Exploring how gender-anonymous voice avatars influence women’s performance in online computing group work

Dominic Kao, Syed T. Mubarrat, Amogh Joshi, Swati Pandita, Christos Mousas, Hai-Ning Liang, Rabindra Ratan

PII: S1071-5819(23)00155-6
DOI: https://doi.org/10.1016/j.ijhcs.2023.103146
Reference: YIJHC 103146

To appear in: International Journal of Human - Computer Studies

Received date: 15 March 2023
Revised date: 15 August 2023
Accepted date: 7 September 2023


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).
Exploring How Gender-Anonymous Voice Avatars Influence Women’s Performance in Online Computing Group Work

DOMINIC KAO, Purdue University, USA
SYED T. MUBARRAT, Purdue University, USA
AMOGH JOSHI, Purdue University, USA
SWATI PANDITA, California Institute of Technology, USA
CHRISTOS MOUSAS, Purdue University, USA
HAI-NING LIANG, Xi’an Jiaotong-Liverpool University, China
RABINDRA RATAN, Michigan State University, USA

We investigate how gender-anonymous voice avatars influence women’s performance in online computing group work. Female participants worked with two male confederates. Voices were filtered according to four voice gender anonymity conditions: (1) All unmasked, (2) Male confederates masked, (3) Female participant masked, and (4) All masked. When only male confederates used masked voices (compared to all unmasked), female participants spoke for a longer period of time and scored higher on computing problems. When everyone used masked voices (compared to all unmasked), female participants spoke for a longer period of time, spoke more words, and scored higher on computing problems. Effects were not significant on subjective measures and one behavioral measure. We discuss the implications for virtual interactions between people.

1 INTRODUCTION

Teams working together remotely became commonplace during the COVID-19 pandemic [25], driven by factors such as cost reduction and flexibility in work-life balance [53], with many companies now offering positions that are permanently remote [157]. Online education is also growing—the global market of $269.87 billion USD in 2021 is expected to increase to $585.48 billion USD in 2027 [136]. Communication platforms such as Zoom, Microsoft Teams, Cisco Webex, Slack, and Discord are now common in businesses, education, and entertainment [17]. However, communication on such platforms may reinforce social inequities in the workplace and online education. For example, multiple studies have found that women experience significantly higher zoom fatigue than men [52, 133]. The present research is concerned with the potential for online teamwork to augment the already high level of disparity for women in computing fields, where fewer than 20% of computer science students are women [73, 81, 141], in part due to historical [71] and cultural biases [2]. Further, stereotypes that women do not belong in technological fields engender hostile and sexist environments that hinder equitable participation [38, 44] and trigger stereotype threat, the phenomenon that reminders of a negative stereotype about a social group increase the risk that members of the group will conform to the stereotype or otherwise perform poorly [89, 103, 162]. Stereotype threat—one of the most widely studied topics in social psychology [146] across various domains [103, 131, 186]—has been implicated in long-standing racial and gender inequalities in academic performance [50, 61, 122], but has not been studied extensively in the context of online teamwork. Hence, the present research addresses this gap, recognizing that gender diversity creates better performing teams [63, 177, 187] and offering a theoretically grounded...
approach to mitigating stereotype threat and promoting more equitable online participation for women in technological team meetings.

In particular, we build on research suggesting that avatars—defined as mediated representations of humans used to facilitate real-time interactions [119]—play an important role in stereotype threat effects within mediated communication. For example, one study found that African-American participants using race-anonymized avatars persisted significantly longer in a cognitive task with other ostensible participants than those using race-revealing avatars [93]. Another study found that both men and women participants who used a female avatar in a threat condition (i.e., in a group with two ostensible male avatars) performed worse on a math task than participants who used a male avatar [96]. Another study found that using a female avatar in VR harmed math task performance and math confidence, while using a male avatar buffered against these stereotype threat effects [129]. In all of these studies, a visual avatar was assigned to the participants, and although a few studies have examined the role of avatar customization in these effects [54, 134], no studies of which we are aware have examined the role of non-visual avatars in stereotype threat effects. Addressing this gap, the present research focuses on mediated representations of user voices, which can be considered voice avatars, and which carry important cues of user identity that are potentially stereotyped, such as gender.

Although voice avatars are a novel technology in the context of online teamwork, there is significant evidence that voice-based communication in online gaming environments triggers hostility and sexism toward women [91, 111]. Further, although participants in an online meeting can use a visual avatar on many platforms, few if any offer the option for voice avatars, and the effects of using such avatars are largely unknown. Hence, the lack of research on voice avatars and stereotype threat represents a significant gap.

The technology required to mask or modify user voices is advancing rapidly [83], so we may expect that team communication platforms will offer these tools in the future, but the best approaches to mitigating stereotyping and stereotype threat through such tools are unclear. As a first step toward exploring this technology and the role of stereotype threat in online team communication, we focus on a simple comparison of gender-revealing and gender-masking voice avatars.\footnote{We recognize that gender is often treated as a non-binary concept [3, 15, 98]. This paper focuses on stereotype threat related to women in computing contexts, using men as the comparison gender category, consistent with previous gender-stereotype research (e.g., [58, 66, 118]). This approach does not intend to dichotomize gender, but leaves room for future research to examine stereotype threat and other gender categorizations.} Hence, for the purposes of this study, we created a faux online meeting platform that allowed our self-identifying female participants to use either a gender-revealing or gender-masking voice avatar to interact with two ostensible male teammates (actually recorded confederates). Results suggest that gender-masking voice avatars buffered the effects of stereotype threat as reflected by the amount participants spoke and performance on computing problems, though subjective measures were not found to be affected. Together, this large-scale preregistered study suggests that stereotype threat is likely rampant in online technical meetings and that voice (and visual) avatars can likely be used to promote gender equity in such teamwork across a variety of contexts (virtual classes, business meetings, and entertainment).

2 RELATED WORK

2.1 Stereotype Threat

Stereotype threat is a type of psychological discomfort triggered when people with a negatively-stereotyped social identity feel at risk of confirming the negative stereotype and consequently experience adverse effects in the stereotype-relevant domain [161, 163]. In the seminal study on
stereotype threat [163], African American students underperformed on the verbal Graduate Records Examination compared to their White counterparts when the test was framed as an assessment of intellectual ability or students were asked to record their race/ethnicity in a questionnaire prior to completing the task. However, no racial performance differences were found when the test was framed as a problem-solving exercise or the race/ethnicity question was administered after the test. This performance difference likely occurred because African-Americans are often negatively stereotyped as inferior to White students in intellectual domains, and thus the test framing that increased the negative stereotype’s salience induced a psychological threat that inhibited their performance. These findings provide evidence that stereotypical views and beliefs about a group can negatively influence the intellectual functioning of individual group members.

