variation

Harnessing data from a large nationally representative sample, we examined variation in mother responses across six demographic variables – *gender, age, race, political ideology, education,* and *income* – and one measure of counterstereotypic exposure. Perhaps most notably, women (73.2%) and men (74.2%) failed to report the mother response to similar degrees ($PH_0 = 0.610$, $p = .907$), suggesting that even members of the stereotyped group are susceptible to stereotypic responding. Similarly, even individuals with a close female family member who *is a doctor or surgeon* were not immune to bias; 67.9% of such participants failed to provide the mother solution (versus 74.1% of participants who were not related to a female physician; $PH_{1+} = 0.718$, $p = .354$).

The proportion of mother responses was also highly similar across political ideology, with 72.9% of conservatives, 74.4% of liberals, and 71.7% of independents failing to provide the mother solution (see Table 2). By contrast, mother responses significantly *decreased* with age, $X^2(1) = 10.92$, $p = .001$, $PH_{1} = 0.957$, and tended to *increase* with

the full sample underrepresented White participants to a slightly larger degree than the final sample.

[8] For a full list of demographic items, see http://tessexperiments.org/pdf/NORCAmeriSpeakTESSStandardDemographicVariables.pdf.

Table 2
Bayesian Regression of Mother Responses by Demographic Membership.

Demographic	Comparison (Intercept vs. Comparison)	M[a]	SD[a]	PH_0[b]	PH_{1-}[c]	PH_{1+}[d]
Gender	Female vs. Male	0.99	0.12	.610	.212	.177
Ideology	Conservative vs. Liberal	0.98	0.14	.516	.290	.194
Ideology	Independent vs. Conservative	0.94	0.15	.437	.421	.141
Ideology	Independent vs. Liberal	0.91	0.14	.414	.492	.094
Age	*n/a*	0.82	0.05	.043	**.957**	.000
Race	White vs. Black	0.60	0.11	.009	**.991**	.0001
Race	White vs. Asian	0.73	0.24	.135	.790	.076
Race	White vs. Latinx	0.57	0.09	.002	**.998**	.000
Race*	White vs. Black	0.68	0.12	.046	**.951**	.003
Race*	White vs. Asian	0.63	0.21	.078	.891	.031
Race*	White vs. Latinx	0.65	0.11	.021	**.978**	.001
Education	High School vs. Some College	1.13	0.18	.367	.089	.544
Education	Some College vs. Associate Degree	1.07	0.18	.429	.170	.401
Education	Associate Degree vs. bachelor's degree	1.64	0.24	.004	.000	**.997**
Education	Bachelor's Degree vs. Professional/Graduate Degree	1.47	0.26	.060	.003	.936
Income	Low vs. Lower Middle	1.49	0.32	.087	.014	.899
Income	Lower Middle vs. Middle	1.51	0.20	.009	.000	**.991**
Income	Middle vs. Upper Middle	1.81	0.35	.007	.000	**.993**
Income	Upper Middle vs. Higher	1.04	0.39	.218	.409	.373
Relation to Female Dr.	No Relation vs. Relation	1.36	0.41	.171	.111	.718
Prevalence Estimates	*n/a*	0.96	0.04	.937	.063	.000
Essentialist Beliefs	Women More Suited vs. Men More Suited	0.47	0.23	.031	**.951**	.018
Essentialist Beliefs	Both Genders Equally Suited vs. Men More Suited	0.56	0.13	.012	**.988**	.001
Essentialist Beliefs	Both Genders Equally Suited vs. Women More Suited	1.38	0.53	.164	.174	.663

[a] Mean and standard deviation of the posterior from the logistic regression results, in odds.
[b] $P(H_0)$ indicates the proportion of the posterior density that fell inside the pre-specified null region of [0.9, 1.1] (supports H_0).
[c] $P(H_{1-})$ indicates the proportion of the posterior density that fell outside and to the left of the null region, suggesting a negative effect, while.
[d] $P(H_{1+})$ indicates the proportion of the posterior density that fell outside and to the right of the null region, suggesting a positive effect.
* Controlling for the effects of education and income.

education and income (see Table 2). Finally, White Americans were conclusively more likely to provide the mother response than Black ($PH_{1-} = 0.991$, $p = .019$) or Latinx Americans ($PH_{1-} = 0.998$, $p = .003$), even after controlling for the effects of income and education (see Table 2).

3.2.4. Statistical knowledge and essentialist beliefs

Does the surgeon = male stereotype stem from statistical information (*frequency-based hypothesis*), or a belief about the true underlying nature of gender as a social category (*essentialist hypothesis*)? Although these possibilities are not mutually exclusive, the data only support the essentialist hypothesis.

Specifically, participants' estimates about the prevalence of male surgeons in the U.S. did not predict their likelihood of providing a mother response, $X^2(1) = 1.16$, $p = .281$, $PH_0 = 0.937$. Conversely, mother responses were predicted by participants' essentialist beliefs (see Table 2). Those who believed that *men are more naturally suited* to be surgeons were 15% more likely to provide a stereotype-consistent response than participants who believed that *both genders are equally suited* to be surgeons ($PH_{1-} = 0.988$, $p = .030$), and 20% more likely to provide a stereotype-consistent response than participants who believed that *women are more naturally suited* ($PH_{1-} = 0.951$, $p = .116$). Similar results were obtained in a multivariate analysis in which the probability of a mother response was predicted simultaneously by prevalence estimates and essentialist beliefs.

To summarize, Study 2 yielded three key results. First, over 80% of Americans failed to report the counterstereotypic response – *the surgeon is the boy's mother*. This constitutes the first prevalence estimate of the surgeon = male stereotype relying on a nationally representative sample, and suggests that this stereotype may be significantly stronger than previously reported. Indeed, although some demographic groups were less susceptible to bias (i.e., younger, White, and participants with higher levels of education and income), the rates of mother responses amongst even these groups remained remarkably low. For example, only 22% of Americans with a bachelor's degree or above reported that the surgeon could be the boy's mother.

Second, despite the strength of this stereotype, introducing gender-neutral language significantly reduced stereotyping. Specifically, as in Study 1, representing the surgeon's offspring in gender-neutral terms ("child") rather than masculine terms ("son") doubled the rate of mother responses. We take this result to suggest that small, incidental shifts in language can produce large shifts in bias. However, in Study 3, an alternative explanation was explored.

Third and finally, beliefs about men and women's natural suitability for surgery, but not participants' estimates about the prevalence of male surgeons in the United States, predicted their likelihood of providing a mother response. These results indicate that if increasing female representation impacts the success rate of the mother response, then it likely does so by erasing the essentialist belief that men are natively more suited to be surgeons. However, given the rigidity of essentialist beliefs (Brandone et al., 2012; Leslie, 2008), this finding also highlights the potential challenges of dismantling this stereotype.

4. Study 3

In Studies 1–2, participants who received the gender-neutral "child" variation of the riddle exhibited significantly less bias relative to both the *male*-gendered (son) and *female*-gendered (daughter) conditions. We place this result alongside others demonstrating that gender-fair language can increase the visibility of women in male-typed roles (Braun et al., 2005). However, these data can be explained another way. Specifically, if the concepts "mother" and "child" were associated in participants' minds, perhaps due to frequent co-occurrences of the two in language, then the word "child" may have activated the concept "mother," thereby increasing the likelihood of providing a mother response.