2.1.1 Domains, Mechanisms, and Triggers of Stereotype Threat. As one of the most widely studied topics in social psychology [146], stereotype threat has been shown to hinder performance in a wide variety of domains, including mathematics, computer science, athletics [16, 50, 61, 122, 131, 143, 163, 167, 174, 186], intelligence tests [40], engineering [14, 99], mathematical reasoning [76, 106, 142, 158], political knowledge [107], memory [72, 97], English [16]. But in addition to directly harming performance [152], stereotype threat affects a variety of other factors, such as sense of belonging [33, 115, 155, 174, 185], motivation [32, 115, 165], interest [156], and career intentions [31, 42, 65]. And although many stereotype threat studies have focused on racial and gender stereotypes, the effect has been found for numerous stigmatized groups, such as Latinos [62], Native Americans [122], older adults [72, 97], and people with low socioeconomic status [40].

There are numerous (debated) psychological mechanisms for stereotype threat [152], such as anxiety, stress, and arousal [16, 19, 123, 124, 162], which may hinder attention [68] and cognitive efficiency [51]. Studies have found evidence of the mediating role of anxiety on stereotype threat outcomes as reflected in measures of communication fluency [108, 125] and psychophysiological activity—such as mean arterial pressure reactivity [19]. skin conductance, surface skin temperature, and diastolic blood pressure [123]. Many situational factors can trigger stereotype threat [75, 116, 164], such as the test’s diagnosticity—the extent to which a test is framed as an evaluation of a negatively stereotyped individual’s skills/abilities [14, 27, 40, 62, 79, 107, 108, 163, 172]—or if a test/exam is perceived as highly important [41, 90, 144]. Stereotype threat results from even subtle cues to test diagnosticity, such as telling participants they will receive performance feedback after a test [102, 145, 163] or that their strengths and weaknesses are being investigated [79, 163].

Furthermore, the type of diagnosticity makes a difference—e.g., women suffered more from stereotype threat when they were led to believe their leadership ability (stereotypically masculine skill) was being evaluated compared to being told relationship maintenance ability (stereotypically feminine skill) was being evaluated [108]. Stereotype threat is also triggered when a stereotyped identity is made salient [11, 12, 19, 24, 46, 80, 86, 108, 109, 120, 123, 138, 158, 156, 168]. For instance, women who were told that men are better than women in mathematical domains exhibited stereotype threat effects during a subsequent computer science programming task [156] and math problems [12, 86]. Implicit cues to stereotype relevance can also trigger stereotype threat, for example, by telling female participants that their performance on a math test would be compared to men [46, 138] or that the objective of an experiment was to investigate differences between men and women in athletics [168] and mathematics [24, 80, 109, 158]. Similarly, job descriptions utilizing more masculine-themed words compared to feminine or neutral words predicted lower belonging for potential female candidates [169] and caused women to report lower scores of expected belongingness and job appeal [57]. And perhaps even more implicitly, being asked to indicate gender, race/ethnicity, or age in questionnaires [41, 106, 145, 153, 163], being in physical [32] or virtual [31] environments decorated with stereotypically masculine objects, or simply being numerically
underrepresented (e.g., women taking a test in a majority-male room) increases the likelihood of stereotype threat effects [9, 76–79, 115, 150].

Relatedly, the presence or behavior of others—such as someone who makes sexist remarks [99] or simply a male experimenter [107]—can also trigger stereotype threat [100, 125, 168]. Similarly, women were more likely to experience stereotype threat after being led to believe they were competing against a man in chess [100] or a shooter video game [54]. Because people possess multiple social identities (e.g., gender, age, race/ethnicity, socioeconomic status, etc.), stereotype threat also occurs when a social identity relevant to the stereotyped context is made more salient than other social identities [137]. For example, Asian women suffered from stereotype threat during a math test after filling out a questionnaire about their gender identity, but not when they answered a questionnaire about their ethnic identity [153]. Similarly, female students primed to be more self-aware about their gender identity performed worse on a standardized spatial reasoning test compared to women who were made aware of their identity as students [106].

2.2 Mitigating Stereotype Threat Online

Given the overwhelming evidence that stereotype threat occurs widely in society, especially among students, researchers should focus on developing approaches to combat it. Small effective steps to mitigate stereotype threat improves overall well-being, reduces worries about being devalued based on group membership (e.g., ethnicity, gender), facilitates meaningful engagement with peers, instructors, and classroom learning activities, and consequently reduces achievement gaps by 30%-40% [37] in the short and long term [35, 36]. Approaches to mitigating stereotype threat generally require altering the psychological environment of the individual [164]. For example, just as framing a test as diagnostic triggers stereotype threat, framing a test as generic will counteract it [163].

In contrast, the present research focuses on stereotype threat mitigation through social-identity mechanisms. Namely, just as increasing a stereotyped social identity’s salience triggers stereotype threat, reducing cues to social identity mitigates it [56]. The present research focuses on a context of performance where such social identity cues are particularly poignant and also malleable: online group collaborations.

Group interactions and collaborations are increasingly taking place online across many contexts, from undergraduate education [126] to academic research conferences [189] to work meetings in general [84]. In all of these contexts, participants create and share content and engage with others socially through digital self-representations that often reflect their social identities [22, 130]. Similar to social identity offline, online social identity can be threatened by social cues in the digital environment, and so approaches to mitigate stereotype threat offline can also work in online contexts [29]. However, digital environments are different because individuals are afforded greater control over how other users perceive them by selectively presenting (or hiding) aspects of their personal identity as a part of their online identity—something that is more difficult to achieve in offline, face-to-face communication [184]. Hence, we focus on how digital self-representations online may be used to mitigate stereotype threat.

2.2.1 Avatars Influence Stereotype Threat. In online communication environments, avatars are the fundamental vehicle of users’ social identities. Drawing from a well-respected definition that applies broadly across media modalities [119], we define avatars here as mediated representations of humans used to interact with others, objects, or environments in real time. Avatars are mediated—not the users themselves, but representations of the users—and hence afford a potential for anonymity. In other words, even if an avatar appears (or sounds) very lifelike and realistic, it may not represent the users’ identity characteristics accurately, so inferences drawn about the user from the avatar’s depiction may be inaccurate. However, despite this rationale, people still treat avatars’ artificially
anthropomorphic characteristics as representative of actual human characteristics and stereotype them accordingly [85, 134]. Such misattribution of human traits to non-human entities likely occurs because people tend to obliviously respond to social cues—including those exhibited by machines—following the social norms developed through human-human interaction [55, 117], a phenomenon referred to as the media equation [135]. Further, when people use avatars, they tend to adopt the avatars’ identity characteristics into their self-perception and then behaviorally conform to associated stereotypes [132, 192]. This phenomenon, referred to as the Proteus effect, helps explain why people are also susceptible to stereotype threat induced with respect to their avatars’ identity characteristics. For example, one study found that both men and women participants who used a female avatar in a threat condition (i.e., in a group with two ostensible male avatars) performed worse on a math task than participants who used a male avatar [96]. Another study found that using a female avatar in VR harmed math task performance and math confidence, while using a male avatar buffered against these stereotype threat effects [129]. And another study with women participants found that those who customized and used a female avatar (compared to a male avatar) in a sword-fighting Wii game performed worse on a competitive math task after the game, but only if perceived avatar embodiment was low [134]. These previous studies of avatars and stereotype threat focused on visual avatar characteristics (i.e., the avatar’s appearance on the screen). No studies of which we are aware have examined the role of non-visual avatars in stereotype threat effects. This is a major gap in the research because voices—and voice avatars—are a highly stereotyped aspect of social identity in online interactions.

2.3 Stereotyping Voice (Avatars)

Voice has a unique capacity to reflect social identity, acting as an “auditory face” that others evaluate [13] to ascertain the speaker’s gender [170, 175], age [149], ethnicity [175], personality [104], and even social status [88, 149]. Hence, voice can trigger “linguistic profiling” [64], racism, and sexism from others [28, 111]. Unlike with visual self-representations (i.e., avatars), most popular online communication platforms—including Zoom, Microsoft Teams, Cisco Webex, and Discord [17]—do not presently offer native tools for users to modify or customize the representation of their voices (i.e., voice avatars). Instead, such platforms simply detect audial information from the microphone input and replicate it through speaker outputs with little or no computational processing that affect how identity cues are presented [26, 79]. Such systems hinder privacy and agency in voice communication, which helps explain why women and minority groups are often reluctant to use their voices online to avoid negative attention and delegitimization of their abilities [39, 91, 110, 111]. This is especially problematic because voice is growing in prominence as a medium of communication not only in online groups, but also in human-computer interaction, such as when communicating with virtual assistants and intelligent agents [30, 34, 47, 48, 148, 190, 191, 193], dictating emails [151], navigating websites [1], and online learning [87, 113, 127].

Voice avatars present a novel opportunity to address this issue by allowing users to decouple identity characteristics presented to others through voices from users’ offline identity characteristics. Although some researchers have examined voice avatars or similar concepts related to voice modification or customization in computational settings (e.g., [82, 83, 121], research on voice avatars in online group communication is practically non-existent. As an early exploratory step in this direction, we focus on masking gender identity through voice avatars in a way that facilitates anonymity for the users, meaning their personal identity cannot be readily ascertained through the information shared online [182, 183]. Studies suggest that such anonymity can mitigate stereotype threat. For example, masking status cues during a choice-dilemma task led to more equitable group discussion between high and low status participants [49]. Similarly, masking gender cues during group discussions among executives led to more equitable decision making between men and
women participants [160]. Research on visual avatars suggests the same pattern. One study found that African-American participants using race-anonymized avatars persisted significantly longer in a cognitive group task than those using race-revealing avatars [93]. Another study found that the African American participants performed worse on a cognitive task when competing (instead of cooperating) with race-revealing avatars [94].

2.4 Hypotheses

Our hypotheses are based on an assumed context of online interaction in which an individual—the potential target of stereotype threat—interacts with group members potentially with voice avatars designed to anonymize social identity. In order to test our general expectation that voice avatars designed to anonymize users should mitigate stereotype threat, we pose a series of hypotheses (all included in our preregistration) about the mechanisms and the stereotype threat of performance outcomes. Regarding mechanisms, we draw from Self-Determination Theory (SDT) [140] to hypothesize that stereotype threat will have a negative effect on competence (e.g., from worse performance [162]), autonomy (e.g., as a correlate of lower motivation [45, 67]), and relatedness (e.g., from a lower sense of belonging [174]). Just as racial anonymization has been found to mitigate stereotype threat effects [93], we hypothesize that gender anonymization may similarly positively affect competence, autonomy, and relatedness. Regarding outcomes, we focus on proportion of speaking contribution, an outcome studied in previous research on group dynamics and stereotypes [49, 160], as well as on facets related to success on the group task.

Additionally, studies have shown that the perception of being in the minority amplifies stereotype threat effects, e.g., [94]. In this context, when all group members are not using gender anonymized voices, this should reinforce identity salience and the perception of being in the minority, augmenting stereotype threat effects. Therefore, we hypothesize that there will be an interaction effect such that when all group members are not using gender anonymized voices, stereotype threat effects will be greater than the summed stereotype threat effects of individual group subsets not using gender anonymized voices.

**Competence**

H1.1: Participant voice masking will lead to higher competence.  
H2.1: Groupmate voice masking will lead to higher competence.  
H3.1: Interaction effect: Participant unmasking will see a greater reduction in competence with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.

**Autonomy**

H1.2: Participant voice masking will lead to higher autonomy.  
H2.2: Groupmate voice masking will lead to higher autonomy.  
H3.2: Interaction effect: Participant unmasking will see a greater reduction in autonomy with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.

**Relatedness**

H1.3: Participant voice masking will lead to higher relatedness.  
H2.3: Groupmate voice masking will lead to higher relatedness.  
H3.3: Interaction effect: Participant unmasking will see a greater reduction in relatedness with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.
More generally, we hypothesize that women will be less susceptible to stereotype threat effects when communicating through gender-anonymous voices in computing group work. Therefore, we hypothesize that gender-anonymous voices will increase duration of speaking (stereotype threat causes anxiety [162] which can hinder communication fluency [108] and decrease speech duration [60, 92, 188]), speed of response (stereotype threat leads to communication tentativeness [108, 124]), correctness of answers (stereotype threat reduces performance [162]), and respect from others (stereotype threat reduces sense of belonging [115, 174]). Because stereotype threat effects result from a complex process involving multiple mechanisms [146], many affective, cognitive, and motivational factors have been hypothesized to mediate stereotype threat effects [131]. We hypothesize that competence, autonomy, and relatedness will mediate stereotype threat effects because these factors are predictors of variables known to be influenced by stereotype threat (e.g., competence and autonomy are theorized to be positive predictors of motivation [105], relatedness influences sense of belonging [112]) [174]. Moreover, competence, autonomy, and relatedness are theorized to be universal needs that motivate behavior and may hence affect our outcomes [139]. (See Figure 3.)

Duration of Speaking
H4.1: Participant voice masking will lead to higher duration of speaking.
H4.2: Groupmate voice masking will lead to higher duration of speaking.
H4.3: Interaction effect: Participant unmasking will see a greater reduction in duration of speaking with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.
H4.4: Competence will mediate H4.1.
H4.5: Competence will mediate H4.2.
H4.6: Autonomy will mediate H4.1.
H4.7: Autonomy will mediate H4.2.
H4.8: Relatedness will mediate H4.1.
H4.9: Relatedness will mediate H4.2.

Speed in Responding
H5.1: Participant voice masking will lead to higher speed in responding.
H5.2: Groupmate voice masking will lead to higher speed in responding.
H5.3: Interaction effect: Participant unmasking will see a greater reduction in speed in responding with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.
H5.4: Competence will mediate H5.1.
H5.5: Competence will mediate H5.2.
H5.6: Autonomy will mediate H5.1.
H5.7: Autonomy will mediate H5.2.
H5.8: Relatedness will mediate H5.1.
H5.9: Relatedness will mediate H5.2.