In Study 3, we tested this possibility in two ways. First, we assessed the existence of a preexisting association between the concepts "mother" and "child" using an Implicit Association Test (IAT; Greenwald et al., 1998). If no association existed, then this finding would undermine the validity of this alternate explanation. Second, we introduced a second gender-neutral condition (kid) alongside the three conditions used in

Studies 1–2 (son, daughter, and child). If both gender-neutral conditions (child and kid) reduced stereotyping to the same degree, then it is unlikely that the reduction observed in the child condition was a consequence of a preexisting association between "mother" and "child" alone. Instead, this finding would support the more general conclusion that gender-neutral language can reduce the expression of the surgeon = male bias.

4.1. Method

4.1.1. Participants

1315 volunteer participants ($M_{age} = 39.47$, 66% White, 68% female) were recruited from the Project Implicit website (https://implicit.harvard.edu).[9] 715 participants with no prior exposure to the riddle were retained for analysis.

4.1.2. Measures and materials

Riddle. Four versions of the surgeon riddle were employed. Specifically, in addition to the three variations used in Studies 1–2 (i.e., son, daughter, and child), a fourth kid variation was introduced. In this kid version, "a father and his kid" were in a car accident.

Demographic Items. Demographic information was collected by Project Implicit before and independently of the current study. For a full list of the demographic information collected, see Supplement 2.

Psychological Correlates. The items were identical to those employed in Study 2.

Mother/Father–Child/Kid IAT. Implicit associations between the concepts of "mother" and "father" with the attributes "child" and "kid" were measured using a seven-block IAT (Greenwald et al., 1998). Specifically, participants used two response keys (E and I) to sort words that represent the categories "mother" (i.e., *mother, mom, mommy, mama*) and "father" (i.e., *father, dad, daddy, papa*); and the attributes "child" (i.e., *child, children, my child, my children*) and "kid" (i.e., *kid, kids, my kid, my kids*). If participants were faster and more accurate when "mother" and "child" (and "father" and "child") shared a response key, then this result was taken to indicate a mother-child/father-kid association. If, however, participants were faster and more accurate when "mother" and "kid" (and "father" and "child") shared a response key, then this result was taken to indicate a mother-kid/father-child association.

4.1.3. Analytic approach

Performance on the IAT was assessed using the improved scoring algorithm (Greenwald et al., 2003) such that higher D scores index relatively stronger associations between "mother" with "child" and "father" with "kid," in line with the predicted direction of the effect. The analytic approach for all other analyses (e.g., condition effects) was identical to that of Studies 1–2.

4.1.4. Procedure

The study was comprised of three parts. In part 1, participants were randomly assigned to read one of the four versions of the surgeon riddle (son, daughter, child, and kid). As in previous studies, participants read the riddle, and subsequently reported (a) the answer to the riddle; (b) whether they had previously encountered some variation of the riddle; and (c) whether they remembered the answer to the riddle. In part 2, participants completed the Mother/Father–Child/Kid IAT. Finally, in part 3, participants responded to three self-report items (see "Psychological Correlates" section above) in a counterbalanced order.

[9] 67 participants were excluded prior to coding the riddle responses because (a) they did not complete the study ($n = 11$) or (b) their response latencies were below 300 ms on more than 10% of all IAT trials ($n = 11$), an indicator of insufficient attention (Greenwald et al., 2003).

4.2. Results and discussion

4.2.1. Mother/Father–Child/Kid IAT

Thus far, the reduction in stereotype expression observed in the child condition is presumed to be a consequence of gender-fair language. However, an alternative explanation is that this reduction is a product of a preexisting association between the concepts "mother" and "child," which renders the mother response more accessible in the child condition. This alternative explanation rests on the assumption that the concept "mother" is more closely associated to "child" than other gender-neutral kinship terms.

In Study 3, we tested this assumption by examining the relative association between the concept "mother" and "child" versus another gender-neutral kinship term: "kid." Consistent with this assumption, participants displayed a slight mother–child/father–kid association, IAT $D = 0.16$, $t(714) = 13.09$, $p < .001$, $d = 0.49$. Thus, if the "child" condition produced a significantly higher rate of mother responses than the gender-neutral "kid" condition, then the reduction observed in Studies 1–2 may have been a consequence of a preexisting association between "mother" and "child," rather than child's status as a gender-neutral descriptor. If, however, if the "child" and "kid" conditions produced similar proportions of mother responses, then it is unlikely that the stereotype reduction observed in the child condition was a consequence of this linguistic association alone. These competing explanations are explored below.

4.2.2. Effects of condition

Replicating Studies 1–2, 82.0% of participants who received the classic son variation failed to report that the surgeon could be female. Crucially, both the child and kid conditions significantly reduced stereotyping relative to the male-gendered son condition, and female-gendered daughter condition (all $PH_{1+}s > 0.995$). Specifically, 62.4% of participants in the child condition and 63.6% of participants in the kid condition failed to report the mother response. Although the proportion of mother responses in the child and kid conditions was highly similar, Bayesian evidence in favor of a null result was only weakly suggestive, $PH_0 = 0.340$, $p = .996$. These data imply that the difference in probability of mother response across the two versions of the riddle is diminishingly small. As such, it seems likely that the child condition implemented in Studies 1–2 increased the rate of mother responses by virtue of being a gender-neutral descriptor rather than (solely) by virtue of a preexisting conceptual association with the concept "mother."[10] Thus, these data lend even greater support to the previously reported result: gender-neutral language can loosen the grip of stereotypes on thought.

5. Study 4

What is the source of the surgeon = male stereotype? Does it stem from statistical information (*frequentist* hypothesis) or a belief about the essence of the social categories men and women (*essentialist* hypothesis)? Although not mutually exclusive, in Study 2, essentialist beliefs about gender differences in suitability for surgery, but not estimates about the gender ratio within the profession, predicted mother responses. Nevertheless, the percentage of participants who successfully generated the

[10] It's worth noting that even if the mother-child/father-kid association observed in the present study was driven by differential associations between father and "kid" over "child" (rather than differential associations between mother and "child" over "kid"), it is difficult to explain this pattern of results through the lens of increased accessibility alone. If father is more associated with "kid" than "child", then a pure accessibility account would predict that the concept father, and by association, the male gender, would be more accessible in the kid condition than in the child condition. This increased accessibility of the male gender should reduce the likelihood of mother responses in the kid condition, relative to the child condition. However, this was not the case; both gender-neutral descriptor conditions produced similar proportions of mother responses.

mother response closely resembles the actual percentage of female general surgeons in the United States, which is 22% (AAMC, 2020, p. 29). As such, while prevalence estimates did not predict mother responses at the individual level, it is conceivable that the group-level success rate in generating the mother response reflects statistical information about the prevalence of both genders in surgery (i.e., base rates). In Study 4, we introduced two new variants of the riddle (doctor and pediatrician) to test this possibility.