Correctness of Responses
H6.1: Participant voice masking will lead to higher correctness of responses.

---

This assumes a gender-anonymous visual representation across all conditions, which was used in this study. This is further discussed in Limitations.

Note that we also test for actual number of words spoken in addition to duration. However, this analysis was not contained in our preregistration and hence is not specified with formal hypotheses. This analysis was instead performed as a follow-up to test whether results with respect to speech duration would translate to the actual number of words spoken. See Section 6.3.4 for additional rationale and methodology.
H6.2: Groupmate voice masking will lead to higher correctness of responses.
H6.3: Interaction effect: Participant unmasking will see a greater reduction in correctness of responses with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.
H6.4: Competence will mediate H6.1.
H6.5: Competence will mediate H6.2.
H6.6: Autonomy will mediate H6.1.
H6.7: Autonomy will mediate H6.2.
H6.8: Relatedness will mediate H6.1.
H6.9: Relatedness will mediate H6.2.

Stereotype Threat Scores
H7.1: Participant voice masking will lead to lower stereotype threat scores.
H7.2: Groupmate voice masking will lead to lower stereotype threat scores.
H7.3: Interaction effect: Participant unmasking will see a greater increase in stereotype threat scores with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.
H7.4: Competence will mediate H7.1.
H7.5: Competence will mediate H7.2.
H7.6: Autonomy will mediate H7.1.
H7.7: Autonomy will mediate H7.2.
H7.8: Relatedness will mediate H7.1.
H7.9: Relatedness will mediate H7.2.

Autonomous Respect Scores
H8.1: Participant voice masking will lead to higher autonomous respect scores.
H8.2: Groupmate voice masking will lead to higher autonomous respect scores.
H8.3: Interaction effect: Participant unmasking will see a greater reduction in autonomous respect scores with respect to participant masking when faux participants are unmasked compared to when faux participants are masked.
H8.4: Competence will mediate H8.1.
H8.5: Competence will mediate H8.2.
H8.6: Autonomy will mediate H8.1.
H8.7: Autonomy will mediate H8.2.
H8.8: Relatedness will mediate H8.1.
H8.9: Relatedness will mediate H8.2.

3 VOICE AVATAR CREATION SOFTWARE
We wanted to develop voice avatar creation software. To do so, as a first step, we sought to leverage voice-changing software that could facilitate the creation of a gender-anonymous voice. Therefore, we (1) reviewed existing voice-changing software; (2) determined that none of the existing voice-changing software were suitable; and (3) developed and validated our own custom voice changer. Details of this process can be found in Supplementary Materials.

4 ONLINE MEETING PLATFORM
We developed an online meeting platform compatible in any modern browser. We built our own platform because we wanted fine-grained control over the application—e.g., embedding our voice avatar creation software and recording meeting analytics. See Figures 1 and 2. Details of the platform and the development process can be found in Supplementary Materials.
5 METHODS

5.1 Study Preregistration

Our study was preregistered on the Open Science Framework (OSF). Hypotheses, experiment design, data collection, sample size, measures, and analyses are contained in our preregistration.

5.2 Conditions

The study used a 2 x 2 factorial design. Participants worked in groups of 3 (participant + 2 male confederates). All participants self-identified as female. Participants were led to believe that the two other group members were real participants. In reality, the two other group members were prerecorded male confederates. Participants were not told explicitly what the gender of the other group members was. We manipulated participant voice (gender-unmasked vs. gender-masked) and the two group members’ voices (gender-unmasked vs. gender-masked). Voice anonymity conditions were as follows:

- **All unmasked**: None of the group members were gender anonymous.
- **Male confederates masked**: Only the two male group members were gender anonymous.
- **Female participant masked**: Only the female participant was gender anonymous.
- **All masked**: All group members were gender anonymous.

Additional details on the specific manner in which stereotype threat was induced, how the male confederate voices were created, and the visual avatar used by meeting participants can be found in Supplementary Materials.
5.3 Measures

Full details and justification for each measure can be found in Supplementary Materials.

5.3.1 Competence. We measured competence using the perceived competence subscale of the Intrinsic Motivation Inventory (IMI) [105].

5.3.2 Autonomy. We measured autonomy using the perceived choice subscale of the IMI [105].

5.3.3 Relatedness. We measured relatedness using the relatedness subscale of the IMI [105].

5.3.4 Stereotype Threat Measure: Duration of Speaking. Duration of speaking was measured by manually inspecting the duration of speech in each participant voice recording, omitting blank silence at the beginning and end.

5.3.5 Stereotype Threat Measure: Speed in Responding. Speed in responding was automatically recorded by the meeting platform. This was the duration of time between when the participant had the ability to begin recording their answer and when the participant actually began recording their answer.

5.3.6 Stereotype Threat Measure: Correctness of Responses. Correctness for each computing problem was determined through audio recordings of participants’ responses. Details on how correctness was assessed can be found in Supplementary Materials.

5.3.7 Stereotype Threat Measure: Stereotype Threat Survey. We used the “negative stereotype concerns” survey from a previous study on stereotypes in computing [103].
5.3.8 Stereotype Threat Measure: Respect From Others. We measured “respect from others” [7, 178]. Additional details can be found in Supplementary Materials.

5.4 Sample Size Determination
We calculated a priori sample size. For 2x2 ANOVAs, we used G*Power 3.1 to conduct a power analysis using an effect size of small (0.1), \( \alpha = 0.05 \), and 95% power. This power analysis found that a sample size of \( N=1302 \) would be necessary.

For the mediation analyses, we use Monte Carlo Power Analysis [147]. We used a model with three parallel mediators, 95% power, 1000 replications, 20000 Monte Carlo draws per rep, and a 95% confidence level. Correlations between variables are set to 0.2 based on the literature available to us, e.g., a validation of the IMI found moderate correlations between subscales [105]. The power analysis found that a sample size of \( N=808 \) would be necessary.

We take the upper bound across both power analyses of \( N=1302 \).

5.5 Participants
1362 participants were recruited from Prolific,\(^4\) with 17 participants removed during data screening (see data screening in Supplementary Materials for rationale). Participants were paid $11 USD per hour. There was no limitation on geographic location. Recruitment details, data screening, participants’ demographics, and participants’ prior computing experience can be found in Supplementary Materials.

5.6 Design
We used a between-subjects factorial design. There were four possible conditions, and each participant was randomly assigned to one condition. Across conditions, there were approximately equal numbers of participants (\( M=336.3, \ SD=9.3 \)).

5.7 Procedure
5.7.1 Overview. Here, we provide a brief overview of the procedure. Additional details including verbatim participant instructions can be found in Supplementary Materials. Participants first opened the online meeting link in their browser. After an audio test and microphone test, participants chose their visual avatar (see Section 5.2). After instructions, the participant was paired with two other “participants” (male confederates). Participants were told they would act as judges, while the other two participants were brainstormers.

During the meeting, the group members were presented each of 5 computing problems one at a time. We took the problems from the “CS4FN Puzzle Book”\(^5\). Specifically, we chose puzzles #7, #10, #13, #15, and #17. Each puzzle was presented one at a time and shown on the screen (see Figure 2). See section Brainstormer Dialogue Snippet and Example Participant Solutions for the specific dialogue for each problem. Each problem was discussed by the two brainstormers, after which the participant was prompted to record an audio clip of their answer. During recording, participants could hear the output from the voice avatar creation software after a delay in order to reinforce the manipulation.

Afterward, participants filled out a post-survey and were debriefed. Full details of the procedure and justification for study choices can be found in Supplementary Materials.

---

\(^4\) We intentionally recruited a slightly larger number of participants than required (\( N=1302 \)) in case of data being removed during screening.

\(^5\) Problems: https://cs4fndownloads.files.wordpress.com/2016/02/cs4fnpuzzlebook11.pdf

Solutions: https://cs4fndownloads.files.wordpress.com/2016/02/cs4fnpuzzles1solutions.pdf
5.7.2 Brainstormer Dialogue Snippet and Example Participant Solutions. A complete listing of brainstormer dialogue for all problems and example participant solutions for all problems can be found in Supplementary Materials.

5.8 Analysis

We used SPSS 23 and the PROCESS macro [69] for data analysis. Factorial 2 x 2 ANOVAs were used to study the effects of participant masking and group member masking on IMI perceived competence (H1), IMI perceived choice (H2), and IMI relatedness (H3). We then performed a parallel mediation analysis with condition as a categorical variable using indicator coding (X), competence (M1), autonomy (M2), relatedness (M3), and duration of speaking (Y) (H4). PROCESS model 4 was used [69]. The parallel mediation was repeated using the different outcomes of interest (Y); speed in responding (H5); correctness of responses (H6); stereotype threat scores (H7); and autonomous respect (H8). Interaction effects for H4–H8 were tested separately using 2 x 2 ANOVAs.

6 RESULTS

6.1 Manipulation Check

The manipulation check consisted of six questions. All manipulation check questions validated the effectiveness of our manipulation. Detailed results of the manipulation check can be found in Supplementary Materials.

<table>
<thead>
<tr>
<th></th>
<th>IMI Competence</th>
<th>IMI Choice</th>
<th>IMI Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participant Unmasked</strong></td>
<td><strong>M</strong></td>
<td><strong>SD</strong></td>
<td><strong>M</strong></td>
</tr>
<tr>
<td>Group Members Unmasked</td>
<td>4.59</td>
<td>1.43</td>
<td>5.94</td>
</tr>
<tr>
<td>Group Members Masked</td>
<td>4.40</td>
<td>1.48</td>
<td>6.05</td>
</tr>
<tr>
<td><strong>Participant Masked</strong></td>
<td><strong>M</strong></td>
<td><strong>SD</strong></td>
<td><strong>M</strong></td>
</tr>
<tr>
<td>Group Members Unmasked</td>
<td>4.64</td>
<td>1.47</td>
<td>6.02</td>
</tr>
<tr>
<td>Group Members Masked</td>
<td>4.55</td>
<td>1.45</td>
<td>5.98</td>
</tr>
</tbody>
</table>

Table 1. Descriptives for IMI competence (H1), IMI choice (H2), and IMI relatedness (H3).

6.2 H1–H3: Effect of Manipulation on Competence, Autonomy, Relatedness

From Tables 1 and 2, factorial 2 x 2 ANOVAs (participant masking x group member masking) found no main effect of participant masking, no main effect of group member masking, and no interaction effect between participant masking and group member masking on competence (H1.1, H1.2, and H1.3 not supported), autonomy (H2.1, H2.2, and H2.3 not supported), and relatedness (H3.1, H3.2, and H3.3 not supported). Therefore, our manipulation did not have a statistically significant effect on competence, autonomy, or relatedness.

6 ANOVAs are considered robust to non-normality, especially at larger sample sizes [18].

7 This means that the first condition (“All unmasked”) will be used as the reference category against which all other conditions will be compared.

8 As specified in our preregistration, these ANOVAs test only for interaction effects and not main effects. The rationale for this is that main effects are already being tested in our mediation analysis, however interaction effects are not.
### Table 2. Results for effects of participant masking and group member masking on IMI competence (H1), IMI choice (H2), and IMI relatedness (H3).

<table>
<thead>
<tr>
<th></th>
<th>IMI Competence</th>
<th>IMI Choice</th>
<th>IMI Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main Effect Participant Masking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>1.801</td>
<td>0.025</td>
<td>0.15</td>
</tr>
<tr>
<td>$p$</td>
<td>0.180</td>
<td>0.874</td>
<td>0.903</td>
</tr>
<tr>
<td>$\eta^2_p$</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Main Effect Group Member Masking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>2.867</td>
<td>0.253</td>
<td>0.000</td>
</tr>
<tr>
<td>$p$</td>
<td>0.091</td>
<td>0.615</td>
<td>0.988</td>
</tr>
<tr>
<td>$\eta^2_p$</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Interaction Effect</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>0.492</td>
<td>1.429</td>
<td>0.087</td>
</tr>
<tr>
<td>$p$</td>
<td>0.483</td>
<td>0.232</td>
<td>0.768</td>
</tr>
<tr>
<td>$\eta^2_p$</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Participant Masking $df=1$, Group Member Masking $df=1$, Interaction $df=1$, Error $df=1341$

### 6.3 H4–H8: Analyses of Direct and Mediated Effect

#### 6.3.1 Assumption Checks. Assumption checks for mediation analyses can be found in Supplementary Materials.

![Mediation Model](image)

**Fig. 3.** Mediation model being tested for H4 through H8.

#### 6.3.2 Hypothesis Tests. We tested the mediation model in Figure 3. We used a 95% bias-corrected confidence interval based on 10,000 bootstrap samples. From Table 3, competence, autonomy, and relatedness do not mediate the effects of voice condition on any outcome variables (H4.4–H4.9, H5.4–H5.9, H6.4–H6.9, H7.4–H7.9, and H8.4–H8.9 not supported). There were direct effects of voice condition on duration of speaking ($c'$) for the comparisons “All unmasked” vs. “Male confederates masked” and “All unmasked” vs. “All masked” (H4.1 supported, H4.2 supported). There were direct effects of voice condition on correctness of responses ($c'$) for the conditions “All unmasked” vs. “Male confederates masked” and “All unmasked” vs. “All masked” (H6.1 supported, H6.2 supported). There were no direct effects ($c'$) of voice condition on speed in responding (H5.1 and H5.2 not supported), stereotype threat scores (H7.1 and H7.2 not supported), and autonomous respect scores (H8.1 and H8.2 not supported). Therefore, we conclude that the voice conditions “Male confederates masked” and “All masked” had a significantly positive effect on duration of speaking compared...
For each outcome variable, the first line presents results and (4) "All masked." We use a multicategorical X with 4 conditions: (1) "All unmasked," (2) "Male confederates masked," (3) "Female participant masked," and (4) "All masked." All presented effects are unstandardized. Significant results are bold.