These variants were developed because the proportion of active physicians who are female (36%) and who specialize in pediatrics (64%) is considerably higher than the proportion of women in surgical specialties (AAMC, 2020, p. 29–31). Accordingly, if mother responses are merely reflective of the proportion of women within surgery, then the proportion of mother responses should be significantly higher when the riddle describes an attending doctor or pediatrician than when it describes a surgeon. In fact, the majority of participants should offer the mother solution in the pediatrician condition, given that almost two-thirds of pediatricians are female. Alternatively, if essentialist beliefs about the category "surgeon" (or, more generally, "doctor") are more influential in driving the effect, modifying the profession label may not increase the probability of mother responses, or at least not to the degree suggested by actual base rates.

5.1. Method

5.1.1. Participants

989 participants (M_{age} = 36.54, 63% White, 62% female) were recruited from the Project Implicit website.[11] 555 participants reported no previous exposure to the riddle and were therefore retained for analyses.

5.1.2. Measures and materials

Riddles. Three variants of the riddle were used. First, the classic surgeon riddle was used. As noted above, this riddle describes three figures – a father, his son, and a surgeon who is also the son's parent. The two other variants were novel and created by varying the profession label of the operant. Specifically, these variations described a doctor ("the attending doctor looks at the boy …") or pediatrician ("the attending pediatrician looks at the boy …") rather than a surgeon ("the attending surgeon looks at the boy …"). In all three conditions, the gender of the child was male.

Demographic Items. As in Study 3, demographic information was collected by Project Implicit.

Psychological Correlates. The self-report questions were identical to Studies 2–3.

Surgeon/Nurse–Male/Female IAT. Participants received a seven-block IAT designed to measure associations of the concept "surgeon" (over "nurse") with male more than female. In this IAT, participants sorted words related to the target categories "surgeon" (i.e., *Surgeon, General Surgeon, Surgery, MD*) and nurse (i.e., *Nurse, Head Nurse, Nursing, RN*), alongside male names (i.e., *Ben, Paul, Daniel, John, Jeffery*) and female names (i.e., *Rebecca, Michelle, Emily, Julia, Anna*).

Explicit Stereotypes. Participants received two items designed to capture their explicit stereotypes about surgery and nursing. Specifically, participants were asked how much they associate SURGERY and NURSING with male and female attributes on a scale of 1 (Strongly Female) to 7 (Strongly Male).

5.1.3. Procedure

The study was comprised of three parts. In part 1, participants were randomly assigned to receive one of three riddle conditions, in which the profession label (surgeon/doctor/pediatrician) of the physician was varied. Then, participants reported (a) the answer to the riddle; (b) whether they had previously encountered some variation of the riddle; and (c) whether they remembered the answer to the riddle. In part 2, participants responded to three self-report questions (see "Psychological Correlates" section above) and two questions about their explicit stereotypes in a counterbalanced order.

Finally, in part 3, participants learned the acronyms "RN" (Registered Nurse) and "MD" (Medical Doctor) – two stimuli used on the Surgeon/Nurse–Male/Female IAT – and then completed a comprehension check question to ensure that they retained the meanings of these acronyms. Following this comprehension check question, participants completed the Surgeon/Nurse–Male/Female IAT.

5.2. Results and discussion

The primary goal of Study 4 was to interrogate the possibility that the group-level success rate in generating the mother response reflects statistical information about the prevalence of both genders in surgery (i.e., base rates). Replicating Studies 1–3, the rate of mother responses in the classic surgeon riddle condition – 14.7% – was similar to the actual percentage of women in surgery (22%; AAMC, 2020). However, the rate of mother responses in the doctor and pediatrician conditions did not veridically track base rates. Specifically, although women constitute 36% of active physicians (AAMC, 2020, p. 29–31), only 9.5% of participants reported that the doctor could be female. In fact, the proportion of mother responses was highly similar across surgeon and doctor conditions, PH_{1-} = 0.890, p = .274, despite surgery touting only a fraction of the gender diversity.

Further evidence against the base rate account was provided by the pediatrician condition: Only 25.4% of participants offered the mother response, despite women constituting 64% of pediatricians in the United States (AAMC, 2020, p. 29–31). Thus, although the proportion successfully generating the mother response was significantly higher than the proportions observed in the surgeon ($PH_{1+} = 0.987$, $p = .030$) or doctor ($PH_{1+} = 1.0$, $p < .001$) conditions, it was dwarfed by the actual percentage of female pediatricians in the United States. Taken together, these data provide strong evidence that the group-level success rate in generating the mother response was not simply a reflection of statistical information. Instead, they suggest that gender stereotypes contributed to the widespread failure to recognize that a medical professional – surgeon, doctor, or pediatrician – could be female in a way that is not reducible to statistical information.

5.2.1. Implicit and self-report measures

The implicit and self-report measures employed in Study 4 were not relevant to the central aim of the study, which was to determine whether the rate of mother responses at the group-level reflects statistical information about the prevalence of both genders in surgery. As such, we report full details in Supplement 4.

6. Study 5

The results of Studies 1–4 demonstrated striking consistency and generalizability: The surgeon = male bias emerged in five independent samples, and across multiple iterations of the riddle (i.e., son, daughter, child, and kid). Further, Study 4 provided evidence to suggest that the rates of mother responses were not simply a reflection of the gender distribution within that field; instead, rates are driven, at least in part, by occupational stereotypes. However, is it possible that participants considered providing the mother response, but declined to report it? That is, because riddles typically require subtle or imaginative answers, might framing the problem as a riddle have suppressed the otherwise potentially obvious mother response? We tested this confounding explanation in Study 5 by framing the problem either as a riddle (as done in Studies 1–4) or, newly, as a story.

[11] 52 participants were excluded prior to coding the riddle responses because (a) they did not complete the study (n = 35) or (b) their response latencies were below 300 ms on more than 10% of all IAT trials (n = 17), an indicator of insufficient attention (Greenwald et al., 2003).

6.1. Method

6.1.1. Participants

663 participants ($M_{age} = 37.44$, 65% White, 69% female) were recruited from the Project Implicit.[12] Participants with no exposure to the riddle ($n = 346$) were retained for subsequent analyses.

6.1.2. Measures materials

Riddle Instructions. In Study 5, we manipulated the study's instructions. Specifically, participants received the classic riddle either framed as a riddle ("Please read this riddle carefully") or as a story ("Please read this story carefully").

Demographic Items. As in Study 3, demographic information was collected by Project Implicit.

Psychological Correlates. The self-report questions were identical to Study 3.

Surgeon/Nurse-Male/Female IAT. Participants received the Surgeon–Nurse/Male–Female IAT described in Study 4.

6.1.3. Procedure

The procedure was identical to Study 4, except that participants were randomly assigned to receive the classic riddle either framed as a riddle ("Please read this riddle carefully"), as it had been done in Studies 1–4, or as a story ("Please read this story carefully").

6.2. Results and discussion

When the problem was framed as a riddle (as it had been done in previous studies), 83.8% of participants failed to provide the mother response. When the problem was newly framed as a story, 87.6% of participants failed to provide the mother response. As such, framing the problem as a story rather than as a riddle did not significantly influence the rate of mother responses, $X^2(1) = 1.01$, $p = .315$, $PH_{1-} = 0.748$. Indeed, the vast majority of participants still failed to report that the surgeon could be the boy's mother, even in the story condition.