### Table 3. Mediation results with voice condition (X), competence (M1), autonomy (M2), relatedness (M3), and each outcome variable (Y).

Regression coefficients aX → M1, bM1 → Y, c' (direct X → Y), c (total X → Y), and aXbM2. We use a multicategorical X with 4 conditions: (1) "All unmasked," (2) "Male confederates masked," (3) "Female participant masked," and (4) "All masked." For each outcome variable, the first line presents results for the comparison between (1) and (2), the second line (1) and (3), and the third line (1) and (4). These specific comparisons are done (and not other comparisons) because we use indicator coding, in which all conditions are compared to the reference condition (in this case, "All unmasked" is the reference condition). See [70] for more details on this approach. All presented effects are unstandardized. Significant results are bold.

<table>
<thead>
<tr>
<th></th>
<th>a1</th>
<th>b1</th>
<th>a1b1</th>
<th>a2</th>
<th>b2</th>
<th>a2b2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Competence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of Speaking</td>
<td>0.103</td>
<td>1.064</td>
<td>0.110</td>
<td>0.103</td>
<td>1.064</td>
<td>0.110</td>
</tr>
<tr>
<td>Speed in Responding</td>
<td>0.051</td>
<td>2.320</td>
<td>0.118</td>
<td>0.082</td>
<td>7.165</td>
<td>0.006</td>
</tr>
<tr>
<td>Stereotype Threat Scores</td>
<td>0.051</td>
<td>0.975</td>
<td>0.050</td>
<td>0.082</td>
<td>7.165</td>
<td>0.005</td>
</tr>
<tr>
<td>Autonomy Respect Scores</td>
<td>0.051</td>
<td>0.975</td>
<td>0.027</td>
<td>0.082</td>
<td>2.277</td>
<td>0.011</td>
</tr>
<tr>
<td>Correctness of Responses</td>
<td>0.103</td>
<td>7.165</td>
<td>0.006</td>
<td>0.103</td>
<td>0.277</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>Autonomy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speed in Responding</td>
<td>0.103</td>
<td>0.054</td>
<td>0.006</td>
<td>0.082</td>
<td>0.054</td>
<td>0.004</td>
</tr>
<tr>
<td>Correctness of Responses</td>
<td>0.082</td>
<td>0.277</td>
<td>0.023</td>
<td>0.082</td>
<td>0.277</td>
<td>0.024</td>
</tr>
<tr>
<td>Autonomous Respect Scores</td>
<td>0.082</td>
<td>0.131</td>
<td>0.011</td>
<td>0.082</td>
<td>0.131</td>
<td>0.005</td>
</tr>
<tr>
<td>Stereotype Threat Scores</td>
<td>0.082</td>
<td>0.131</td>
<td>0.005</td>
<td>0.082</td>
<td>0.131</td>
<td>0.005</td>
</tr>
<tr>
<td>Relatedness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of Speaking</td>
<td>0.051</td>
<td>0.511</td>
<td>0.010</td>
<td>0.051</td>
<td>0.511</td>
<td>0.010</td>
</tr>
<tr>
<td>Speed in Responding</td>
<td>0.100</td>
<td>0.051</td>
<td>0.004</td>
<td>0.051</td>
<td>0.151</td>
<td>0.008</td>
</tr>
<tr>
<td>Correctness of Responses</td>
<td>0.050</td>
<td>0.151</td>
<td>0.000</td>
<td>0.050</td>
<td>0.151</td>
<td>0.000</td>
</tr>
<tr>
<td>Stereotype Threat Scores</td>
<td>0.050</td>
<td>0.151</td>
<td>0.000</td>
<td>0.050</td>
<td>0.151</td>
<td>0.000</td>
</tr>
</tbody>
</table>

* significant at p < 0.05; ** significant at p < 0.01; *** significant at p < 0.005; significant aXbM2 based on 95% CI.
to “All unmasked.” The voice conditions “Male confederates masked” and “All masked” also had a significantly positive effect on correctness of responses compared to “All unmasked.” Descriptives for each variable are in Table 4.

### 6.3.3 Interaction Effects

Interaction effects for H4–H8 were tested separately using factorial 2 x 2 ANOVAs (participant masking x group member masking).

- Interaction effects were not significant for speed in responding ($F_{[1, 1341]} = 0.064$, $p = 0.801$, $\eta^2_p = .000$) (H4.3 not supported), duration of speaking ($F_{[1, 1341]} = 0.102$, $p = 0.750$, $\eta^2_p = .000$) (H5.3 not supported), correctness of responses ($F_{[1, 1341]} = 0.104$, $p = 0.748$, $\eta^2_p = .000$) (H6.3 not supported), stereotype threat scores ($F_{[1, 1341]} = 0.023$, $p = 0.880$, $\eta^2_p = .000$) (H7.3 not supported), and autonomous respect scores ($F_{[1, 1341]} = 0.223$, $p = 0.637$, $\eta^2_p = .000$) (H8.3 not supported).

### 6.3.4 Analysis Extension: Speech Word Count

Stereotype threat has been shown to increase anxiety, which may influence duration of speech without actually influencing the number of words spoken—e.g., nervously speaking quickly [159]. Therefore, we performed an additional analysis on actual number of words spoken by each participant during the meeting. This was a small oversight of our preregistration—this is the only analysis not part of the preregistration. The purpose of this analysis is to confirm whether the voice conditions affect not only duration of speaking, but also the actual number of words spoken.

First, a researcher blind to conditions manually listened and recorded the number of words spoken in each participant audio recording. Filler sounds (e.g., “ah,” “uh,” “um”) were excluded [20]. The researcher made two separate passes over all recordings to ensure an accurate word count. Next, the average number of words spoken per problem was calculated for each participant. This average number of words was used as the outcome variable (Y). We again tested the same mediation model in Figure 3 with identical parameters to Section 6.3.2. From Table 5, we can see that competence, autonomy, and relatedness did not mediate the effects of voice condition on number of words spoken. However, there was a direct effect of voice condition on number of words spoken ($c'$) for the comparison “All unmasked” vs. “All masked.” Therefore, we conclude that the voice condition “All masked” had a significantly positive effect on number of words spoken. Similar to H4–H8, an interaction effect was tested separately using a factorial 2 x 2 ANOVA (participant masking x group member masking). The interaction effect was not significant ($F_{[1, 1341]} = 0.016$, $p = 0.899$, $\eta^2_p = .000$). Descriptives can be found in Table 4.