These data indicate that the low proportion of mother responses observed in Studies 1–4, as well as past work (Belle et al., 2021; Kollmayer et al., 2018; Reynolds et al., 2006; Skorinko, 2018; Stoeger et al., 2004) cannot be attributed to the framing of the problem as a "riddle." These data are instead consistent with the conclusion that the low rates of mother responses reflected a strong surgeon = male stereotype.

6.2.1. Implicit and self-report measures

As in Study 4, the implicit and self-report measures were not relevant to the central aim of Study 5, which was to interrogate whether might framing the problem as a riddle suppressed mother responses. As such, we report full details in Supplement 4.

7. Study 6

Study 5 demonstrated that participants' failure to report that a surgeon could be a woman was not an artifact of the task's framing. Instead, this failure seems to reflect a pervasive gender stereotype that connects surgery more with men than women. In Study 6, we tested whether the effects of the same surgeon = male stereotype similarly emerge in a sample recruited from India. India is characterized by a similar gender ratio in surgery as the U.S. (India Department of Higher Education, 2020), and provides a pipeline to U.S. medicine (Ranasinghe, 2015). However, India is less WEIRD (Western, Educated, Industrialized, Rich, and Democratic) than the United States. Because there is substantial variability in experimental findings across WEIRD and non-WEIRD populations, with WEIRD populations often producing outlying effects (see, for example, Henrich et al. 2010), this study provided an important test of generalizability.

7.1. Method

7.1.1. Participants

205 participants ($M_{age} = 32.78$, 22% female, 96% Asian) residing in India at the time of the study were recruited from Amazon Mechanical Turk (MTurk), and paid $0.25 in exchange for 3–5 min of their time. Participants with no previous exposure to the riddle ($n = 160$) were retained for analysis.

7.1.2. Measures and materials

Riddle. All participants received the classic riddle, in which a father and his son were in a car accident.

Demographic Items. The demographic items were identical to those administered in Study 1.

7.1.3. Procedure

The procedure was identical to that of Study 1, except that all participants received the classic son variation of the surgeon riddle.

7.2. Results and discussion

Although WEIRD populations often produce outlying effects (see, for example, Henrich et al. 2010), the results of Study 6 replicated the results obtained in Studies 1–5. Specifically, in Study 6, 65.6%, 95-percent CI: [58.1%, 72.7%], of participants failed to provide the mother solution. Notably, the pervasiveness of this effect was similar to the value of 71.9% obtained in the U.S. convenience sample that was also recruited from Mturk (Study 1), providing initial evidence that these results generalize to a non-WEIRD culture with similar gender imbalances in surgery.

It is worth noting that the rates of failure observed in the present sample were lower than the rates observed in Studies 4–5. However, these deviations are likely the product of sample characteristics, rather than cultural differences. The sample obtained in Study 6 was relatively young (81% were under 40 years old) and highly educated (91% had a bachelor's degree or above), two features that were related to increased rates of mother solutions in Study 2 (see Table 2). Moreover, as discussed above, the landscape of Indian medicine is similar to that of American medicine in ways that are relevant to the current report: gender parity exists at the medical school level, but women are vastly underrepresented in surgical specialties. Thus, these data indicate that the surgeon = male stereotype likely exists outside of the United States. In fact, given the sample characteristics of Study 6, and the results of the nationally representative sample recruited in Study 2, the present study may even underestimate the true magnitude of bias in India.

8. Study 7

The underrepresentation of women in medicine is especially pronounced in surgery, with surgical specialties contributing to 7 of the 10 ten specialties with the largest gender gaps (AAMC, 2020, p. 29–31). However, in other specialties (e.g., pediatrics) and subfields of medicine (e.g., nursing), men are similarly underrepresented. For instance, only 9% of registered nurses are male in the United States (Smiley et al., 2021). In Study 7, we introduced a nurse variant of the riddle to test the generalizability of the findings obtained from Studies 1–6 in an additional way, namely by estimating the magnitude of the nurse = female (rather than the surgeon = male) stereotype.

[12] 19 participants were excluded prior to coding the riddle responses because (a) they did not complete the study ($n = 11$) or (b) their response latencies were below 300 ms on more than 10% of all IAT trials ($n = 3$), an indicator of insufficient attention (Greenwald et al., 2003).

8.1. Method

8.1.1. Participants

199 participants (M_{age} = 36.81, 46% female, 75% White) residing in the United States at the time of the study were recruited from Amazon Mechanical Turk (MTurk), and paid $0.25 in exchange for 3–5 min of their time. Participants with no previous exposure to the riddle (n = 122) were retained for analysis.

8.1.2. Measures and materials

Riddle. All participants received the nurse riddle. The nurse riddle is a variation of the surgeon riddle in which a mother and her daughter (rather than a father and his son) are in a car accident, and the attending nurse (rather than the surgeon) refuses to treat (rather than operate on) the daughter (rather than son). Specifically, the nurse riddle is as follows:

> *A **mother** and her **daughter** are in a car accident. The **mother** dies on the spot. The **daughter,** badly injured, is taken to the hospital. In the ER, the attending **nurse** looks at the **girl** and says, "I cannot **treat** this **girl. She's my daughter!**" How is this possible?*

Demographic Items. The demographic items were identical to those administered in Study 1.

8.1.3. Coding of text responses

Responses[13] were coded using the same procedure outlined in Study 1, except that the answer of interest was "father," rather than "mother." Accordingly, instead of privileging mother responses, we privileged father responses – the analog to a mother response for the nurse riddle. That is, if father was included by the participant as a possible solution, then that response was categorized as "father" regardless of the order in which it was mentioned. Otherwise, answers were categorized on the basis of the first possibility mentioned. Finally, participants' answers were dichotomized such that responses that included "father" as a possibility were assigned a value of 1, and all other responses were assigned a value of 0. This dichotomized value, referred to as "father response" below, was the primary dependent variable of interest in Study 7.

8.1.4. Procedure

The procedure was identical to that of Study 1, except that all participants received the nurse riddle, rather than one of three versions of the surgeon riddle.

8.2. Results and discussion

8.2.1. Magnitude of bias

As in Studies 1–6, the data revealed a robust gender occupational stereotype. Only 16.4% 95-percent CI: [9.7%, 23 0.1%] of participants provided the counterstereotypic response: *The nurse is the girl's father*. By contrast, 60.7% of participants reported that the nurse was female (e.g., "the mother that died is the stepmother, that is the only explanation I can find."), and 23.0% of participants offered a nonsensical response (e.g., "the nurse may be confused"; "the mother was not in the car?"). As in previous studies, this result was not confined to men or women. Both men (17.5%) and women (15.3%) failed to report the father response at rates that were statistically indistinguishable, PH_0 = . 153, p = .742. Overall, these data resemble the results obtained in Studies 1–6, and suggest that gender stereotypes may similarly disallow imagining a *male nurse*.

[13] Interestingly, some participants mistakenly reported the identity of the doctor (e.g., "the doctor is the father"), rather than the nurse (e.g., "the nurse is the father"). Though the number of such responses was relatively small (N_{full} = 12, $N_{no\ exposure}$ = 6), the main analyses were conducted with and without these participants to interrogate whether the estimate of father responses was inflated by participants' mistaken interpretation of the question. However, including these participants did not substantially alter the pattern of results obtained. As such, these participants were retained.