---

**Table 4. Descriptives for outcomes in H4 through H8 and number of words spoken. Duration of Speaking, Speed in Responding, and Number of Words Spoken are averages per problem.**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>PARTICIPANT UNMASKED</th>
<th>PARTICIPANT MASKED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GROUP MEMBERS</td>
<td>GROUP MEMBERS</td>
</tr>
<tr>
<td></td>
<td>UNMASKED</td>
<td>MASKED</td>
</tr>
<tr>
<td>Duration of Speaking (sec.)</td>
<td>36.52</td>
<td>40.74</td>
</tr>
<tr>
<td>Speed in Responding (sec.)</td>
<td>82.59</td>
<td>149.82</td>
</tr>
<tr>
<td>Correctness of Responses (0–15)</td>
<td>6.48</td>
<td>3.69</td>
</tr>
<tr>
<td>Stereotype Threat Scores (1–7)</td>
<td>3.16</td>
<td>3.07</td>
</tr>
<tr>
<td>Autonomous Respect Scores (1–7)</td>
<td>3.82</td>
<td>3.95</td>
</tr>
<tr>
<td>Number of Words Spoken</td>
<td>30.89</td>
<td>32.39</td>
</tr>
</tbody>
</table>

---

9See footnote 8.
were significant when female participants alone used a gender-masked voice avatar. Group members word count) and correctness (i.e., higher scores on problems). However, none of these outcomes were significant when female participants alone used a gender-masked voice avatar. Group members using a gender-masked voice avatar resulted in higher duration of speaking and higher problem scores, but not higher spoken word count. Significant positive effects on all three outcomes (speaking duration, spoken word count, problem scores) only occurred when everyone used gender-masked voice avatars. Therefore, while there may be benefits when a subset of meeting participants are

7 DISCUSSION

Masking visual identity cues in online avatars (e.g., race for African-American participants in anagram-solving [93], gender for female participants in mathematics [94]) positively influences problem-solving performance in stereotype-relevant contexts [93, 94]. However, it is not known if masking of identity cues in online avatars (e.g., race for African-American participants in anagram-solving [93], gender for female participants in mathematics [94]) positively influences problem-solving performance in stereotype-relevant contexts [93, 94]. However, it is not known if

Table 5. Mediation results with voice condition (X), competence (M1), autonomy (M2), relatedness (M3), and number of words spoken (Y). Regression coefficients $a_x$ (X $\rightarrow$ M1), $b_x$ (M1 $\rightarrow$ M2), $c'$ (direct X $\rightarrow$ Y), $c$ (total X $\rightarrow$ Y), and $a_xb_x$. We use a multicategorical X with 4 conditions: (1) “All unmasked,” (2) “Male confederates masked,” (3) “Female participant masked,” and (4) “All masked.” For each outcome variable, the first line presents results for the comparison between (1) and (2), the second line (1) and (3), and the third line (1) and (4). These specific comparisons are done (and not other comparisons) because we use indicator coding, in which all conditions are compared to the reference condition (in this case, “All unmasked” is the reference condition). See [70] for more details on this approach. All presented effects are unstandardized. Significant results are bold.

<table>
<thead>
<tr>
<th>COMPETENCE</th>
<th>AUTONOMY</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$b_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NUMBER OF WORDS SPOKEN</th>
<th>DIRECT EFFECT</th>
<th>TOTAL EFFECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$b$</td>
<td>$a_ib_a$</td>
</tr>
<tr>
<td>$c'$</td>
<td>$c$</td>
<td></td>
</tr>
</tbody>
</table>

*significant at $p < 0.05$; **significant at $p < 0.01$; ***significant at $p < 0.005$; significant $a_xb_x$ based on 95% CI.
using a gender-masked voice avatar, our results suggest that everyone in the meeting should use
gender-masked voice avatars to maximize outcomes.\textsuperscript{10}

7.1 The Benefits of Greater Participation and Performance to Diversity
Male-dominated organizations in computing fields can be discouraging for women (e.g., startups
with fraternity-like cultures and sexism) \cite{166}. Repeatedly experiencing stereotype threat can lead
to a negative cycle of diminished performance, confidence, and interest \cite{59}. Ultimately, this leads
to disidentification from a career \cite{59}. However, our study suggests that voice gender anonymity
can have a positive effect on both participation and performance in solving computing problems.
Therefore, voice avatars may be one path forward toward improving diversity.

7.2 Applications of Identity-Masked Voice Avatars Online
The results of this study suggest that identity-masked voice avatars can mitigate stereotype threat
in group-based problem solving. When all group members used an identity-masked voice avatar,
female participants participated more actively in the online collaborative activity. We view identity-
masked voice avatars as applicable for institutions and corporations where meetings between
employees are increasingly virtual \cite{128} and where stereotype threat may be present in group
decision-making processes wherein stakeholders differ in status \cite{49, 171}. In these instances,
identity-masked voice avatars (coupled with visual anonymity) may empower more individuals to
voice their opinions and share critical information. As a consequence, decision-making processes
can be made more robust and foster more equitable discussions.

In addition to applications in online collaborative work and employee meetings, identity-masked
voice avatars may also have significant benefits in online learning environments. Previous research
has demonstrated that gender bias is rampant in these environments, manifesting in lower peer
ratings of female students’ work \cite{23, 114}, lower instructor grades of female students’ work \cite{74}, and
lower student evaluations of female instructors \cite{4, 101}. Gender anonymization within these online
learning environments, including identity-masked voice avatars, may help mitigate long-standing
gender biases. Within the specific context of computing, the underrepresentation of women may
be influenced by a perceived lack of similarity with those in the field \cite{33}. Even subtle gender
cues, such as objects in a classroom, can affect female participation \cite{32}. Given that voice carries
rich identity cues, identity-masked voice avatars present a promising avenue for exploration in
addressing gender inequities in computing.

7.3 Anonymous Voice Avatars for Everyone in the Group Maximizes Benefits
Research has shown that viewing oneself as being in the minority amplifies stereotype threat effects
\cite{95}. Within this context, our results suggest that everyone should use gender anonymous voice
avatars to maximize outcomes. This minimizes the potential sources of stereotype threat. There is,
however, a broader ethical question of whether we \textit{should} use anonymous voice avatars.

7.4 Should We Use Anonymous Voice Avatars?
This study may suggest that people should be encouraged to use gender-anonymous avatars in
online learning platforms. However, the social implications of anonymizing gender online to combat
stereotype threat are fraught—this techno-solutionist approach does not address the underlying
cultural problem. Hence, we suggest that this study’s results be interpreted as support for a solution

\textsuperscript{10}The only statistically significant difference between “Male confederates masked” and “All masked” is spoken word count.
However, there is another reason to use voice avatars that mask everyone. In a real virtual group, having only group
members’ voices gender-masked (from every user’s perspective) would create asymmetries in what each user hears. This
could be confusing and presents a usability issue.
that utilizes avatars (both visual and audial) to deemphasize gender cues and instead encourage users to express other aspects of their identities that are more relevant to the context of interaction.

7.5 Anonymous Voice Avatars Did Not Affect All Outcomes

Voice gender anonymity did not have a significant impact on competence, autonomy, relatedness, speed in responding, perceived stereotype threat scores, and autonomous respect scores.