9. General discussion

In the present work, seven studies were conducted to understand the magnitude and nature of a specific occupational stereotype: *surgeon = male*. Specifically, the surgeon riddle was used to understand the extent to which kidnappers and clergy, but not mothers, come to mind. This stereotype was selected for examination because it has been cited as a major deterrent for women pursuing a surgical career (Cochran et al., 2013). Therefore, insights gleaned from this report may not only inform future interventions designed to increase female participation in surgery, but also may generalize to other occupational stereotypes.

Building from literature demonstrating the usefulness of the surgeon riddle as a measure of stereotyping (Belle et al., 2021; Kollmayer et al., 2018; Reynolds et al., 2006; Skorinko, 2018; Stoeger et al., 2004), the present paper used this now famous riddle to ask four key questions: (1) What is the magnitude of the surgeon = male stereotype?; (2) What are the demographic correlates of this stereotype?; (3) What interventions modulate stereotype expression?; and (4) What is the underlying representation of the stereotype?

9.1. What is the magnitude of the surgeon = male stereotype?

Regarding the first question of magnitude, Studies 1–6 provided strong and consistent evidence for the robustness of the surgeon = male stereotype. Specifically, the majority of participants failed to report that the surgeon could be female in a nationally representative sample of the United States (Study 2), four US convenience samples (Study 1, Studies 3–5), and a convenience sample recruited from India (Study 6). The generalizability of this effect was demonstrated in two additional ways. First, the result persisted when the task was framed as a story rather than a riddle (Study 5), thereby eliminating an important confounding explanation: Participants considered the mother solution but declined to report it because riddles tend to require subtle and more imaginative answers. Second, low rates of counterstereotypic responding were similarly observed when the task focused on another gender stereotype within medicine: nurse = female (Study 7). Taken together, these data speak to the replicability and generalizability of these results.

9.2. What are the demographic correlates of this stereotype?

Given the robustness of the surgeon = male stereotype, we asked: are all demographic groups equally susceptible to the bias? To answer this question, we harnessed data from the nationally representative sample recruited in Study 2. Specifically, variability was examined across six demographic groups: gender, age, race, political ideology, education, and income.

Although men and conservative participants offered mother responses at similar rates as women and liberal participants, respectively, we observed variability across age, race, education, and income. Specifically, stereotypic responding significantly increased with age; White Americans were conclusively more likely to provide the mother response than Black or Latinx Americans; and lower levels of education and income were generally associated with greater stereotypic responding.

Nevertheless, it is worth noting that even amongst groups that were associated with reduced stereotypic responding, the rates of mother response remained low. For example, amongst participants with a bachelor's degree or higher, only 22% reported that the surgeon could be female in response to the classic riddle. In fact, almost 70% of participants with a close female family member who *is a doctor or surgeon* failed to provide the mother solution. Thus, these data reinforce and refine the main conclusion reached above: The surgeon = male stereotype is remarkably robust.

9.3. What interventions modulate stereotype expression?

As described in the introduction, stereotypes can perpetuate gender imbalances by (a) creating psychological barriers that dissuade individuals from pursuing gender-incongruent occupations (e.g., Eagly et al., 2000; He et al., 2019); (b) leading the stereotyped group to underperform (e.g., Nguyen and Ryan, 2008; Ryan and Nguyen, 2017; Shewach et al., 2019; Steele and Aronson, 1995; Zigerell, 2017); and (c) biasing evaluations of performance (e.g., Moss-Racusin et al., 2012). Thus, identifying manipulations capable of *modulating* the expression of stereotypes, such as the surgeon = male stereotype, is of both theoretical and practical interest.

In the present work, we examined the efficacy of gender-fair language, a strategy that aims to reduce stereotyping by eliminating asymmetries in referring to and addressing men and women (e.g., Mucchi-Faina, 2005; Stahlberg et al., 2007). Specifically, this minimal intervention involved incidental exposure to either a female-gendered (daughter) or gender-neutral (child) figure. Across three studies (Studies 1–3), incidental exposure to a gender-neutral figure (child) reduced the rate of stereotypic responding by approximately 50%, relative to the classic male-gendered son condition. These data suggest that gender-fair language can modulate the expression of at least one occupational stereotype: surgeon = male. Indeed, in Study 3, an alternative explanation for this reduction – it emerged because of a preexisting association between the concepts "mother" and "child" – was eliminated.

Unexpectedly, the gender-neutral child condition was more effective in reducing stereotyping than even the daughter condition, which included multiple mentions of the female gender. Although this result was not anticipated, it converges with a growing body of literature demonstrating that gender-fair language (e.g., replacing "policeman" with "police officer") can increase the visibility of women in male-typed roles (Horvath and Sczesny, 2016; Stahlberg et al., 2007).

In addition to efficacy concerns, language neutralization may also be preferred relative to the explicit use of feminine forms for at least two reasons. First, phrases such as "he *or she*", reinforce the gender binary. By contrast, gender-neutral language, such as generic third-person singular pronouns (e.g., "they"), can refer to a broader spectrum of gender identities. Second, the explicit use of feminine forms can have adverse outcomes for women. For example, in gendered languages such as Italian and German, occupational titles with female suffixes are perceived to be lower status (e.g., Merkel et al., 2012), and are often accompanied by a slightly derogatory connotation (e.g., Marcato and Thüne, 2002). Future work might explore whether feminine linguistic markers produce similar effects in natural gender languages like English.

9.4. What is the underlying representation of the stereotype?

Although we demonstrated the efficacy of gender-fair language in momentarily reducing bias in Studies 1–3, this manipulation may or may not durably change the underlying surgeon = male representation. However, if the underlying representation of this stereotype is known, then interventions can be designed to target the basic mechanisms underlying this bias. To understand the underlying representation of this stereotype, we examined whether statistical information (i.e., the perceived prevalence of men and women in surgery), essentialist beliefs (i.e., the belief that a certain gender is "naturally more suited for surgery"), or both, predicted stereotype expression (Study 2). Although not mutually exclusive possibilities, essentialist beliefs but not estimates about the prevalence of male surgeons in the U.S. predicted participants' likelihood of providing a mother response.

Although prevalence estimates at the *individual* level did not predict mother responses, might mother responses at the *group* level reflect statistical information about the prevalence of both genders in surgery? If this were the case, then altering the professional label of the operant to a profession with greater female representation (e.g., pediatrics) should have significantly increased the proportion of mother responses. Providing evidence against a base rate account, participants produced similar rates of mother responses, regardless of whether the riddle described a female doctor or surgeon (Study 4). In fact, only 25.4% of participants reported that a pediatrician could be female, despite the fact that 64% of pediatricians are female (AAMC, 2020, p. 29–31). As such, these data suggested that rates of mother responses at the *group* level did not veridically track statistical realities. Instead, they indicate that while observational learning may have *produced* the surgeon = male stereotype, it is likely *sustained* by essentialist beliefs, rather than statistical information.

9.5. Practical implications

Beyond these theoretical contributions, this work has substantial practical implications. Occupational stereotypes can create psychological barriers that dissuade individuals from pursuing gender-incongruent occupations (Olsson and Martiny, 2018), and may signal to women that members of their gender lack the skills necessary to be successful in male-dominated fields (e.g., STEM; Eagly et al., 2000). Further, even when women choose to pursue gender-incongruent roles, stereotypes may exacerbate existing disparities (e.g., by biasing evaluations of performance; Goldin and Rouse, 2000; MacNell et al., 2015; Moss-Racusin et al., 2012). These outcomes likely reduce the quality of talent that enters medicine, a deleterious consequence for any profession that seeks to attract the best individuals.

Indeed, the perception of surgery as "masculine" and as an "old boys' club" has been documented as a major deterrent to women pursuing a surgical career (Bellodi, 2004; Cochran et al., 2013; Hill et al., 2014), and across seven studies, we empirically demonstrated the robustness of the surgeon = male stereotype across situations and cultures. At the same time, this work illuminated the mechanisms that underlie participants' ability (and inability) to report that the surgeon could be a woman.

For one, these data indicated that the surgeon = male stereotype is likely *sustained* by essentialist beliefs rather than statistical information. This finding warrants attention because, unlike beliefs that stem from observational learning, essentialist beliefs are not intrinsically tied to statistical realities. As a result, the surgeon = male stereotype may persist even in the face of changing gender distributions within surgery. However, if female representation in surgery remains low, then this may reinforce the essentialist belief that "men (not women) are surgeons" (Eagly et al., 2000), further preventing women from pursuing a surgical career. Thus, these data suggested that interventions aimed at increasing representation should be applied in concert with those designed to dismantle essentialist beliefs about suitability in surgery.

In addition to potentially informing future interventions aimed at changing the underlying representation of this stereotype, the present work demonstrated the efficacy of a minimal intervention in temporarily alleviating the bias. Specifically, in three independent samples (Studies 1–3), the introduction of gender-fair language, i.e., using gender-neutral descriptors (e.g., "child," "kid") rather than gendered descriptors (e.g., "son," "daughter"), significantly increased the rates of counterstereotypic responding.

This data suggested that employing gender-neutral language in surgery, and perhaps medicine, more generally, may weaken the mental connection between surgery and men over women. For instance, consider a primary care doctor who is referring their patient to a surgeon. To make their language more gender-fair, the doctor could refer to the surgeon as "Dr. [Last Name]" and "they" when describing their expertise and qualifications. That is, the primary care doctor could refrain from using gendered pronouns (e.g., she) or first names, both of which may reveal the surgeons' gender, and potentially bias patient decision making.

While such practices may seem self-evident, in the United Kingdom and the Republic of Ireland, the custom of using the title "Dr" rather than "Mr" or "Mrs" to address surgeons was only instated within the past two decades (Loudon, 2000; Whelan and Woo, 2004). Further, even to-

day, professional titles are not used equally across genders (Files et al., 2017), and women are less likely than men to be referred to using their last name only (Atir and Ferguson, 2018). Thus, creating organizational guidelines around the use of pronouns and professional titles may alleviate bias.

10. Conclusion

Across seven studies, including a first nationally representative estimate, we demonstrated the robustness of the surgeon = male stereotype. Additionally, these studies shed light onto the mechanisms and generalizability of the effects. Specifically, they empirically demonstrated that (a) the effect emerges reliably across variations of the riddle (Study 5 and 7), and across cultures (Study 6); and (b) this stereotype is likely sustained by essentialist beliefs, rather than statistical information. Further, these data provided consistent evidence that incidental exposure to gender-neutral descriptors (e.g., "child") can, at least temporarily, alleviate bias. Thus, these data advance the understanding of a specific occupational stereotype, and in doing so, enrich our understanding of one path by which gender disparities in surgery are maintained.

Data sharing statement

All data files and analysis scripts used in this project are available for download from the Open Science Framework (https://osf.io/2fmt5/). Additionally, all measures and materials used in the present paper can be found in Supplement 2.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Kirsten N. Morehouse: Visualization, Formal analysis, Writing – original draft. **Benedek Kurdi:** Visualization, Formal analysis, Writing – original draft. **Ece Hakim:** Visualization, Writing – original draft. **Mahzarin R. Banaji:** Visualization, Writing – original draft.

Acknowledgments

We thank Charlotte Ruhl and Swathi Kella for their research assistance, and Tessa E. S. Charlesworth for conceptual feedback and support with study development.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cresp.2022.100044.

References

ABC, 2010. Good Morning America September 12. ABC News.
Aitkin, M., Boys, R.J., Chadwick, T., 2005. Bayesian point null hypothesis testing via the posterior likelihood ratio. Stat. Comput. 15 (3), 217–230. doi:10.1007/s11222-005-1310-0.
Association of American Medical Colleges (2005). Figure 1: representation of women M.D.s in academic medicine, 1965-2004. https://www.aamc.org/media/9991/download.
Association of American Medical College (2019). Figure 1. Percentage of applicants to U.S. medical schools by sex, academic years 1980-1981 through 2018-2019. https://www.aamc.org/data-reports/workforce/interactive-data/figure-1-percentage-applicants-us-medical-schools-sex-academic-years-1980-1981-through-2018-2019
Association of American Medical College (2020). 2020 Physician Specialty Data Report.; 2021 SRI A3273-25. https://hollis.harvard.edu/permalink/f/1mdq5o5/TN_cdi_proquest_statistical_2021_A3273_25.
Atir, S., Ferguson, M.J., 2018. How gender determines the way we speak about professionals. Proc. Natl. Acad. Sci. 115 (28), 7278–7283. doi:10.1073/pnas.1805284115.
Banaji, M.R., Hardin, C.D., 1996. Automatic stereotyping. Psychol. Sci. 7 (3), 136–141. doi:10.1111/j.1467-9280.1996.tb00346.x.
Belle, D., Tartarilla, A.B., Wapman, M., Schlieber, M., Mercurio, A.E., 2021. I can't operate, that boy is my son!": gender schemas and a classic riddle. Sex Roles 1–11. doi:10.1007/s11199-020-01211-4.
Bellodi, P.L., 2004. The general practitioner and the surgeon: stereotypes and medical specialties. Rev. Hosp. Clín. 59, 15–24. doi:10.1590/s0041-87812004000100004.
Brandone, A.C., Cimpian, A., Leslie, S.J., Gelman, S.A., 2012. Do lions have manes? For children, generics are about kinds rather than quantities. Child Dev. 83 (2), 423–433. doi:10.1111/j.1467-8624.2011.01708.x.
Braun, F., Sczesny, S., & Stahlberg, D. (2005). Cognitive effects of masculine generics in German: an overview of empirical findings. 30(1), 1–21. doi:10.1515/comm.2005.30.1.1.
Campbell, C., Horowitz, J., 2016. Does college influence sociopolitical attitudes? Sociol. Educ. 89 (1), 40–58. doi:10.1177/0038040715617224.
Carnes, M., Devine, P.G., Baier Manwell, L., Byars-Winston, A., Fine, E., Ford, C.E., Forscher, P., Isaac, C., Kaatz, A., Magua, W., Palta, M., Sheridan, J., 2015. The effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. Acad. Med. J. Assoc. Am. Med. Coll. 90 (2), 221–230. doi:10.1097/ACM.0000000000000552.
Catapano, R., Tormala, Z.L., Rucker, D.D., 2019. Perspective taking and self-persuasion: why "putting yourself in their shoes" reduces openness to attitude change. Psychol. Sci. doi:10.1177/0956797618822697.
Charlesworth, T.E., Banaji, M.R., 2019. Patterns of implicit and explicit attitudes II. Long-term change and stability, regardless of group membership. Am. Psychol. doi:10.1177/0956797618813087.
Charlesworth, T.E., Banaji, M.R., 2021. Patterns of implicit and explicit stereotypes III: long-term change in gender stereotypes. Soc. Psychol. Pers. Sci. doi:10.1177/1948550620988425.
Charlesworth, T.E., Yang, V., Mann, T.C., Kurdi, B., Banaji, M.R., 2021. Gender stereotypes in natural language: word embeddings show robust consistency across child and adult language corpora of more than 65 million words. Psychol. Sci. 32 (2), 218–240. doi:10.1177/0956797620963619.
Cheryan, S., Siy, J.O., Vichayapai, M., Drury, B.J., Kim, S., 2011. Do female and male 1008 role models who embody STEM stereotypes hinder women's anticipated success in 1009 STEM? Soc. Psychol. Pers. Sci. 2 (6), 656–664. doi:10.1177/1948550611405218, 1010.
Cochran, A., Hauschild, T., Elder, W.B., Neumayer, L.A., Brasel, K.J., Crandall, M.L., 2013. Perceived gender-based barriers to careers in academic surgery. Am. J. Surg. 206 (2), 263–268. doi:10.1016/j.amjsurg.2012.07.044.
Devine, P.G., Forscher, P.S., Austin, A.J., Cox, W.T.L., 2012. Long-term reduction in implicit race bias: a prejudice habit-breaking intervention. J. Exp. Soc. Psychol. 48 (6), 1267–1278. doi:10.1016/j.jesp.2012.06.003.
Devine, P.G., Forscher, P.S., Cox, W.T., Kaatz, A., Sheridan, J., Carnes, M., 2017. A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. J. Exp. Soc. Psychol. 73, 211–215. doi:10.1016/j.jesp.2017.07.002.
Eagly, A.H., 1987. Reporting sex differences. Am. Psychol. 42 (7), 756–757. doi:10.1037/0003-066x.42.7.755.
Eagly, A.H., Wood, W., Diekman, A.B., 2000. Social role theory of sex differences and similarities: a current appraisal. In: Eckes, T., Trautner, H.M. (Eds.), The Developmental Social Psychology of Gender. Erlbaum, Mahwah, NJ, pp. 123–174. doi:10.4324/9781410605245-12.
Eagly, A.H., Wood, W., van Lange, P., Kruglanski, A., Higgins, T.E., 2012. Social Role Theory. In P.A. Van Lange, A.W. Kruglanski, E.T. Higgins Handbook of theories of social psychology. volume 2, pp. 458–476, SAGE Publications Ltd.. doi:10.4135/9781446249222.n49.
Etz, A., Gronau, Q.F., Dablander, F., Edelsbrunner, P.A., Baribault, B., 2018. How to become a Bayesian in eight easy steps: an annotated reading list. Psychon. Bull. Rev. 25 (1), 219–234. doi:10.3758/s13423-017-1317-5.
Files, J.A., Mayer, A.P., Ko, M.G., Friedrich, P., Jenkins, M., Bryan, M.J., …, Hayes, S.N., 2017. Speaker introductions at internal medicine grand rounds: forms of address reveal gender bias. J. Women's Health 26 (5), 413–419. doi:10.1089/jwh.2016.6044.
Fiske, S.T., Neuberg, S.L., 1990. A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation. Adv. Exp. Soc. Psychol. 23, 1–74. doi:10.1016/s0065-2601(08)60317-2.
Forscher, P.S., Mitamura, C., Dix, E.L., Cox, W.T.L., Devine, P.G., 2017. Breaking the prejudice habit: mechanisms, timecourse, and longevity. J. Exp. Soc. Psychol. 72, 133–146. doi:10.1016/j.jesp.2017.04.009.
Freese, J., Visser, P.Principal Investigators, 2010. Data collected by time-sharing experiments for the social sciences. NSF Grant, 818839.
Fuesting, M.A., Diekman, A.B., 2017. Not by success alone: role models provide 1104 pathways to communal opportunities in STEM. Pers. Soc. Psychol. Bull. 43 (2), 163–176. doi:10.1177/0146167216678857.
Gelman, S.A., Taylor, M.G., Nguyen, S.P., 2004. Mother-child conversations about gender: understanding the acquisition of essentialist beliefs: IV. Talk about categories versus individuals (generics vs. non-generics). Monogr. Soc. Res. Child Dev. doi:10.1111/j.1540-5834.2004.06901005.x.
Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. Am. Econ. Rev. 90 (4), 715–741. doi:10.1257/aer.90.4.715.
Greenwald, A.G., McGhee, D.E., Schwartz, J.L., 1998. Measuring individual differences in implicit cognition: the implicit association test. J. Pers. Soc. Psychol. 74 (6), 1464. doi:10.1037/0022-3514.74.6.1464.
Greenwald, A.G., Nosek, B.A., Banaji, M.R., 2003. Understanding and using the implicit association test: I. An improved scoring algorithm. J. Pers. Soc. Psychol. 85 (2), 197. doi:10.1037/h0087889.

Hammond, M.D., Cimpian, A., 2017. Investigating the cognitive structure of stereotypes: generic beliefs about groups predict social judgments better than statistical beliefs. J. Exp. Psychol. Gener. 146 (5), 607. doi:10.1037/xge0000297.

He, J.C., Kang, S.K., Tse, K., Toh, S.M., 2019. Stereotypes at work: occupational stereotypes predict race and gender segregation in the workforce. J. Vocat. Behav. 115, 103318. doi:10.1038/466029a.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. Most people are not WEIRD. Nature 466 (7302). doi:10.1038/466029a, 29-29.

Hill, E.J., Bowman, K.A., Stalmeijer, R.E., Solomon, Y., Dornan, T., 2014. Can I cut it? Medical students' perceptions of surgeons and surgical careers. Am. J. Surg. 208 (5), 860–867. doi:10.1016/j.amjsurg.2014.04.016.

Horvath, L.K., Sczesny, S., 2016. Reducing women's lack of fit with leadership positions? Effects of the wording of job advertisements. Eur. J. Work Organ. Psychol. 25 (2), 316–328. doi:10.1080/1359432x.2015.1067611.

India. Ministry of Human Resource Development. Department of Higher Education. (2020). All India survey on higher education 2019-20. https://ruralindiaonline.org/en/library/resource/all-india-survey-on-higher-education-aishe-2019-20/

Jussim, L., 2012. Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy doi:10.1093/acprof:oso/9780195366600.001.0001.

Koenig, A.M., Eagly, A.H., 2014. Evidence for the social role theory of stereotype content: observations of groups' roles shape stereotypes. J. Pers. Soc. Psychol. 107 (3), 371. doi:10.1037/a0037215.

Kollmayer, M., Pfaffel, A., Schober, B., Brandt, L., 2018. Breaking away from the male stereotype of a specialist: gendered language affects performance in a thinking task. Front. Psychol. 9, 985. doi:10.3389/fpsyg.2018.00985.

Kruschke. (2015). Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan (2nd ed.). Academic Press is an imprint of Elsevier. doi:10.1016/B978-0-12-405888-0.09999-2.

Lai, C.K., Marini, M., Lehr, S.A., Cerruti, C., Shin, J.E.L., Joy-Gaba, J.A., …, Nosek, B.A., 2014. Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. J. Exp. Psychol. Gener. 143 (4), 1765. doi:10.1037/a0036260.

Lai, C.K., Skinner, A.L., Cooley, E., Murrar, S., Brauer, M., Devos, T., …, Nosek, B.A., 2016. Reducing implicit racial preferences: II. Intervention effectiveness across time. J. Exp. Psychol. Gener. 145 (8), 1001. doi:10.1037/xge0000179.

Leblebicioglu, G., Metin, D., Yardimci, E., Cetin, P.S., 2011. The effect of informal and formal interaction between scientists and children at a science camp on their images of scientists. Sci. Educ. Int. 22 (3), 158–174.

Leblebicioglu, G., Cetin, P., Eroglu Dogan, E., Metin Peten, D., Capkinoglu, E., 2021. How do science camps affect middle grade students' image of scientists? Res. Sci. Technol. Educ. 39 (3), 285–305. doi:10.1080/02635143.2020.1740667.

Leslie, S.J., 2008. Generics: cognition and acquisition. Philos. Rev. 117 (1), 1–47. doi:10.1215/00318108-2007-023.

Loudon, I., 2000. Why are (male) surgeons still addressed as Mr? BMJ 321 (7276), 1589–1591. doi:10.1136/bmj.321.7276.1589.

MacNell, L., Driscoll, A., Hunt, A.N., 2015. What's in a name: exposing gender bias in student ratings of teaching. Innov. Higher Educ. 40 (4), 291–303. doi:10.1007/s10755-014-9313-4.

Marcato, G., Thüne, E.M., 2002. Gender and female visibility in Italian. In: In Gender Across Languages: The Linguistic Representation of Women and Men, 2. John Benjamins Publishing Company, pp. 187–217. doi:10.1075/impact.10.14mar.

Merkel, E., Maass, A., Frommelt, L., 2012. Shielding women against status loss: the masculine form and its alternatives in the Italian language. J. Lang. Soc. Psychol. 31 (3), 311–320. doi:10.1177/0261927X12446599.

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J., Handelsman, J., 2012. Science faculty's subtle gender biases favor male students. Proc. Natl. Acad. Sci. 109 (41), 16474–16479. doi:10.1073/pnas.1211286109.

Mucchi-Faina, A., 2005. Visible or influential? Language reforms and gender (in) equality. Soc. Sci. Inf. 44 (1), 189–215. doi:10.1177/0539018405050466.

Nicholl, D. (Writer), LaHendro, B. (Director), & Livingston, R. (Director). (1972, March 11). Gloria and the Riddle (Season 3, Episode 4) [TV series episode]. In N. Lear (Executive Producer), *All in the Family*. CBS. https://www.imdb.com/title/tt0509876/.

Nosek, B.A., Smyth, F.L., Hansen, J.J., Devos, T., Lindner, N.M., Ranganath, K.A., …, Banaji, M.R., 2007. Pervasiveness and correlates of implicit attitudes and stereotypes. Eur. Rev. Soc. Psychol. 18 (1), 36–88. doi:10.1080/10463280701489053.

Nguyen, H.H.D., Ryan, A.M., 2008. Does stereotype threat affect test performance of 1314 minorities and women? a meta-analysis of experimental evidence. J. Appl. Psychol. 93 (6), 1314–1334. doi:10.1037/a0012702.

Olsson, M., Martiny, S.E., 2018. Does exposure to counterstereotypical role models influence girls' and women's gender stereotypes and career choices? A review of social psychological research. Front. Psychol. 9, 2264. doi:10.3389/fpsyg.2018.02264.

Pew Research Center. (2016). A wider ideological gap between more and less educated adults. *Pew Research Center: US Politics and Policy*. https://www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/.

Prasada, S., Khemlani, S., Leslie, S.J., Glucksberg, S., 2013. Conceptual distinctions amongst generics. Cognition 126 (3), 405–422. doi:10.1016/j.cognition.2012.11.010.

R Core Team (2021). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ranasinghe, P.D., 2015. International medical graduates in the US physician workforce. J. Osteopath. Med. 115 (4), 236–241. doi:10.7556/jaoa.2015.047.

Reynolds, D., Garnham, A., Oakhill, J., 2006. Evidence of immediate activation of gender information from a social role name. Q. J. Exp. Psychol. 59 (5), 886–903. doi:10.1080/02724980543000088.

Ryan, A.M., Nguyen, H., 2017. Publication bias and stereotype threat research: a reply to Zigerell. J. Appl. Psychol. 102, 1169–1177. doi:10.1037/apl0000242.

Salles, A., Awad, M., Goldin, L., Krus, K., Lee, J.V., Schwabe, M.T., Lai, C.K., 2019. Estimating implicit and explicit gender bias among health care professionals and surgeons. JAMA Netw. Open 2 (7). doi:10.1001/jamanetworkopen.2019.6545, e196545-e196545.

Sczesny, S., Formanowicz, M., Moser, F., 2016. Can gender-fair language reduce gender stereotyping and discrimination? Front. Psychol. 7, 25. doi:10.3389/fpsyg.2016.00025.

Shewach, O.R., Sackett, P.R., Quint, S., 2019. Stereotype threat effects in settings with features likely versus unlikely in operational test settings: a meta-analysis. J. Appl. Psychol. doi:10.1037/apl0000420.

Skorinko, J.L.M., 2018. Riddle me this: using riddles that violate gender stereotypes to demonstrate the pervasiveness of stereotypes. Psychol. Learn. Teach. 17 (2), 194–208. doi:10.1177/1475725717752181.

Smiley, R.A., Ruttinger, C., Oliveira, C.M., Hudson, L.R., Allgeyer, R., Reneau, K.A., …, Alexander, M., 2021. The 2020 national nursing workforce survey. J. Nurs. Regul. 12 (1), S4–S96. doi:10.1016/S2155-8256(21)00027-2.

Stoeger, H., Ziegler, A., David, H., 2004. What is a specialist? Effects of the male concept of a successful academic person on the performance in a thinking task. Psychol. Sci. 46 (4), 514–530.

Stahlberg, D., Braun, F., Irmen, L., Sczesny, S., 2007. Representation of the sexes in language. In: Fiedler, K. (Ed.), Social Communication. Psychology Press, pp. 163–187.

Steele, C.M., Aronson, J., 1995. Stereotype threat and the intellectual test performance of African Americans. J. Pers. Soc. Psychol. 69 (5), 797. doi:10.1037/0022-3514.69.5.797.

WBUR, 2013. Blindspot: Hidden Biases of Good People February 18. Radio Boston.

Whelan, C., Woo, H.H., 2004. Mister or Doctor? What's in a name? Med. J. Aust. 181 (1). doi:10.5694/j.1326-5377.2004.tb06151.x.

Zigerell, L.J., 2017. Potential publication bias in the stereotype threat literature: comment on Nguyen and Ryan (2008). J. Appl. Psychol. 102, 1159–1168. doi:10.1037/apl0000188.