With the exception of speed in responding, these were all self-reported subjective outcomes. Interestingly, even though we see an actual difference in levels of competence between conditions (i.e., scores on computing problems), we do not see a corresponding difference in self-reported competence. Stereotype threat, through increased rumination, arousal, and efforts to self-regulate, causes cognitive depletion [10]. This cognitive depletion from stereotype threat has been shown to make people poor judges of their own competence [173]. As such, one potential explanation is that stereotype threat is undermining participants’ assessments. Research into other measures used in our study (i.e., autonomy, relatedness, autonomous respect scores) may shed light on whether the accuracy of other self-reported measures might be inaccurate in some cases. Since SDT was used as a theoretical lens to hypothesize that stereotype threat would have a negative effect on competence, autonomy, and relatedness, future research could utilize alternative measurement approaches for each of these facets to further elucidate the effects of anonymous voice avatars—especially for autonomy and relatedness, for which there were no non-subjective measures in the present study.

Because subjective stereotype threat scores did not differ significantly across conditions, one might argue that stereotype threat was not the cause of differences between conditions. However, given the highly controlled nature of our study and validation of the voice avatar creation software, we believe this is unlikely. Variation between conditions was minimized except for the presence or absence of stereotypical gender cues in the voice-changing output. This was further confirmed through manipulation checks. We believe that a more likely cause is that stereotype threat was a subconscious factor that influenced participants’ performance and participation but was not salient enough to manifest in the self-reported scale. The vast majority of previous research on stereotype threat has focused on actual performance effects [162] rather than self-reported measures. Furthermore, previous research has shown that with regards to individuals under stereotype threat conditions, there is a significant divergence between self-reports regarding stereotype threat outcomes and objective stereotype threat outcomes [21]. For example, in one study stereotype-threatened individuals demonstrated significantly more anxiety, but this did not manifest as significantly more self-reported anxiety [21]. As such, divergences between objective and self-reported stereotype threat measures is supported by previous studies.

Regarding speed in responding, our hypothesis that voice gender anonymity will lead to faster response times may be incorrect. Although we expected participants who are less anxious to also be less tentative (i.e., less hesitant in communicating), it is also possible that such participants would have more confidence and therefore take their time to think carefully about their responses before answering. Study design could have also played a role in this result: Because participant responses were recorded and participants were not engaged in a real-time conversation, this could have encouraged participants to take more time to think about their responses. Here, we did not find any effect of our voice avatar manipulation on speed in responding. Nevertheless, our manipulation had a significant effect on women’s participation and performance on computing problems.

7.6 Voice Avatar Application Areas: Education, Entertainment, Work

There are several potential application areas of voice avatars we envision. In education, this might include teaching through virtual reality [5], online meeting platforms [43], and massive open online courses [179]. In entertainment, this might include online multiplayer games. For example,
anonymous voice avatars may help reduce toxic behavior between players [176, 180, 181]. In work, this might include virtual meetings between employees [128]. Although we are optimistic about the potential of voice avatars, there are a few caveats.

Firstly, our results on voice avatars may not generalize to all situations. For example, in a virtual organizational meeting in which the identity of all meeting participants is known in advance, it is not clear what level of effectiveness anonymous voice avatars would have.

Secondly, our meeting platform did not contain a video feed of participants’ faces. This absence allowed us to study voice-only avatars in isolation, an important first step. Voice-only avatars are highly applicable in audio calls (e.g., Zoom call-in, regular phone calls, Discord voice chat), social virtual reality, and online gaming. Here, we have shown that the effects of voice avatars alone are impactful. Nevertheless, video feeds are commonplace in many online meetings [8]. As such, an appropriate next research direction is to study the combination of visual and audial avatars, i.e., visual avatars that align with their audial avatar counterparts, e.g., identity-masked visual avatars.

8 LIMITATIONS
Despite the carefully controlled nature of our experiment, there are a few limitations that should be considered for future studies.

Participants were not engaged in live interaction with the other two group members (i.e., the two male confederates). However, according to the manipulation check, participants perceived the interaction as an online meeting. Future smaller-scale studies might consider live interaction.

According to our manipulation checks, participants in self-unmasked conditions still felt their voice was somewhat masked. As a result, the effects of self-masking might be understated in our study (note that our conditions were still successful at inducing higher or lower levels of perceived self-masking—see Manipulation Checks; furthermore, our validation study in Section 3 confirms that the unmasked voice avatar is of a discernible stereotypical gender).

Our study recruited only participants self-identifying as female. More research is needed on non-binary gender identities and anonymous voice avatars. Furthermore, investigating diverse group configurations (smaller vs. larger groups, variations in group demographics) would help researchers understand relevant contextual factors.

9 CONCLUSION
In a large-scale preregistered study, we investigated gender-anonymous voice avatars in online computing group work. Female participants worked with two ostensible male participants (actually recorded confederates). There were four voice gender anonymity conditions: (1) All unmasked, (2) Male confederates masked, (3) Female participant masked, and (4) All masked. When only male confederates used gender-anonymous voices (compared to all unmasked), female participants spoke for a longer period of time and scored higher on computing problems. When everyone used gender-anonymous voices (compared to all unmasked), female participants spoke for a longer period of time, spoke more words, and scored higher on computing problems. We did not find significant effects on subjective measures and one behavioral measure we deemed to poorly reflect stereotype threat. Our findings demonstrate that anonymous voice avatars are a potential avenue for supporting diversity in virtual group work. We discussed several potential application areas of anonymous voice avatars (e.g., virtual classrooms, online multiplayer games, business meetings). This study presents a first step in understanding the effects of anonymous voice avatars.

However, people have become increasingly comfortable with turning off their video cameras during online meetings [8]. This can be due to the stress of having to appear constantly attentive, wanting to multitask, and connectivity issues [6]. Furthermore, research has shown that having the video camera on during online meetings increases fatigue and reduces engagement and that such effects are disproportionately worse for women [154].
ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation under the award number IIS #2113991. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

SUPPLEMENTARY MATERIALS

Supplementary materials: https://osf.io/ejzcw/.
Preregistration: https://osf.io/zmyhr/.

REFERENCES


and Technology. 1–2.


Morgan Smith. 2022. Twitter, Reddit and 8 other companies offering permanent remote or hybrid work? And hiring right now. https://www.cnbc.com/2022/04/13/10-companies-that-switched-to-permanent-hybrid-or-remote-work-and-hiring-right-now.html


Highlights:

- We developed voice avatar creation software that transforms voice input in real time to mask stereotypical gender characteristics.
- We perform a study (N=1362) comparing four voice anonymity conditions in online computing group work in which female participants worked with two male confederates.
- When only male confederates used gender-anonymous voices, female participants spoke longer and scored higher on computing problems.
- When all group members used gender-anonymous voices, female participants spoke longer, spoke more words, and scored higher on computing problems.
Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: