Interacting with Computers, 2025, pp. 1–27

doi: 10.1093/iwc/iwaf033
Accepted for publication in Interacting with Computers
Paper

PAPER

Toward Accounting for the Effects of Gender Socialization in Quantitative Research in Human-Computer Interaction

Nazeli Hagen[1], Luke W. Miratrix[2] and Krzysztof Z. Gajos[3,*]

[1]Harvard College, Cambridge, MA, 02138, USA, [2]Harvard Graduate School of Education, Cambridge, MA, 02138, USA and [3]Harvard School of Engineering and Applied Sciences, 150 Western Ave., Allston, MA, 02134, USA
[*]Corresponding author. kgajos@seas.harvard.edu
FOR PUBLISHER ONLY Received on 3 July 2024; accepted on 16 April 2025

Abstract

In quantitative HCI research, gender is typically represented as a single categorical variable and data from non-binary participants are frequently excluded from analyses. Meanwhile, many scholars argue that gender is a complex, multidimensional construct, and that overly simplistic operationalization of gender risks that our theories will generalize poorly, have limited explanatory power, and will exclude experiences of individuals whose gender identities are not included in our analyses. In this work, we modeled gender as inclusive of multiple dimensions of gender socialization and we operationalized gender socialization through a subset of the items from the Conformity to Masculine Norms Inventory (CMNI). We replicated three studies of basic cognitive abilities (theory of mind, mental rotation, spatial working memory) that previously showed gender differences. For two of the studies, adding CMNI variables significantly and substantially improved the explanatory power of regression models. Also, in those studies, more than half of the effect of binary gender was mediated through the CMNI variables. These results suggest that gender socialization rather than categorical gender explain a substantial part of the individual differences on some cognitive tasks. Consequently, differences in task performance associated with gender categories may not be universal, i.e., they may not generalize to people from other cultures or eras where people are socialized into their gender roles differently. Instead, including multidimensional representations of gender may produce more accurate and more generalizable models. Given that our results also showed that CMNI might not model non-binary participants the same way as men and women, it remains an open question what specific instruments should be used to represent gender in quantitative analyses.

Key words: feminist HCI, sex difference, gender difference, gender socialization, masculinity, femininity

Key Messages

- Gender socialization is significantly and substantially associated with differences in performance on tasks that involve such low-level cognitive abilities as theory of mind and spatial memory.
- Measures of gender socialization can be used in addition to categorical gender to improve explanatory power of regression models of participant behavior.
- In some cases, measures of gender socialization can be used *instead of* categorical gender variable in regression models without loss of explanatory power.

1. Introduction

Gender is now understood to be a complex, multidimensional construct shaped substantially by socialization (Rode, 2011; Keyes et al., 2021; Stumpf et al., 2020; Carothers and Reis, 2013; Nielsen et al., 2021). A considerable body of research has demonstrated that many gender differences in behavior (ranging from low-level cognitive differences in information processing to high-level social behaviors) are attributable, at least in part, to differences in socialization (e.g., Eagly and Wood (2013); Nazareth et al. (2013)). This has important implications for quantitative research in which gender is used as either a predictor or a covariate when analyzing behavior: Because gender socialization is imperfectly correlated with gender categories, and because it differs across individuals, cultures, and history, *any* gender categorization is only an imperfect proxy for some other more precise and more relevant constructs (Jaroszewski

et al., 2018; Keyes et al., 2021). Consequently, quantitative research that models gender solely in terms of categories risks producing models that have low explanatory power and that may not generalize across cultures or across time.

Despite this, in quantitative Human-Computer Interaction (HCI) research, gender is typically captured as a single categorical variable (recent examples include Hwang and Won (2024); Umbach et al. (2024); Wang et al. (2024a); Mahmood and Huang (2024); Wang et al. (2024b)). Moreover, data from non-binary participants are frequently excluded from the analysis (e.g., Mahmood and Huang (2024); Wang et al. (2024b)). When interpreting results, some quantitative studies — particularly those relating to complex social phenomena — do address the possibility that gender differences in observed behavior may be impacted by societal factors such as gender role expectations (Huang et al., 2018; Kimbrough et al., 2013), or gender stereotypes (Wijenayake et al., 2019). However, HCI publications relating to gender differences in perception, motor performance, or basic cognitive tasks are less likely to suggest that anything other than innate differences between sexes might explain the differences (e.g., Borkin et al. (2013); Czerwinski et al. (2002); Huber and Gajos (2020); Jing et al. (2012); Ross et al. (2006); Tan et al. (2003); Yamauchi et al. (2015)).

A core motivation for our agenda, of which this manuscript is the first step, is to align quantitative methods in HCI with the conceptual advances regarding gender. In pursuit of this goal, we operationalized gender socialization through two existing instruments: a subset (7 items) of the Conformity to Masculine Norms Inventory (CMNI) (Mahalik et al., 2003) which captures individual differences in gender attitudes, and the Gender Gap Index (GGI) (World Economic Forum, 2020) which characterizes differences in gender attitudes at the country level. We then replicated three studies that had been shown previously to produce reliable differences in performance between men and women: the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001) (women perform better), the Mental Rotations Test (Peters et al., 1995) (men perform better), and the Spatial Working Memory Task (Duff and Hampson, 2001) (women perform better). We then investigated to what extent these differences could be explained by differences in gender socialization. Specifically, we conducted two types of statistical analyses for each of the 3 studies: First, we conducted mediation analyses to quantify what fraction of the information carried by the categorical gender variable could be explained by measures of individual gender socialization (CMNI items). Second, we conducted multiple regression analyses, successively adding variables to quantify how much additional explanatory power was provided by both individual (CMNI) and country-level (GGI) measures of gender socialization beyond what was captured by the categorical gender variable.

In all 3 studies (N=248,256, N=7,723, and N=7,730, respectively), the effect of binary gender on the outcome variables was significantly mediated through CMNI. Additionally, in all 3 studies, regression models that included CMNI in addition to binary gender and baseline covariates explained significantly more variability than models that did not include CMNI. Further, in all 3 studies, models that included GGI in addition to binary gender and covariates also significantly improved the explanatory power of the models. In two of the studies (Reading the Mind in the Eyes and Spatial Working Memory), the effects were not only statistically significant but also meaningful in magnitude: more than half of the total effects was mediated through CMNI and the relative benefits of including CMNI and GGI in regression analyses were as large or larger than the explanatory benefit of the binary gender variable. In the Mental Rotation Test, the results were statistically significant but small in magnitude.

In an additional analysis conducted in the context of our largest study with substantial international participation (the Reading the Mind in the Eyes Test), we found statistically significant associations between GGI and individual measures of CMNI—in most cases, greater gender parity (i.e., higher GGI scores for countries where our participants grew up and lived recently) was associated with less conformity with masculine norms (i.e., lower CMNI scores).

Across the 3 studies, in models that included only two gender categories we found no substantial interaction effects between gender and CMNI, indicating that people who identified with either of the binary genders were modeled similarly by CMNI. However, in the Reading the Mind in the Eyes study (our largest study and the only one with a substantial number of participants identifying as non-binary), when we modeled gender as having three categories (woman, man, non-binary), we found a significant and substantial interaction effect between gender and CMNI. Additional analyses showed that participants identifying as non-binary were modeled differently by CMNI than participants who identified with either of the binary genders.

At the high level, our results contribute additional evidence that gender socialization is significantly associated with differences in behavior on several fundamental cognitive tasks, including some (like spatial working memory) that may not appear to be affected by social factors. Our results also demonstrate limitations of CMNI and point to the need for further research. Specifically, with respect to the quantitative research in HCI, our results have the following key implications:

1. Gender socialization is significantly and substantially associated with differences in performance on tasks that involve such low-level cognitive abilities as theory of mind and spatial memory. Because gender socialization differs across cultures and across time, relying solely on gender categories to demonstrate gender effects risks that the results may not hold universally. Thus, to ensure validity and generalizability of empirical claims related to the impact of gender on behavior, quantitative HCI research needs to develop methods to account for gender socialization.
2. Given the results of the mediation analyses that show that more than half of the effect of binary gender is mediated through CMNI for some studies, measures of individual gender socialization could be used *instead of* categorical gender variables for some studies. However, CMNI appears to capture different underlying constructs for binary and non-binary individuals. Thus, to facilitate inclusion of all genders in statistical analyses, there is a need for a new instrument that more universally captures individual differences related to gender socialization.
3. Both individual (CMNI) and country-level (GGI) measures of gender socialization can be used in addition to categorical gender to improve explanatory power of regression models of participant behavior. Besides studies concerned directly with impacts of gender socialization on behavior, such measures may be of value as covariates to control for gender socialization differences in between-subjects comparisons.

2. Related Work

2.1. Gender as a Social Construct

Contemporary literature frequently distinguishes between sex and gender (Tannenbaum et al., 2019). Sex refers to biological attributes at birth which, for humans, includes 3 categories: male, female and intersex (although this last category—which recognizes individuals who are born with some characteristics of both male and female sexes (Fausto-Sterling, 2000)—is not yet universally included). Gender, in turn, "refers to the socially constructed roles, behaviours, expressions and identities of girls, women, boys, men, and gender diverse people" (Canadian Institutes of Health Research, 2023). This said, it is frequently acknowledged that sex and gender can interact in complex ways (Boerner et al., 2018; Van Anders et al., 2015; Steensma et al., 2013; Wood and Eagly, 2012). It is also the case that, as defined above, gender is a very broad concept. Thus, in different research projects gender is conceptualized differently, often responding to the specific needs of the project. Examples of specific variables considered include caregiver strain, work strain, social support (and 4 more) in (Nielsen et al., 2021) or personal income, number of hours per week spent doing housework, stress at home (and 4 more) in (Pelletier et al., 2016). Some researchers also posit that gender may include rapidly-changing constructs such as "changes in hormones, abilities, roles, and socially imposed expectations" (Boerner et al., 2018) and, thus, may need to be conceptualized as comprising both state and trait components (Boerner et al., 2018; Morgenroth and Ryan, 2021). Finally, it is also the case that in day-to-day research practice the distinction between sex and gender is not yet fully established. As observed by (Boerner et al., 2018), there are still numerous research articles that "use sex and gender interchangeably or use the word gender to refer to sex differences."

In our work, we conceptualized gender as a stable trait and we wanted to focus on aspects of gender that are distinct from sex. Specifically, we chose to focus on gender socialization. Thus, let us first summarize some foundational work that elucidated the concept of the social construct of gender and that developed links between gender socialization and behavior. Judith Butler, for example, defined gender as a "stylized repetition of acts" (Butler, 1988); instead of being something intrinsically defined and predetermined, socialized gender can be understood as a constructed identity that is performed and that exists only through its performance. There is no fundamental truth or predefined idea of what socialized gender is. Instead, it is produced through social interactions and is only defined through the acts performed as a representation of it. People are taught to perform these acts and are continuously exposed to similar acts being performed repeatedly around them, and these acts are what come to constitute the definition of how men and women behave.

Understanding the socialized aspects of gender as a set of acts repeatedly performed suggests then that gender socialization is inseparable from the social context in which it is embedded. This social context includes the social structures which "constrain certain enactments of gender and enable others" (Morawski, 1987). With the performance of gender comes a significant social power and a definitive hierarchy. To behave convincingly as a woman means that you are given the social power of a woman and affected by sexism and those who expect you to behave a certain way (as feminine, submissive, etc.). Therefore, gender socialization is not just about one individual performance but a group performance as well in how others treat you and which gender they see being performed. These repeated acts also come to represent skills and abilities, so it is often through these gendered performances and acts that people learn the skills associated with their gender role which is what often results in men and women being "good" at different skills and tasks.

Social learning (or socialization) theories of gender differences focus on the impact of gender norms and constructs that exist in someone's social environment and suggest the importance of observation, imitation, and reinforcement in internalizing ideas about gender roles and gender differences (Mischel, 1966). For example, within the home, parental figures (and even older siblings) often act as examples for children (mothers/sisters for girls, fathers/brothers for boys) as to how to behave as a man or woman (McHale et al., 2003, 2001). Previous studies have found that girls whose mothers did not work, for example, exhibited more feminine behavior (Gold and Andres, 1978; Hoffman, 1974), and boys who had no father figure exhibited less masculine behavior (Russell and Ellis, 1991; Stevenson and Black, 1988). Socialization, then, specifically impacts gendered behaviors in children to the extent that parents encourage or discourage certain gendered behaviors and (intentionally or unintentionally) impress onto their children these gendered identities and attitudes (Lytton and Romney, 1991). In a similar way, mass media can also have a large influence on constructing these ideas of gender-typical behavior and what is acceptable and unacceptable (Signorielli, 2001; Morgan, 1982, 1987).

And, indeed, there is now quantitative evidence demonstrating that differences in socialization explain at least some of the gender differences in performance on a variety of low-level cognitive tasks. For example, there is a significant association between participation in masculine-associated spatial activities like car repair, carpentry, or building model planes (Newcombe et al., 1983) and performance on the mental rotation test (Nazareth et al., 2013). Individual differences in confidence also mediate the sex differences in performance on that test (Estes and Felker, 2012).

Building on all of the above work, in our research, we aimed to represent gender using variables that capture attitudes, predispositions and/or accumulated experiences that reflect gender socialization. We made no further a priori commitments regarding what those specific variables should be and, instead, we decided to seek an established instrument that is reasonably contemporary, that is broadly used, and for which some empirical evidence exists to support its validity. We explain our final choice in Section 3.1.

2.2. Gender in Quantitative HCI Research

We are not aware of any papers in the quantitative HCI literature (through 2024) that represented gender in statistical analyses other than as a single categorical variable with two possible values: men and women. It is possible that exceptions exist but if they do, they are rare. To illustrate the general trend with a small but systematically acquired sample, we searched for papers from three 2024 general HCI conferences (ACM CHI, ACM CSCW and ACM UIST) that included "gender" in the paper title or as an author keyword. We then reviewed these papers and kept those that used quantitative methods and that included some representation of gender in their analyses. We were left with 2 papers from ACM CHI 2024 (Hwang and Won, 2024; Umbach et al., 2024) and 3 papers from ACM CSCW (Mahmood and Huang, 2024; Wang et al.,

2024a,b). We did not find any relevant papers in ACM UIST. All papers represented gender solely as a binary variable. In two of the papers, data from participants identifying as non-binary were collected but they were excluded from analysis (Mahmood and Huang, 2024; Wang et al., 2024b). A "significant imbalance in participant gender" (Mahmood and Huang, 2024) or "very small number of individuals" (Wang et al., 2024b) identifying as non-binary were given as the reasons for not including these individuals in the analyses. Exceptions might exist, of course, but we are not aware of them.

However, when *interpreting* the quantitative results, some researchers are beginning to adopt more elaborate perspectives on gender. For example, Huang et al. (2018) analyzed behavior of participants on a social networking site QQ. Although they represented gender as a binary variable in their statistical analyses, they based their research questions on and analyzed their results through the lens of social role theories of gender. As another example, Wijenayake et al. (2019) also modeled gender as a binary variable but founded their research in the observation that certain topics of interest are stereotyped as masculine or feminine.

One prominent strand of research in quantitative HCI research where we detected an emerging shift in how gender might be represented in quantitative analyses is a body of work collectively referred to as Gender HCI (Beckwith et al., 2006). The general approach taken by Gender HCI researchers was first to synthesize existing social science research on gender differences in problem solving and information processing. Examples of such differences included self-efficacy, motivations, attitude toward risk, and information processing strategies (Burnett et al., 2016). Next, researchers conducted quantitative studies assessing whether—in the context of existing software—people achieved different outcomes depending on their gender or qualities such as self-efficacy. For example, in the context of debugging moderately complex spreadsheets, they found that women were less likely than men to use helpful debugging features that they were previously unfamiliar with (Beckwith et al., 2005). For women in that study, there was a significant positive association between self-efficacy and effectiveness on the debugging task but there was no such association for men. Despite the differences in adoption of debugging features, there was no difference between men and women in how many bugs they identified and solved. However, because women relied more on formula editing than men, they introduced more new bugs (which they did not fix) than men. The researchers hypothesized that these differences materialized because the software, which was designed by a male-dominated industry, was implicitly optimized for the personal qualities and problem solving strategies that are prevalent among men. In a subsequent experiment, they provided compelling evidence for this hypothesis: they redesigned the spreadsheet interface to support individuals with lower self-efficacy and a lower propensity to tinker to explore and adopt new features. The results showed that both men and women benefited from the redesign (compared to the conventional baseline) and that on several outcome measures the differences between men and women were reduced (Grigoreanu et al., 2008).

In all the Gender HCI papers mentioned so far, gender was initially conceptualized as a binary variable. However, recent research (Guizani et al., 2022) descended from this line of work expanded the scope to "inclusivity bugs" (rather than differential outcomes for men and women) and is beginning to focus analytically on the individual characteristics (such as self-efficacy, motivations, attitude toward risk, etc.) relevant to the question at hand rather than binary gender or any other social categories. This body of work is a valuable illustration of how within the same intellectual agenda (design for equitable outcomes among users of computer software) researchers used both coarse gender categories and rich, theoretically-motivated multidimensional representations to capture relevant differences among users.

One more compelling reason for rethinking the representation of gender in quantitative research was offered by Hu and Kohler-Hausmann (2020) in a paper directed at the machine learning community. Their key point was that some variables that are currently considered as external to gender (e.g., a choice of major in college among contemporary women in the US) may need to be considered as *constitutive* features of gender instead. The illustrative example that they bring up is an analysis of admission rates to a competitive university. If choice of major is conceptualized as separate from gender, then a statistical analysis that analyzes admission rates for men and women while controlling for major will show that men and women have equal chances of getting in. However, if we consider that to be a contemporary woman in the US is to be encouraged to prefer humanities over STEM (and the opposite for men) then we are likely to conclude that women have lower chances of university admission because the majors that they are socialized to prefer have lower admission rates. Or, to leave out binary gender out completely, individuals socialized to prefer humanities over STEM have lower chances of university admission than those socialized to prefer STEM over humanities.

3. Hypotheses and Approach

Our hypotheses explore concrete implications of the general claim that gender is, at least in part, socially constructed. The first implication is that people who share a gender identity share experiences of gender socialization. Thus:

H1: We expect that gender socialization is what accounts, at least in part, for the behavior differences between men and women measured in the studies we replicate. Therefore, we hypothesize that the effects of binary gender categories on behavior are *mediated* through individual differences in gender socialization.

In statistics, mediation analysis allows one to quantify to what extent one variable influences another indirectly through a third variable (the mediator) (MacKinnon et al., 2007). In our case, we are interested to what degree a person's gender identity impacts that person's performance on cognitive tasks *indirectly* through gender socialization.

Next, individuals who share a gender identity also have unique experiences of gender socialization that make them develop in distinct ways. For example, previous work argued that girls with mothers working outside the home exhibited less feminine behavior, and boys without a father exhibited less masculine behavior (Gold and Andres, 1978; Hoffman, 1974; Russell and Ellis, 1991; Stevenson and Black, 1988) (these studies only considered families with one male and one female parent). Thus:

H2: We expect that gender categories capture only a fraction of the differences in gender socialization among individuals. Therefore, we hypothesize that regression models that include a measure of individual gender socialization in addition to binary gender categories will explain more of the differences in behavior than models that only include binary gender.

Lastly, we also expect that country-level differences in gender norms and conceptions of masculinity and femininity will result in group-level differences in gender socialization across different countries. This hypothesis builds on previous work demonstrating country-level differences in conformity to gender norms. For example, Tager and Good (2005) documented that Italian male students showed significantly less conformity to masculine norms than American students did, especially those norms which related to individuality, something much more heavily emphasized in the US. Additionally, Sánchez-López and Cuéllar-Flores (2011) argued that Spanish female students showed significantly less conformity to feminine norms than American students did, except for those related to the family, a heavily feminized role in Spanish culture. Because of these cross-country socialization and gender norm differences (and because our participants come from over 200 countries representing a diversity of national cultures), we expect that country-level measures of gender equality and socialization will provide additional predictive and explanatory power in the analysis of the performance differences in these tasks. Thus:

H3: We hypothesize that regression models that include a measure of country-level differences in gender socialization in addition to binary gender will explain more of the differences in behavior than models that only include binary gender.

3.1. Measuring Individual Differences in Gender Socialization

Previous studies have attempted to measure gender socialization largely through understanding how masculine or feminine someone is. In other words, researchers have created studies and questionnaires that attempted to quantify how much someone conforms to either masculine or feminine norms. The original questionnaires developed that initially gained popularity were the Bem Sex Role Inventory (Bem, 1981) and the Personal Attributes Questionnaire (Spence et al., 1974). The Bem Sex Role Inventory (BSRI) focused on having participants identify specific characteristics and traits they possessed, but it was criticized for its now somewhat outdated constructs (Auster and Ohm, 2000) and its use of two different scales to measure masculinity and femininity (where high scores on both was "androgynous" and low scores on both was "undifferentiated") (Pedhazur and Tetenbaum, 1979; Hoffman and Borders, 2001). Whether masculinity and femininity should be measured on separate scales or one spectrum has also been an area of debate, as separating them ignores the relation between them and the way in which they have been defined in relation to each other. The Personal Attributes Questionnaire (PAQ) similarly used multiple scales to measure masculinity and femininity separately, but it also added a third scale to measure bipolar masculinity-femininity (i.e., a spectrum from masculine to feminine). The PAQ also asked participants to rate each personality trait as it fit into the masculine or feminine ideal, in addition to rating themselves on the level with which they possessed the trait, but the test has been similarly criticized for its age and the potential outdated constructs it evaluates (Smiler and Epstein, 2010). A more recent questionnaire, the Traditional Masculinity and Femininity (TMF) scale, instead addresses the constructs of masculinity and femininity as a whole by asking participants directly to define themselves and different aspects of their lives as more masculine or feminine (Kachel et al., 2016).

Another recent instrument is the Conformity to Masculine Norms Inventory (CMNI) (Mahalik et al., 2003). The (CMNI) has been found to be an effective way of characterizing the conformity to gender norms of both men and women and has been found to be generally consistent across different genders and ethnicities/cultures within a country (Hsu and Iwamoto, 2014; Kivisalu et al., 2015; Mahalik et al., 2003; Parent and Moradi, 2009; Parent and Smiler, 2012). This questionnaire asks participants to rate themselves as to how much they agree with a statement representing some construct of masculinity.

According to a recent review (Horstmann et al., 2022) of how sex and gender are operationalized in quantitative health-related research, the BSRI and the CMNI are the two most frequently used instruments for assessing gender socialization.

Because BSRI is substantially older than CMNI and its constructs have been found somewhat outdated by some researchers (Auster and Ohm, 2000), we chose to use CMNI.

Specifically, we used a subset of the recent 29-item variant of the Conformity to Masculine Norms inventory (CMNI-29) (Hsu and Iwamoto, 2014). CMNI-29 has 8 factors. We excluded one of the factors, Power Over Women, because we were concerned that the questions drawn from that factor (e.g., "Women should be subservient to men") would be too off-putting to many participants. For the remaining seven factors, we selected one question per factor, each time picking the question that had the highest factor loading in (Parent and Moradi, 2009). The questions were presented on a 6-point Likert scale anchored at the end points with "Strongly disagree" and "Strongly agree". The questions used in our studies are shown in Table 1.

The choice to use only a subset of CMNI-29 with one item per factor was motivated by our use of the LabintheWild.org platform (Reinecke and Gajos, 2015) to host our studies. As we explain in Section 4.1.2, LabintheWild hosts studies that are completed by unpaid volunteers. To motivate participation, the studies end by giving participants their results, they are also kept brief, and most survey questions are optional. While abbreviating the CMNI questionnaire increases the variance in the responses, it also tends to increase the number of people who complete the studies and who answer all questions. The higher variance in the responses means that the effect sizes are smaller but the large numbers of participants (over 6,000 in each of the primary analyses) makes it possible to detect these smaller effect sizes.

3.2. Measuring Country-Level Differences in Gender Socialization

There are a number of measures of country-level gender parity (Else-Quest and Grabe, 2012). Following some recent work (Zentner and Mitura, 2012), we chose the Global Gender Gap Index (GGI) (World Economic Forum, 2020) because it is designed to be independent of the absolute level of income in any country, thus decoupling gender attitudes from general affluence. GGI is a composite measure capturing gender gaps in access to resources in several domains: economic, educational, health and political. Although GGI does not directly measure gender socialization, it is very likely that it captures it indirectly. Because individual gender socialization continues throughout one's lifetime, we asked participants to report what country they spent most of their childhood in and also in what country they spent most of the past 5 years. Consequently, we included 2 GGI variables in relevant analyses: one for the country of childhood and one for the country where the person resided for most of the past 5 years. We used the 2020 GGI scores. GGI scores range from 0 to 1, with higher scores indicating greater gender parity.

Factor	Question
Emotional Control	25r. I like to talk about my feelings.
Winning	27r. More often than not, losing does not bother me.
Playboy	36. It would be enjoyable to date more than one person at a time.
Violence	41r. No matter what the situation I would never act violently.
Self-Reliance	43. It bothers me when I have to ask for help.
Risk Taking	8. I enjoy taking risks.
Heterosexual Self-Presentation	24. It would be awful if people thought I was gay.

Table 1. CMNI questions used in our study. Item numbers from (Hammer et al., 2018) are shown. Suffix 'r' indicates reverse-coded questions.

3.3. Selection of Studies to Replicate

We replicated 3 studies with well-established gender effects: the Reading the Mind in the Eyes Test (one of the leading instruments for measuring theory of mind—the ability to attribute emotional and cognitive states to others), the Mental Rotations Test, and the Spatial Working Memory Task. We chose these tests because they pertain to fairly basic cognitive abilities and the HCI studies of such basic abilities are the least likely to consider social components of gender. We also chose these studies because of their varying associations to masculine and feminine norms and more socialized behavior, which allows us to see the varying impact of gender socialization on different types of tasks. Specifically, we expected that the effects of gender socialization would be particularly prominent in the context of the Reading the Mind in the Eyes Test. This is because there is some evidence that Theory of Mind can be improved through training (Kloo and Perner, 2008; Kidd and Castano, 2013) and from a young age women tend to be socialized (and pressured), more than men, to be cooperative, "nice", and attentive to other people's feelings (Thorne and Luria, 1986). In contrast, we expected that the effects of gender socialization would be harder to detect in the performance on the other two tests—although there is some evidence that experience impacts performance on each of them, the links to *gender* socialization are less obvious.

3.3.1. Reading the Mind in the Eyes Test

The Reading the Mind in the Eyes test (Baron-Cohen et al., 2001) consists of the participant being shown 37 pictures showing just the eyes part of people's faces (1 picture used for practice, the remaining 36 used for the actual assessment). For each picture, participants are given 4 emotions and asked to tell which emotion the eyes are showing. Previous researchers have argued that men were less accurate and less sensitive in labeling and processing facial expressions and emotions, specifically when the emotions or mental states are represented only by eye stimuli (Kirkland et al., 2013; Montagne et al., 2005).

Women have been found to outperform men on the Reading the Mind in the Eyes Test across countries (Greenberg et al., 2023). The performance on this test has been connected to biological mechanisms such as prenatal testosterone (Chapman et al., 2006), intranasal oxytocin administration (Domes et al., 2007), as well as other genetic patterns (Warrier et al., 2015; Uzefovsky et al., 2019), but new research suggests that performance on the test is only partly genetic (Stewart and Kirkham, 2020). This suggests that perhaps social factors can be an additional contributing factor. For example, one study argued that there was a difference in performance in adolescents relative to their smartphone usage (Stewart and Kirkham, 2020).

3.3.2. Mental Rotations Test

The Mental Rotations Test (MRT) (Vandenberg and Kuse, 1978) consists of the participant being shown 20 shapes. For each shape, participants are given 4 other shapes and are asked to assess which 2 shapes are identical to the original shape (regardless of rotation). Previous researchers have shown that men perform better than women on this test and that these differences are not solely related to time-based performance factors (Geary et al., 2000; Masters, 1998; Peters, 2005; Peters et al., 1995). Past research largely categorizes these distinctions as due to sex, but a few previous studies have produced results showing that the relation between a person's gender and MRT score is mediated to some extent by social factors and life experiences such as the number of masculine spatial activities the participant had engaged in as youth, the confidence of the participant, and the age of the participant, which all speak to impact of socially learned behaviors (Nazareth et al., 2013; Estes and Felker, 2012; Geiser et al., 2008). Similarly, researchers have also explored how a country's culture impacts mental rotation ability through differences in mathematics curricula and even the visual/pictorial aspects of the languages (Jannsen and Geiser, 2011; Sakamoto and Spiers, 2014).

One study looked at the impact of gender socialization and norms on spatial ability through both gendered personality traits and behaviors and found that only one specific category of stereotypically masculine personality traits (those related to agency) contributed significantly to spatial ability (as measured by the MRT) (Saucier et al., 2002). This study points to the possibility of both biology and socialization as significant contributors to spatial ability.

3.3.3. Spatial Working Memory Task

The Spatial Working Memory Task (Duff and Hampson, 2001) consists of the participant being shown a 4×5 board which has a different color dot "hidden" in each square. Participants are asked to find all 10 pairs of dots by color by clicking on each square to look at each dot, with the constraints that they are only able to look at two dots at a time and that the dots will be hidden again after each guess regardless of whether the colors match. This specific task tests spatial working memory and has been shown to produce sex differences, with women performing better in terms of working memory errors and completion time (Duff and Hampson, 2001). Past related research has argued that women have a better spatial memory, specifically object location memory and object identity memory (Alexander et al., 2002; Lejbak et al., 2009; Levy et al., 2005; McBurney et al., 1997; Neave et al., 2005; Spiers et al., 2008; Tottenham et al., 2003). Additionally, two studies have argued that this spatial memory ability depends on age and that women only surpass men in adulthood, raising the question again of how

Figure 1. An example stimulus from the Reading the Mind in the Eyes test

this skill is learned over time and not necessarily only biologically determined (Barnfield, 1999; Voyer et al., 2007). One study even found evidence that childhood experiences of gender nonconformity was related to some types of object location memory, pointing again to socialized behaviors and life experiences as a contributor to these skills (Hassan and Rahman, 2007).

4. Study 1: Reading the Mind in the Eyes Test

4.1. Methods

4.1.1. Task

The Reading the Mind in the Eyes Test consists of the participant being shown 37 pictures showing just the eyes part of people's faces. As shown in Figure 1, for each picture, participants are presented with 4 emotions and asked to decide which emotion the person is feeling. The first picture is treated as a practice task and is excluded from score calculations.

4.1.2. Procedure

The test was launched on the LabintheWild.org platform (Reinecke and Gajos, 2015) for conducting behavioral studies with unpaid online volunteers. Participants on LabintheWild are incentivised to participate in the studies in exchange for a chance to see their results and to compare themselves to others. The quality of the data collected on LabintheWild has been shown to match those collected in traditional laboratory settings (Reinecke and Gajos, 2015; Huber and Gajos, 2020; Li et al., 2018, 2020). In comparison to paid platforms like the Amazon Mechanical Turk, LabintheWild provides access to larger and more diverse participant samples (Reinecke and Gajos, 2015) and LabintheWild volunteer participants have also been shown to provide more reliable data and exert themselves more (Ye et al., 2017; August and Reinecke, 2019).

The landing page for the test offered basic information about the test and it was followed by a detailed consent page. After that, participants filled a demographics questionnaire where all questions were optional.

Next, participants were presented with a brief set of instructions and moved to the 37 items of the Reading the Mind in the Eyes Test. Afterwards, they were presented with the abbreviated CMNI questionnaire, followed by a page asking how well they understood the words used to describe emotions in the study, if they experienced any difficulties during the test, or if they had any comments. At the very end, they were presented with the results page showing how they scored on the Reading the Mind in the Eyes Test and how their score compared to others.

4.1.3. Participant Recruitment

Most participants arrived at the test site through the LabintheWild.org main page, mentions on social media (mostly by other participants), or web search results. We did not conduct a formal power analysis to determine the number of participants for the study. This is because a number of past studies launched on LabintheWild attracted very large numbers of participants, substantially exceeding numbers that would have been indicated by power analyses (e.g., 16,000 participants in (Gajos and Chauncey, 2017), 229,000 in (Gajos et al., 2020), 305,000 in (Greenberg et al., 2023)) with specific numbers varying depending on the popular appeal of the topic of the study and the amount of time the researchers were willing to wait for the results. Because participants do not receive monetary compensation, there is also no resource constraint that would motivate capping recruitment. Instead, we set an informal threshold of at least 1,000 participants. We set this threshold because this number is typically sufficient to detect even small effect sizes (on the order of Cohen's $d = 0.2$) in between-group comparisons (assuming that participants are distributed approximately equally between the two groups) using common statistical techniques like t-test, multiple regression or ANOVA.

4.1.4. Approvals

All studies reported in the manuscript were approved by the Internal Review Board at Harvard University (protocol number IRB20-0578).

4.1.5. Design and Analysis

This was an observational study in which we examined associations between participants' self-reported gender, measures of gender socialization, and their score on the Reading the Mind in the Eyes Test.

Our primary variables were:

- Self-reported gender. Following Spiel et al. (2019), the gender question in our demographics questionnaire offered 5 options: Male, Female, Non-binary, Prefer to self-describe, Prefer not to say. Participants who selected Prefer to self-describe were given an option to write in an answer.
- Reading the Mind in the Eyes Test score (0–36).
- Responses to the 7 CMNI questions (each on a 6-point Likert scale, with 6 always associated with the high conformity to masculine norms).
- GGI (last 5 years): Gender Gap Index for the country where the person spent most of the past five years. As part of the demographics questionnaire we asked each participant where they had lived for most of the past five years.
- GGI (childhood): Gender Gap Index for the country where the person grew up. As part of the demographics questionnaire, we also asked participants "In what country did you live most of your childhood?" (and we instructed them to pick one that influenced them the most if they grew up in more than one country).

We also included the following covariates in our analyses:

- Education ("What is the highest level of education you have received or are pursuing?") {Pre-high school, High school, College, Masters or professional degree, PhD (Doctorate)}
- Age. Because many cognitive abilities improve quickly in childhood and decline slowly through adulthood (see, e.g., Germine et al. (2011)), we followed Germine et al. (2011) and represented age using $\log_{10}(\text{age})$ and $\log_{10}(\text{age})^2$ terms in our models.
- Comprehension of the English words used to describe emotions in the Reading the Mind in the Eyes Test {"I am a native speaker of English", "I am not a native speaker, but I recognized all the words used to describe emotions in the study", "I recognized almost all the words used to describe emotions in the study", "I recognized only some of the words used to describe emotions in the study"}.

To enable additional intersectionality-related analyses, we also collected two other variables:

- National culture. If a participant reported that they currently lived in the same country in which they lived for most of their childhood, we set that country as their national culture. Participants for whom the two answers differed, were not included in the analyses involving national culture.
- Race and ethnicity. Participants were asked to choose any combination of the following options: "Asian or Asian American", "Black or African American", "Latino / Latina or Hispanic", "Native American, American Indian or Alaska Native", "Pacific Islander or Native Australian", or "White". We have also provided an option for participants to self describe their ethnicity. Because different cultures define their ethnic boundaries differently and because social meanings of these categories are culture-dependent, we have only asked these race and ethnicity questions of participants who accessed the study from within the United States.

As conceptually illustrated in Figure 2, we conducted a mediation analysis using the R package mma (Yu and Li, 2017; Yu et al., 2019) (version 10.3-2). Under some structural assumptions, this analysis identifies what proportion of the relationship between the self-reported binary gender (independent variable) and the score on the Reading the Mind in the Eyes Test (dependent variable) can be ascribed to socialization as measured by CMNI (mediator variables). Under a mediation model, identifying with a particular binary gender category leads to particular gender-specific socialization experiences as captured by CMNI scores. Subsequently, the differences in socialization lead to differences in performance on the Mind in the Eyes test. This is reported as the "indirect effect." The "direct effect," by contrast, captures the relationship between gender and the test score that is not explained by the intermediate CMNI. We used a non-parametric bootstrap to compute the 95% confidence intervals for the effects estimated with the mediation analysis. We considered the presence of an indirect effect significantly different from 0 to be evidence in support of **H1**.

We compared a sequence of linear regression models to determine if adding gender socialization (as measured with CMNI or GGI) significantly improved the explanatory power of the model compared to simpler models containing only covariates and binary gender. We report adjusted R^2 (denoted $\bar{R}^2$), i.e., the fraction of the variance explained by each model, as a measure of the explanatory power of each model. We used ANOVA for statistical comparisons of the models. In the context of our study, we considered **H2** or **H3** supported if adding CMNI or GGI items to a model that already included covariates and the binary gender variable significantly improved the explanatory power of the model.

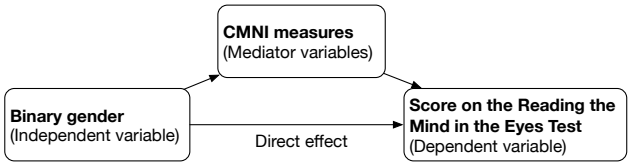


Figure 2. A mediation model can identify how much of the total effect of the independent variable (binary gender in our case) on the dependent variable (score on the Reading the Mind in the Eyes test) is mediated through mediator variables (CMNI measures). The mediated part of the effect is referred to as the indirect effect while the remaining part is referred to as direct effect.

Besides statistical significance, when analyzing the regression results, we also considered the magnitude of the benefits conferred by including CMNI or GGI in the analysis relative to the benefits of including binary gender. Specifically, the baseline for our comparisons is the difference in $\bar{R}^2$ between a model that includes baseline covariates with gender and a model that includes only baseline covariates. Given that the effects of binary gender are considered scientifically meaningful for this test, we consider the explanatory benefits of CMNI or GGI relative to binary gender as a benchmark.

4.2. Results

4.2.1. Participants

248,256 participants completed the test, responded to the CMNI items and provided their education level, reported being 13 years old or older, and responded to the comprehension question. Of those, 243,558 who identified as either men or women were included in the mediation analyses. The additional 4,698 participants who chose the "Non-binary" option were included in some of the additional analyses. Participants who chose to self-describe were excluded because a detailed analyses of their responses indicated that some of them were non-binary individuals (e.g., "non-binary femme"), while some others were trolling (e.g., "Attack helicopter") or protesting the inclusion of non-binary options in the questionnaire (e.g., "there are only 2 genders"), and yet others could not be unambiguously categorized—a situation observed in some other settings (Jaroszewski et al., 2018). Participants who did not share what country they grew up in or where they lived for the past five years, or whose countries were not included in the Gender Gap Index, were omitted from the regression analyses. Participants reported having grown up in over 200 different countries and territories, of which 150 had a Gender Gap Index score available. The top six most represented countries of childhood were United States (103,853), United Kingdom (19,486), Canada (10,328), Australia (9,977), India (8,157), and Philippines (7,414). The distribution of countries where people were living for the past 5 years was similar. 88.4% of the participants reported currently living in the same country in which they grew up.

4.2.2. Preliminary Analyses

We conducted a separate linear regression for each CMNI item with gender (Man, Woman, Non-binary) as the sole predictor. Using ANOVA, for each item, we observed a significant

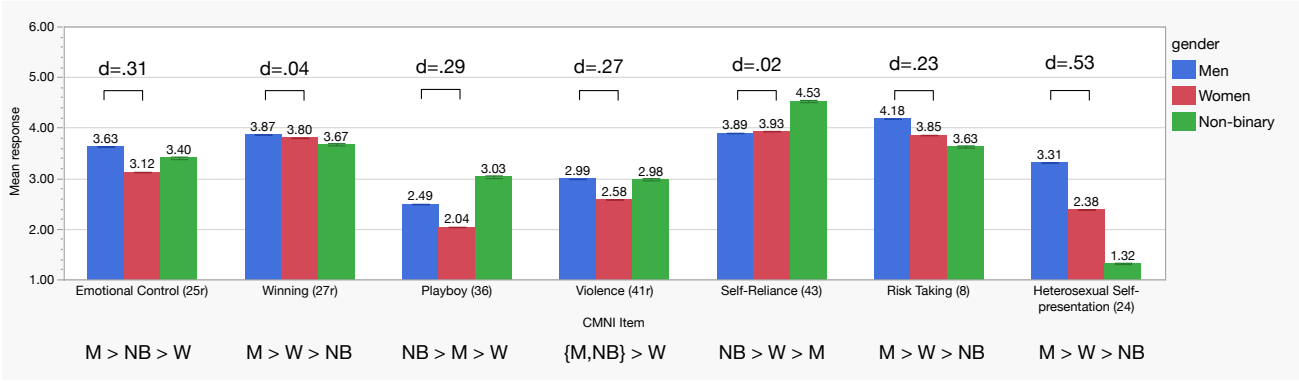


Figure 3. Mean CMNI responses by gender. Higher scores indicate higher conformity to masculine norms. Error bars show 95% Confidence Intervals. Absolute Cohen's d effect size is given for differences between men and women. The comparisons below show results of the post hoc Tukey HSD comparisons: the > symbols indicates statistically significant differences while genders sharing braces are not significantly different from each other. Total N=248,256.

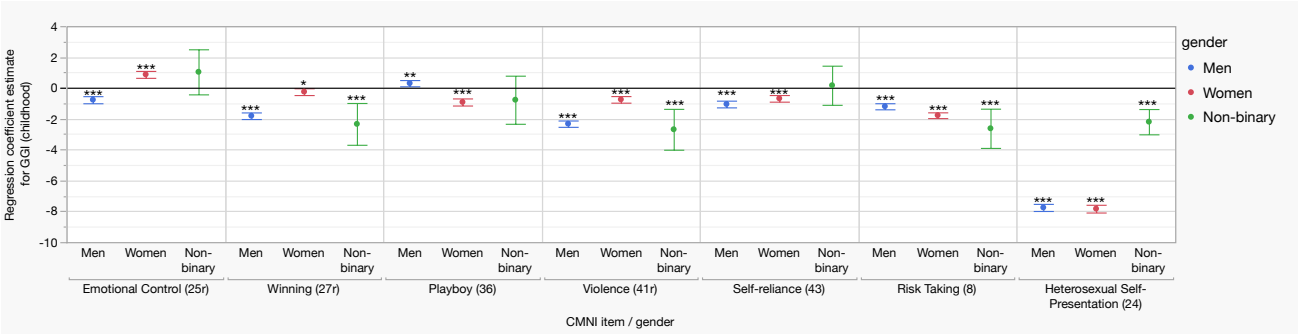


Figure 4. Regression coefficients from separate models for each gender/CMNI item combination predicting CMNI from GGI. Results indicate that there is a statistically significant association between most CMNI items and GGI (childhood). In most cases, greater gender parity (higher GGI) results in less conformity to masculine norms (as indicated by negative regression coefficients). Results for GGI (last 5 years) were similar and are not reported. Error bars show 95% Confidence Intervals. Total N=252,939. * $p < .05$, ** $p < .01$, *** $p < .001$

main effect (at $p < .0001$) of gender on the response to the CMNI. The mean responses are shown in Figure 3 together with the results of a post hoc Tukey HSD analysis. For differences between men and women, we also included effect sizes (as Cohen's d) to help illustrate the extent to which CMNI captures differences between binary genders. This analysis showed that participants who identified as men scored significantly higher on all CMNI questions except Self-Reliance (item 43) than people who identified as women. For Self-Reliance, women scored higher than men but the magnitude of the difference was negligible ($d = .02$). These results confirm that CMNI generally captures differences in contemporary binary gender norms in our sample.

As shown by the results of the post hoc Tukey HSD analyses, people who identified as non-binary had CMNI scores that varied differently than the scores of people who identified with either of the binary genders. We will return to these results in Discussion, but for now we note that they indicate that CMNI items may measure different underlying constructs or particular life experiences of non-binary individuals compared to people who identify as either men or women.

Next, we conducted linear regression analyses with GGI (childhood) as the sole predictor separately for each CMNI item and gender category. As shown in Figure 4, there was a statistically significant effect of GGI (childhood) on all CMNI items for women and for men. Because there were many fewer non-binary participants (N=4,796) than either men or women in this analysis, only 4 effects were statistically significant for this group. In most cases, greater gender parity in the country were one grew up (i.e., greater GGI score) was associated with lower conformity to masculine norms (i.e., lower CMNI scores) as indicated by the predominantly negative regression coefficients. These results indicate that there are links between society-level norms and individual gender attitudes. Because 88.4% of our participants reported currently living in the same country in which they grew up, results for GGI (last 5 years) were similar to those for GGI (childhood) and are not reported.

Finally, as shown in Table 2, the correlations among responses to the CMNI questions are small ($r < .3$) or very small ($r < .1$). As expected, given that we selected one item per factor, each question appears to measure a distinct construct. Because these responses are only minimally correlated, we will treat them as independent in subsequent analyses.

4.2.3. Main Results

As reported in Table 3, our mediation analysis showed a statistically significant total indirect effect indicating that the impact of binary gender on the score in the Reading the Mind in the Eyes Test was mediated through aspects of socialization captured by the CMNI. The CMNI variable most responsible for the results was the Heterosexual Self-presentation. The total indirect effect accounted for an estimated 61.7% of the total effect. This result supports hypothesis **H1**.

	(27r)	(43)	(41r)	(24)	(8)	(25r)
Winning (27r)						
Self-Reliance (43)	0.10****					
Violence (41r)	0.16****	0.05****				
Heterosexual Self-Presentation (24)	0.09****	0.07****	0.09****			
Risk Taking (8)	-0.01****	-0.06****	0.06****	0.07****		
Emotional Control (25r)	0.08****	0.22****	0.12****	0.14****	-0.10****	
Playboy (36)	0.01****	0.01****	0.09****	0.00	0.13****	-0.05****

Table 2. Correlations among responses to different CMNI questions for participants who identified as either men or women. * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$

As shown in Table 4, the model in Step 3a, which includes the baseline covariates, binary gender and CMNI, is a significantly better fit than the model in Step 2a ($\Delta\bar{R}^2 = 0.020$; $F(7, 231479) = 178.28$; $p < 10^{-15}$) which includes only the covariates and binary gender. This result is meaningful as the magnitude of this benefit is several times larger than the benefit of adding binary gender to a model that only includes baseline covariates ($\Delta\bar{R}^2 = 0.005$). The results of this analysis support hypothesis **H2**: CMNI captures relevant information that goes beyond what is already captured by one's binary gender.

Similarly, the model in Step 3b, which includes GGI in addition to covariates and binary gender, is a significantly better fit than the model in Step 2a ($\Delta\bar{R}^2 = 0.004$; $F(2, 231484) = 445.14$; $p < 10^{-15}$). The magnitude of this improvement in fit is nearly as large as the improvement due to inclusion of binary gender. This result indicates that GGI also adds relevant information and supports hypothesis **H3**.

The model that includes both CMNI and GGI in addition to covariates and binary gender (Step 5), is a significantly better fit than either a model that included only CMNI (Step 3a) or the one that included only GGI (Step 3b) indicating that CMNI and GGI capture some complementary information. This said, the magnitude of the improvement in fit due to adding GGI to a model that already includes CMNI and binary gender is relatively small ($\Delta\bar{R}^2 = 0.001$).

The model in Step 4a, which extends the model in Step 3a by including interaction terms between binary gender and CMNI questions, significantly improves the goodness of fit compared to the model in Step 3a ($\Delta\bar{R}^2 < .001$; $F(7, 231472) = 29.762$; $p < 10^{-15}$) but the magnitude of this improvement is negligible. Because the interaction terms allow for a different relationship between CMNI and outcome for men and women, the lack of a meaningful difference in goodness of fit between the model with the interaction terms and the one without indicates that a single model (without interaction terms) can be used for both men and women together. Similarly, the model in Step 4b, which extends model in Step 3b by adding interaction terms between GGI measures and gender, significantly improved goodness of fit ($\Delta\bar{R}^2 < .001$; $F(2, 231482) = 72.143$; $p < 10^{-15}$) but the magnitude of this improvement was also negligible. We conclude that a single model can be used for both men and women to measure the impact of GGI.

4.2.4. Additional Analyses: Intersectionality

Our social identities are multidimensional and these dimensions combine in complex ways to create distinct identities. Each combination of dimensions, such as gender and race, can create a distinct social identity that cannot be understood by studying each dimension in isolation. The importance of directly studying such *intersectional* identities was prominently illustrated in the computing research community by the "Gender Shades" paper (Buolamwini and Gebru, 2018). In that paper, Buolamwini and Gebru (2018) demonstrated that computer vision algorithms for gender recognition performed particularly poorly for dark-skinned women — a result that was not detectable by analyzing these algorithms' performance separately by gender and skin tone. Only an intersectional analysis, that looked at each combination of these factors separately, made the inequities apparent.

An analogous equity-related concern in our research is whether the modeling insights that we have demonstrated on the sample as a whole apply equally well to the individual subgroups that comprise it. Specifically, so far our study has demonstrated that, for our combined sample, variables that quantitatively capture aspects of gender socialization explain some of the information captured by the binary gender variable (**H1**) and add further relevant information not captured by binary gender (**H2**). Will these results hold if we intersect gender with other dimensions of identity? To begin to answer this question, we conducted an additional set of post hoc analyses.

Methods.

In these additional analyses, we consider intersections of gender and national culture, and gender and race/ethnicity for participants from the United States. We acknowledge at the outset that relying on any socially-constructed categories has limitations when trying to conceptualize intersectional identities (McCall, 2005; Rankin and Thomas, 2019). Thus, we consider the following analysis as preliminary.

There are multiple ways of operationalizing intersectionality in statistical analyses, three of which are discussed by Scott and Siltanen (2017). Following one of the suggested approaches, we decided to conduct our analyses by disaggregating data by another aspect of identity (i.e., national culture or race), repeating our earlier analyses separately for each subgroup, and then qualitatively comparing the results.

For these analyses, we included subgroups for which we had at least 2,000 participants. For smaller groups, we found that the confidence intervals in the mediation analyses were too large to enable meaningful comparisons.

Results.

There were 11 countries (Australia, Brazil, Canada, France, Germany, India, Netherlands, Philippines, Russia, United Kingdom, and United States) and 5 race/ethnicity identities among US participants (Asian or Asian American; Black or African American; Latino / Latina or Hispanic, a multi-ethnicity identity of Latino / Latina or Hispanic and White; and White) with at least 2,000 participants each.

For the mediation analyses (related to hypothesis **H1**) across national cultures (Table 11 in Appendix A), there was a significant indirect effect of CMNI for all countries analyzed, though

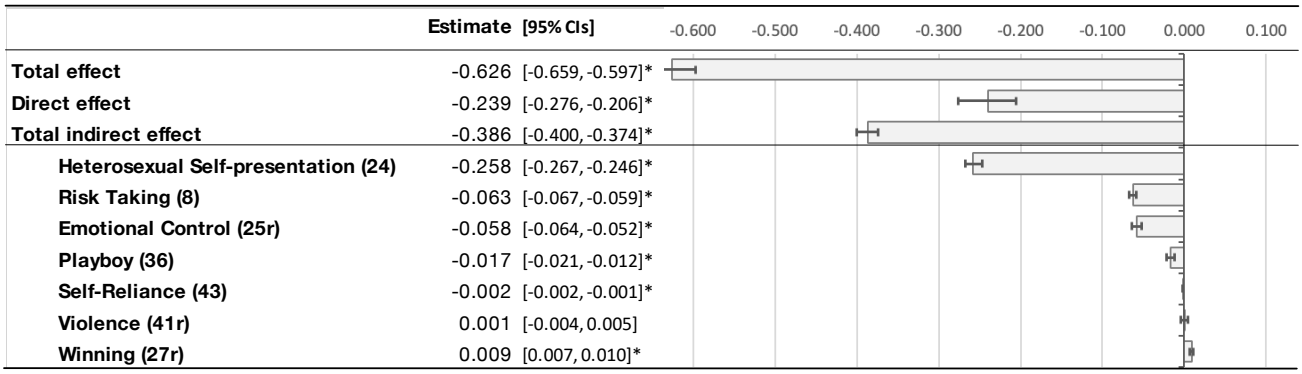
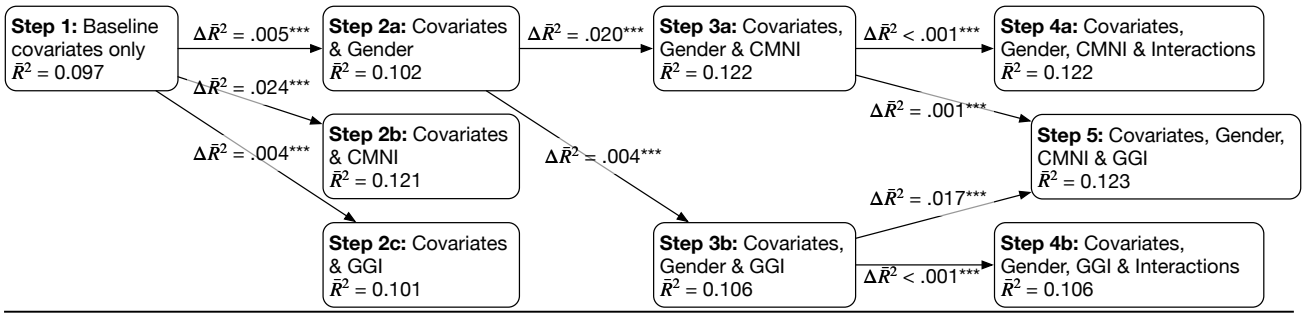


Table 3. Results of mediation analysis for the Reading the Mind in the Eyes Test showing the effect of binary gender (identifying as a man) on the test score. Overall, men scored 0.626 points lower than women (the Total effect). The Total *indirect* effect, which shows how much of the gender effect was mediated through aspects of gender socialization captured by the CMNI, is estimated to be 61.7% of the total effect and is significantly different from zero. * $p < .05$. The 95% Confidence Intervals were estimated using a casewise bootstrap.



	Step 2a Gender	Step 3a Gender & CMNI	Step 3b Gender & GGI	Step 5 Gender, CMNI & GGI
(Intercept)	−2.86 (0.61)***	3.74 (0.62)***	−7.44 (0.63)***	0.21 (0.64)
Education: high school	0.69 (0.06)***	0.72 (0.06)***	0.68 (0.06)***	0.71 (0.06)***
Education: college	1.28 (0.07)***	1.28 (0.06)***	1.31 (0.07)***	1.30 (0.06)***
Education: masters	1.53 (0.07)***	1.50 (0.07)***	1.54 (0.07)***	1.51 (0.07)***
Education: PhD	2.07 (0.08)***	1.95 (0.08)***	2.11 (0.08)***	1.98 (0.08)***
Comprehension: understood all words	−1.93 (0.03)***	−1.76 (0.03)***	−1.89 (0.03)***	−1.73 (0.03)***
Comprehension: understood most words	−1.51 (0.02)***	−1.38 (0.02)***	−1.49 (0.02)***	−1.38 (0.02)***
Comprehension: understood some words	−3.42 (0.04)***	−3.15 (0.04)***	−3.31 (0.04)***	−3.09 (0.04)***
$\log_{10}$(age)	37.63 (0.85)***	30.55 (0.85)***	37.20 (0.85)***	30.55 (0.85)***
$\log_{10}(\text{age})^2$	−12.03 (0.29)***	−9.79 (0.29)***	−11.94 (0.29)***	−9.81 (0.29)***
Gender (man)	−0.63 (0.02)***	−0.24 (0.02)***	−0.63 (0.02)***	−0.26 (0.02)***
Winning (27r)		0.12 (0.01)***		0.12 (0.01)***
Self-Reliance (43)		0.04 (0.01)***		0.04 (0.01)***
Violence (41r)		0.00 (0.01)		0.01 (0.01)
Heterosexual Self-Presentation (24)		−0.28 (0.00)***		−0.26 (0.01)***
Risk Taking (8)		−0.20 (0.01)***		−0.19 (0.01)***
Emotional Control (25r)		−0.12 (0.01)***		−0.12 (0.01)***
Playboy (36)		−0.04 (0.01)***		−0.03 (0.01)***
GGI (country of childhood)			4.48 (0.38)***	3.27 (0.38)***
GGI (main residence in last 5 years)			2.30 (0.40)***	1.51 (0.40)***
R^2	0.102	0.122	0.106	0.123
$\bar{R}^2$	0.102	0.122	0.106	0.123
Num. obs.	231,497	231,497	231,497	231,497

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 4. Results of regression analyses for the Reading the Mind in the Eyes Test. The diagram above shows the model hierarchy and summarizes the model comparison results. Details of a subset of the models are shown in the table (standard errors in parentheses; pre-high school was the reference value for Education; native speaker was the reference for Comprehension). The key results are the significant difference between Steps 3 and 2, which demonstrate that both CMNI and GGI capture relevant information that goes beyond what is already captured by binary gender. $\bar{R}^2$ denotes adjusted R^2.

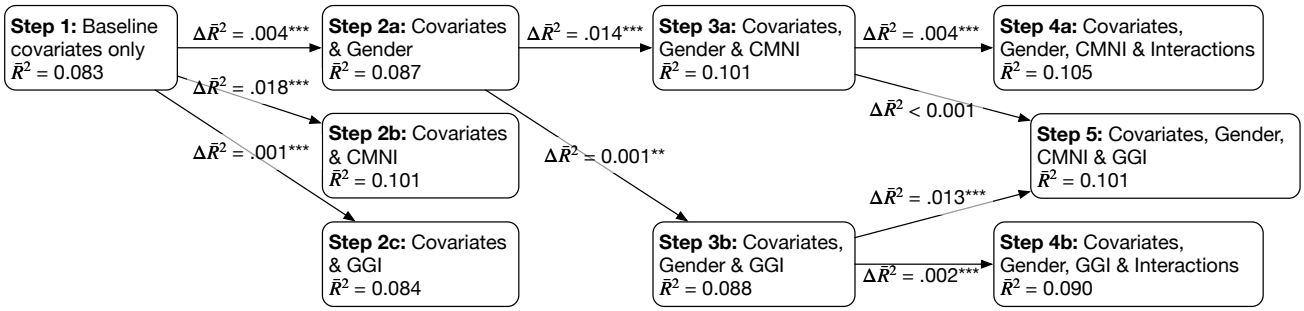


Figure 5. Overview of the model comparison results of regression analyses that included 3 gender categories: men, women and non-binary. Of interest now is the difference between Step 4a (model that includes interactions between categorical gender and CMNI) and Step 3a (which does not include the interaction terms). This difference is significant and the magnitude of the improvement in model fit (captured through $\Delta\bar{R}^2$) is as large as improvement in model fit due to adding categorical gender (Step 2a) to a model that only included covariates. $\bar{R}^2$ denotes adjusted R^2.

it ranged in magnitude from CMNI mediating 23% of the total effect for Russia to 99% for Netherlands. In other words, in Russia, gender socialization as captured by CMNI explained only about a quarter of the effect of binary gender on the performance on the test. In Netherlands, in contrast, *all* of the effect of binary gender was mediated through CMNI. For all national cultures analyzed, Heterosexual Self-presentation had the largest effect except in the United Kingdom where Emotional Control was estimated to have a slightly larger effect (not significantly different from Heterosexual Self-presentation, though). In most countries, more than half of the indirect effect was accounted for by Heterosexual Self-presentation. The three exceptions were France, United Kingdom, and India. In France and United Kingdom, the effect of Emotional Control was very close to Heterosexual Self-presentation, whereas in India the two other large contributors were Winning and Risk Taking.

For the mediation analyses across race/ethnicity identities among US participants (Table 10), there was also a significant indirect effect of CMNI for all groups. For people who identified as Asian or Asian American, 40% of the binary gender effect was mediated through CMNI while for people who identified as Black or African American, the entire effect was mediated through CMNI. For all groups, Heterosexual Self-presentation was the most impactful dimension. For participants who identified as white, Emotional Control was also notable (half of the magnitude of Heterosexual Self-presentation) while for other groups, no other dimension rose to prominence.

For the regression analyses (related to hypothesis **H2**), Table 12 shows that for all race/ethnicity groups included, adding CMNI to the analyses resulted in similar model fit improvements (as shown by $\Delta\bar{R}^2$ Step 3a - Step 2a and $\Delta\bar{R}^2$ Step 2b - Step 1). Inclusion of CMNI also improved model fits for all countries analyzed, though the magnitude of the improvement varied considerably (Table 13). Specifically the $\Delta\bar{R}^2$ achieved by adding CMNI to a model that already included binary gender and covariates ranged from .006 in Russia and Brazil to .025 in India. For all race/ethnicity identities and for all countries except for Russia, the magnitude of the improvement in fit due to adding CMNI to a model that already included gender and covariates was larger than the initial benefit of adding binary gender to a model that only included covariates.

4.2.5. Additional Analyses: Generalizability

Because in the preceding analyses CMNI appeared to explain more variance than binary gender (see R^2 for Step 2b vs Step 2a in Table 4), we asked if future analyses that needed to include the effect of gender socialization as a covariate could use CMNI *instead of* categorical gender so that having many fewer non-binary participants than men or women (a stated reason for excluding non-binary participants from analyses in some research, e.g., (Mahmood and Huang, 2024; Wang et al., 2024b)) would not stand in the way of including non-binary participants in analyses.

To answer this question, we repeated the comparison between Steps 3a and 4a, but this time we included the participants who identified as non-binary in addition to those who identified as either men or women. We also took a random sample of the data for men and women so that we would have an equal number of samples for each gender category (4,434 per gender category; 13,302 total). The results are summarized in Figure 5. As a baseline for evaluating the magnitude of the effect of including interaction terms, the improvement in model fit due to adding categorical gender to a model that already included baseline covariates (the difference between Step 2a and Step 1) was $\Delta\bar{R}^2 = .004$ ($F(2, 13290) = 33.776$; $p < 10^{-15}$).

We found that the new model in Step 4a (which included interaction terms between gender and CMNI) significantly improved model fit over the model in Step 3a, which did not include the interaction terms ($\Delta\bar{R}^2 = .004$; $F(14, 13269) = 5.615$; $p < 10^{-10}$). The magnitude of the improvement due to adding the interaction terms ($\Delta\bar{R}^2 = .004$) was as large as the magnitude of the improvement due to adding categorical gender. Thus, we consider this effect to be meaningful. In particular, there were significant interactions between gender and Risk Taking, and between gender and Emotional Control—in both cases, participants who identified as non-binary were modeled significantly differently from participants who identified as either men or women.

To examine these differences further, we fit separate models for men, women and non-binary participants (Table 5, this time using all available data). Inspecting the three models, we note that for all three gender categories, Heterosexual Self-Presentation has the highest coefficient value among the 7 CMNI variables. However, for men and women, the second most impactful CMNI variable is Risk Taking while for non-binary participants it is Emotional Control.

These results indicate that it may not be accurate to use a single model with CMNI instead of categorical gender to simultaneously model people who identify with either of the binary genders and those who identify as non-binary.

	Men	Women	Non-binary
(Intercept)	−4.92 (0.89)***	11.41 (0.82)***	−21.37 (4.89)***
Education: high school	0.68 (0.09)***	0.74 (0.08)***	0.01 (0.39)
Education: college	1.23 (0.09)***	1.36 (0.09)***	0.18 (0.41)
Education: masters	1.40 (0.09)***	1.63 (0.09)***	0.32 (0.44)
Education: PhD	1.87 (0.11)***	2.09 (0.10)***	0.17 (0.51)
$\log_{10}$(age)	41.31 (1.24)***	20.71 (1.14)***	65.08 (6.89)***
$\log_{10}(\text{age})^2$	−13.34 (0.42)***	−6.58 (0.38)***	−21.19 (2.39)***
Comprehension: understood all words	−1.74 (0.04)***	−1.73 (0.04)***	−1.69 (0.25)***
Comprehension: understood most words	−1.39 (0.03)***	−1.35 (0.03)***	−0.70 (0.17)***
Comprehension: understood some words	−3.23 (0.05)***	−3.03 (0.05)***	−2.58 (0.41)***
Winning (27r)	0.13 (0.01)***	0.11 (0.01)***	0.13 (0.04)**
Self-Reliance (43)	0.05 (0.01)***	0.04 (0.01)***	−0.02 (0.05)
Violence (41r)	0.06 (0.01)***	−0.05 (0.01)***	0.00 (0.05)
Heterosexual Self-Presentation (24)	−0.27 (0.01)***	−0.29 (0.01)***	−0.36 (0.08)***
Risk Taking (8)	−0.19 (0.01)***	−0.20 (0.01)***	0.06 (0.05)
Emotional Control (25r)	−0.06 (0.01)***	−0.17 (0.01)***	−0.31 (0.04)***
Playboy (36)	−0.05 (0.01)***	−0.02 (0.01)**	−0.02 (0.04)
R^2	0.120	0.116	0.097
$\bar{R}^2$	0.120	0.116	0.094
Num. obs.	112,757	130,801	4,698

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5. Separate models for the score on the Reading the Mind in the Eyes Test for participants who identified as men, women, or non-binary. Standard errors in parentheses; pre-high school was the reference value for Education; native speaker was the reference for Comprehension. $\bar{R}^2$ denotes adjusted R^2.

5. Study 2

In Study 2, we replicated the Mental Rotations Test (Peters et al., 1995) and the Spatial Working Memory Task (Duff and Hampson, 2001).

5.1. Methods

5.1.1. Task

In this study, participants were asked to first complete the Mental Rotations Test followed by the Spatial Working Memory Task.

The Mental Rotations Test consisted of 10 trials. In each trial (Figure 6 Left), participants were shown a three-dimensional shape (the reference) and 4 additional shapes. Participants were asked to identify which 2 of the additional shapes were identical to the reference shape, just rotated.

The Spatial Working Memory Task consisted of the participant being shown a 4 × 5 board which had a different dot "hidden" in each square (Figure 6, Right). There were 10 pairs of dots, each pair had its own color and letter. Participants were asked to find all 10 pairs of dots by clicking on each square to look at each dot, with the constraints that they were only able to look at two dots at a time and that the dots would be hidden again after each guess regardless of whether the colors matched.

5.1.2. Procedure

This study was also launched on LabintheWild. As with the previous study, the landing page offered basic information about the test and it was followed by a detailed consent page. After that, participants filled a demographics questionnaire where all questions were optional.

Next, participants were presented with instructions for the Mental Rotations Test, which included one practice trial, followed by the actual test. After that, they were given an option to take a break before being presented with the instructions and the main interface for the Spatial Working Memory Task.

After completing both tests, participants were presented with the abbreviated CMNI questionnaire, followed by a page asking if they experienced any difficulties during the test, or if they had any comments. At the very end, they were presented with the results page showing how they scored on both tests and how their scores compared to those of other participants.

5.1.3. Participant Recruitment

Most participants arrived at the test site through the LabintheWild.org main page, mentions on social media (by other participants), or web search results.

5.1.4. Design and Analysis

We analyzed the data separately for each test using the same methods as in Study 1.

The main performance measure for the Mental Rotations Test was the average number of correct choices for each trial (max 2).

For the Spatial Working Memory Task, we measured the total number of attempts (one attempt = a pair of squares uncovered) taken to find all pairs.

Our preliminary analyses showed that age was the only covariate that significantly improved the statistical models for both tests so this is the only covariate used in our analyses. As before, the relationship between age and performance was best modeled as log-quadratic for both tests.

5.2. Results

5.2.1. Participants

7,723 participants (3,744 women, 3,713 men, 266 non-binary) completed the Mental Rotation Test and provided their gender and age. The Gender Gap Index could be assigned to 6,993

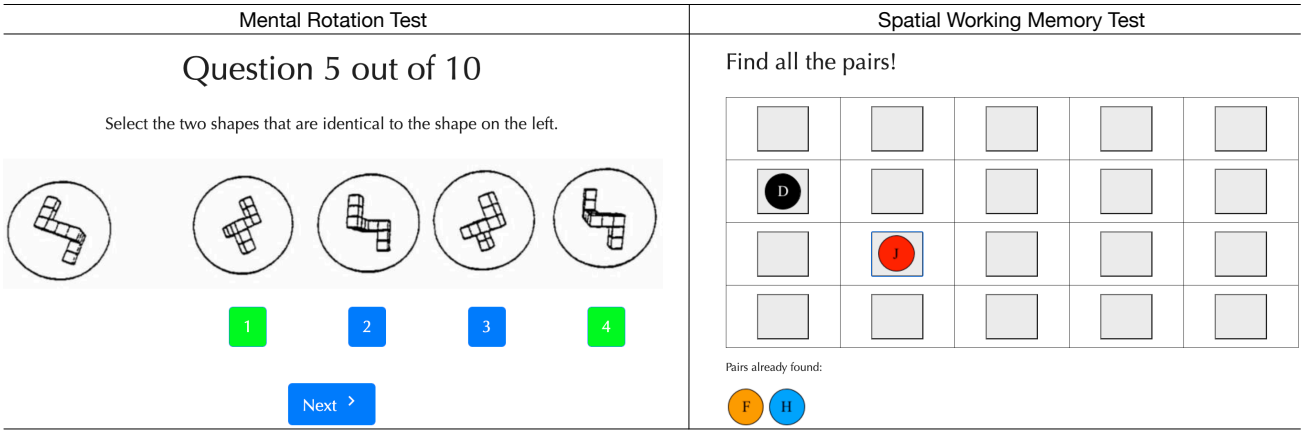


Figure 6. Left: A trial in the Mental Rotations Test (a participant selected options 1 and 4). Right: The Spatial Working Memory Task interface.

participants (3,382 women, 3,376 men, 235 non-binary). For the Spatial Memory Working Test, 7,730 participants (3,746 women, 3,717 men, 267 non-binary) completed the task and provided their gender and age. The Gender Gap Index could be assigned to 7,000 participants (3,384 women, 3,380 men, 236 non-binary).

Participants who completed the Mental Rotation Test reported having grown up in 166 different countries, with the top 6 most frequently listed being: United States (3,267), United Kingdom (859), Canada (442), Australia (362), Sweden (225), and India (217). 87.5% participants reported having spent most of the past 5 years in the same country in which they grew up. The numbers for the Spatial Memory Working Test are nearly identical as the two tests were administered together.

5.2.2. Main Results: Mental Rotations Test

As shown in Table 6, in our mediation analysis we observed a statistically significant total indirect effect indicating that the impact of binary gender on the score in the Mental Rotations Test is mediated through aspects of socialization captured by the CMNI. The magnitude of the indirect effect is relatively small: 11.9% of the magnitude of the total effect. Interestingly, the indirect effect is negative, while the total effect is positive. This suggests that identifying as a man is associated with a higher score on this test, but being *socialized* to conform to current masculine norms is associated with a *decrease* in performance. This result supports hypothesis **H1**, but the support is weak (small relative magnitude of the indirect effect) and we acknowledge that we did not expect gender identity and gender socialization to have opposite effects.

As shown in Table 7, the model in Step 3a, which includes age, binary gender and CMNI, is a significantly better fit than the model in Step 2a ($\Delta R^2 = 0.016$; $F(7, 6747) = 19.631$; $p < 10^{-15}$) which includes only age and binary gender. The relative magnitude of this improvement in model fit is only 16.3% of the improvement in fit due to adding binary gender to a model that only includes baseline covariates ($\Delta R^2 = 0.098$; $F(7, 6754) = 756.91$; $p < 10^{-15}$). This result weakly supports hypothesis **H2**: CMNI captures relevant information that goes beyond what is already captured by one's binary gender but the relative benefit in terms of model fit is small compared to the benefits of using a categorical gender variable.

The model in Step 4a, which extends the model in Step 3a to include interaction terms between binary gender and CMNI questions did not significantly improve the explanatory power of the model ($\Delta R^2 = .001$; $F(7, 6740) = 1.5428$; n.s.) indicating that men and women can be modeled together.

Adding GGI to a model that already included age and gender (Step 3b vs Step 2a) significantly improved model fit ($\Delta R^2 = 0.013$; $F(2, 6752) = 51.043$; $p < 10^{-15}$) but the magnitude of the improvement was only 13.3% of the benefit of adding the categorical gender variable to a model that already contained baseline covariates. Thus hypothesis **H3** is weakly supported for the Mental Rotations Test. Adding interaction effects between gender and GGI (Step 4b vs Step 3b) did not significantly improve model fit ($\Delta R^2 = . - 001$; $F(7, 6750) = 0.1301$; n.s.), indicating that, as with CMNI, men and women can be modeled together.

5.2.3. Main Results: Spatial Working Memory Task

As shown in Table 8, our mediation analysis showed a statistically significant total effect of binary gender on the number of attempts taken to complete the Spatial Working Memory Task: consistent with prior research, men performed worse (needed to make more attempts) than women. Consistent with hypothesis **H1**, there was also a statistically significant indirect effect of binary gender on the number of attempts taken to complete the Spatial Working Memory Task. The indirect effect accounted for 57.2% of the total effect. This result indicates that a substantial part of the total gender effect observed on this task was mediated through aspects of gender socialization captured by the CMNI.

As shown in Table 9, both hypotheses **H2** and **H3** are supported: adding CMNI to a model that contained only baseline covariates and binary gender significantly improved model fit ($\Delta R^2 = .012$ between Steps 3a and 2a; $F(7, 6753) = 12.593$; $p < 10^{-15}$). Similarly, adding GGI to the model in Step 2a also significantly improved model fit ($\Delta R^2 = .002$ between Steps 3b and 2a; $F(2, 6758) = 8.3879$; $p < 10^{-15}$). The magnitudes of these improvements are meaningful compared to the improvement in model fit due to adding binary gender to a model that contained only baseline covariates (Step 2a vs Step 1) which was $\Delta R^2 = .002$ ($F(2, 6760) = 14.369$; $p = 0.0002$).

The model that included *both* CMNI and GGI in addition to binary gender and baseline covariates (Step 5) was significantly better than the model in Step 3a which lacked GGI but the magnitude of the improvement was negligible ($\Delta R^2 < 0.001$; $F(2, 6751) = 3.3354$; $p = 0.0357$). This suggests that GGI did

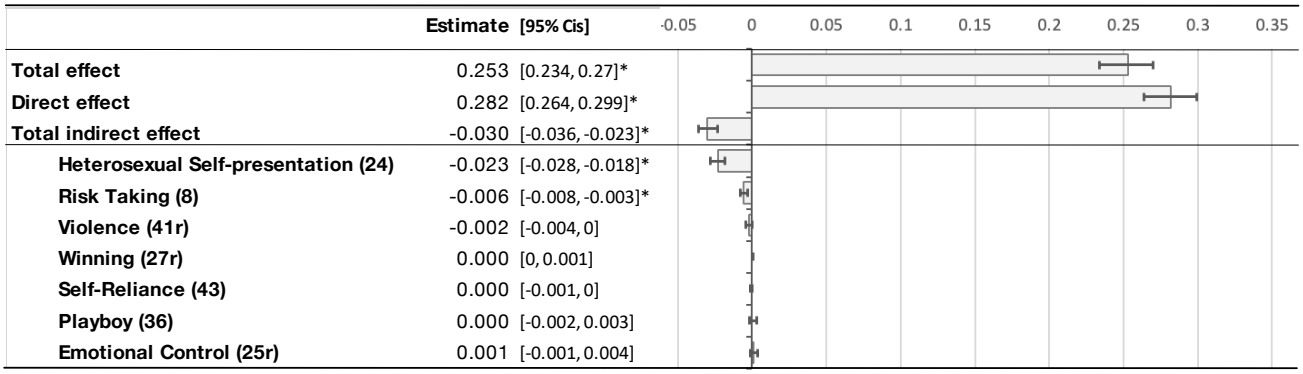
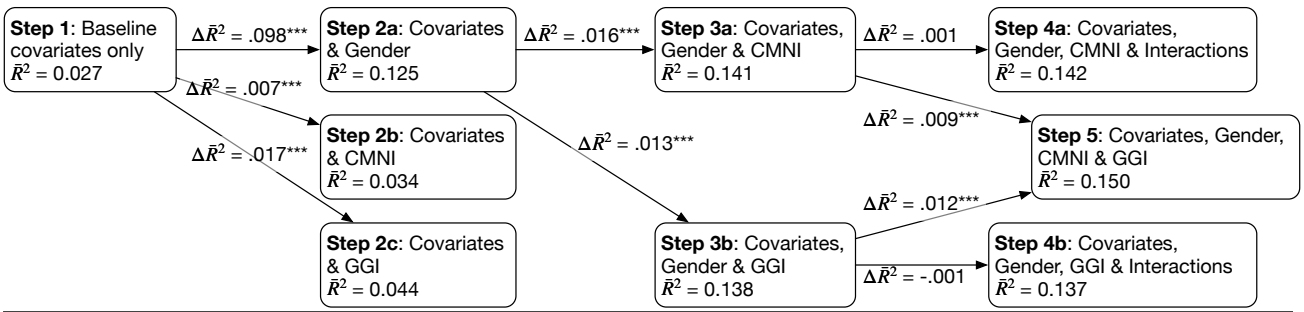


Table 6. Results of mediation analysis for the Mental Rotations Test showing the effect of gender (identifying as a man) on the average number of correct responses per trial. Overall, men gave 0.253 more correct responses per trial than women (the total effect). The total *indirect* effect, which shows how much of the gender effect was mediated through aspects of gender socialization captured by the CMNI, is significantly different from zero, but it points in the opposite direction from the total effect. This indicates that identifying as a man confers an advantage on this test, but being *socialized* to conform to masculine norms is associated with a decrease in performance. * $p < .05$. The 95% Confidence Intervals were estimated using a casewise bootstrap.



	Step 2a Gender	Step 3a Gender & CMNI	Step 3b Gender & GGI	Step 5 Gender, CMNI & GGI
(Intercept)	−1.26 (0.29)***	−0.78 (0.30)**	−2.09 (0.30)***	−1.53 (0.31)***
$\log_{10}$(age)	3.77 (0.40)***	3.26 (0.40)***	3.78 (0.39)***	3.33 (0.40)***
$\log_{10}(\text{age})^2$	−1.33 (0.13)***	−1.16 (0.13)***	−1.34 (0.13)***	−1.19 (0.13)***
Gender (man)	0.26 (0.01)***	0.28 (0.01)***	0.25 (0.01)***	0.28 (0.01)***
Winning (27r)		0.00 (0.00)		0.00 (0.00)
Self-Reliance (43)		0.01 (0.00)*		0.01 (0.00)*
Violence (41r)		−0.01 (0.00)		−0.00 (0.00)
Heterosexual Self-Presentation (24)		−0.03 (0.00)***		−0.02 (0.00)***
Risk Taking (8)		−0.02 (0.00)***		−0.02 (0.00)***
Emotional Control (25r)		0.00 (0.00)		0.00 (0.00)
Playboy (36)		−0.00 (0.00)		−0.00 (0.00)
GGI (country of childhood)			0.99 (0.21)***	0.83 (0.21)***
GGI (main residence in last 5 years)			0.16 (0.22)	0.13 (0.21)
R^2	0.125	0.143	0.138	0.152
$\bar{R}^2$	0.125	0.141	0.138	0.150
Num. obs.	6758	6758	6758	6758

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 7. Results of regression analyses for the Mental Rotations Test. The diagram above shows the model hierarchy and summarizes the results. Model in Step 3a is a significantly better fit than the model in Step 2a indicating that CMNI captured relevant information not already captured by binary gender. Adding GGI to the model that already included covariates and gender (Step 3b vs Step 2a) also significantly improved model fit. Standard errors in parentheses; pre-high school was the reference value for Education; native speaker was the reference for Comprehension. $\bar{R}^2$ denotes adjusted R^2.

	Estimate	[95% CIs]
Total effect	1.810	[1.053, 2.599]*
Direct effect	0.775	[-0.024, 1.585]
Total indirect effect	1.035	[0.742, 1.35]*
Heterosexual Self-presentation (24)	0.795	[0.58, 1.015]*
Risk Taking (8)	0.192	[0.094, 0.288]*
Violence (41r)	0.137	[0.023, 0.241]*
Playboy (36)	0.012	[-0.107, 0.128]
Winning (27r)	0.010	[-0.016, 0.035]
Self-Reliance (43)	0.009	[-0.012, 0.028]
Emotional Control (25r)	-0.103	[-0.224, 0.018]

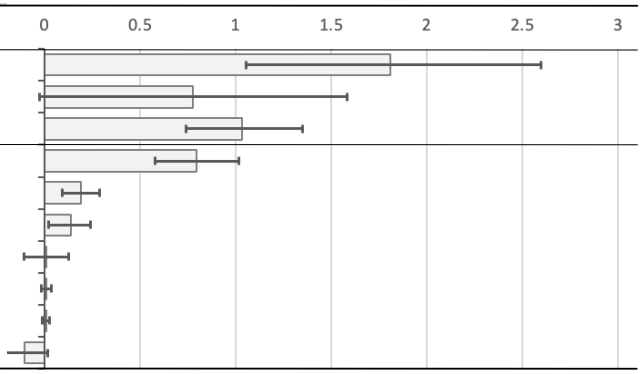
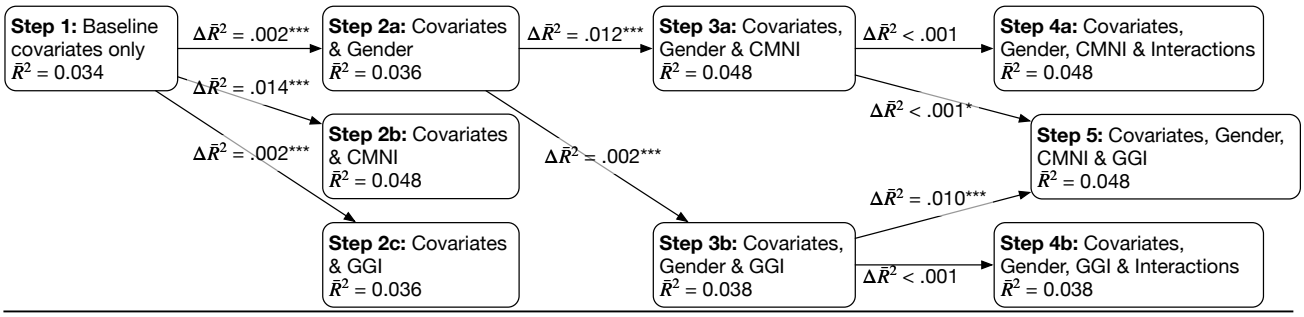


Table 8. Results of mediation analysis for the Spatial Working Memory Task. Men made on average 1.81 more attempts to complete the task than women (the total effect). The total indirect effect was significantly different from zero, indicating that the gender effect was mediated through aspects of gender socialization captured by CMNI. * $p < .05$. The 95% Confidence Intervals estimated using a casewise bootstrap.



	Step 2a Gender	Step 3a Gender & CMNI	Step 3b Gender & GGI	Step 5 Gender, CMNI & GGI
(Intercept)	195.43 (13.15)***	177.17 (13.46)***	210.93 (13.67)***	187.86 (14.09)***
$\log_{10}$(age)	−218.44 (17.94)***	−200.41 (18.16)***	−218.02 (17.95)***	−200.86 (18.17)***
$\log_{10}(\text{age})^2$	77.20 (6.04)***	71.43 (6.10)***	77.18 (6.04)***	71.64 (6.11)***
Gender (man)	1.59 (0.42)***	0.56 (0.44)	1.69 (0.42)***	0.67 (0.45)
Winning (27r)		0.16 (0.14)		0.16 (0.14)
Self-Reliance (43)		−0.16 (0.14)		−0.16 (0.14)
Violence (41r)		0.34 (0.14)*		0.33 (0.14)*
Heterosexual Self-Presentation (24)		0.96 (0.13)***		0.91 (0.13)***
Risk Taking (8)		0.61 (0.15)***		0.59 (0.15)***
Emotional Control (25r)		−0.20 (0.13)		−0.19 (0.13)
Playboy (36)		0.01 (0.14)		0.01 (0.14)
GGI (country of childhood)			−12.68 (9.47)	−6.19 (9.45)
GGI (main residence in last 5 years)			−9.21 (9.85)	−7.84 (9.80)
R^2	0.037	0.049	0.039	0.050
$\bar{R}^2$	0.036	0.048	0.038	0.048
Num. obs.	6764	6764	6764	6764

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 9. Results of regression analyses for the Spatial Working Memory Task. The diagram above shows the model hierarchy and summarizes the results. The key results are the significant differences between Steps 3a/3b and 2a, which demonstrate that both CMNI and GGI capture relevant information that goes beyond what is already captured by binary gender. Standard errors in parentheses; pre-high school was the reference value for Education; native speaker was the reference for Comprehension. $\bar{R}^2$ denotes adjusted R^2.

not include much relevant information that was not already captured by CMNI.

As shown in the model hierarchy diagram above Table 8, adding interaction terms between CMNI and binary gender ($\Delta R^2 < 0.001$; $F(7, 6746) = 1.01$; n.s) or between GGI and binary gender ($\Delta R^2 < 0.001$; $F(2, 6756) = 0.7834$; n.s) did not significantly improve model fits suggesting that men and women can be modeled together.

5.2.4. Additional Analyses

Because of a much smaller sample size for these tests compared to the Reading the Mind in the Eyes Test, we did not conduct the additional analyses for these tests.

6. Discussion

Prior research in HCI and cognitive psychology has demonstrated that men and women systematically perform differently on a range of cognitive tasks. We have reproduced three such tasks. Our results present several converging strands of evidence showing that quantitatively modeling gender socialization helps account for the observed differences.

First, in all three studies, a significant portion of the effect of binary gender was mediated through the individual differences in gender socialization as measured by a subset of the Conformity to the Masculine Norms Inventory (CMNI).

Further, in all three studies, when regression models of task performance included both the binary gender and CMNI, they explained significantly more variance in the task performance than models that did not include CMNI. These results suggest that there exist individual differences in gender socialization (as captured by CMNI) that are consequential for task performance and that are not captured when gender is represented as a categorical variable.

Lastly, we showed that society-level differences in attitudes toward women (as measured by the Gender Gap Index or GGI) were also significantly associated with how people from different countries performed on all three tasks. These results, combined with the statistically significant associations between GGI and CMNI observed in Study 1, add further evidence that socialization contributes to the development of consequential gender differences.

For two of the studies (Reading the Mind in the Eyes Test and the Spatial Working Memory Task), the observed effects were not only statistically significant but also meaningfully large. For those studies, more than half of the effect of binary gender was mediated through CMNI and the additional variance explained by CMNI and GGI in regression analyses was as large or larger than the variance explained by the binary gender variable. For the Mental Rotation Test, all the effects were relatively small and might be of limited scientific or practical significance.

Our results add support to previous research (e.g., Carothers and Reis (2013)) arguing that in some quantitative research gender may be best modeled as a multidimensional construct rather than a single categorical variable. Although Heterosexual Self-Presentation was primarily responsible for the mediation effects in all of our studies, in the regression analyses at least two distinct CMNI dimensions were significantly associated with the participants' performance in each of the studies.

This said, based on our results, we are not confident that CMNI is the right instrument to model gender socialization. There are two reasons for this. First, Heterosexual Self-Presentation—the most impactful CMNI construct in all three studies—is a more complex construct than it at first appears. Second, our analyses suggest that CMNI captures relevant gender-related experiences and attitudes of non-binary individuals differently from how it models binary men and women. These concerns impact the precision and universality of CMNI as a tool for modeling gender. We address each of these two concerns below.

Additionally, our preliminary analyses related to intersectional identities in the context of the first study yielded mixed results. On the one hand, when we disaggregated the data by national culture (11 countries) or when we disaggregated the data from US participants by race and ethnicity (5 groups), we found support for hypotheses H1 and H2 in all the subgroups. However, despite some indication that CMNI generalizes across race and ethnicity (Kivisalu et al., 2015; Hsu and Iwamoto, 2014) the degree to which CMNI mediated the effect of gender categories and the degree to which CMNI helped model participants' performance in the regression analyses varied considerably across these subgroups.

6.1. Heterosexual Self-Presentation

In all three studies, heterosexual self-presentation shows up as the strongest and most significant of the CMNI constructs we included in our questionnaire when modeling the behavior of people who identify with binary genders. Although heterosexual self-presentation seems to be a specific and narrow part of gender socialization and performance, it is in fact tied to several other gender norms and broader concepts. Most importantly, this factor is likely tied to anti-femininity (e.g., Wilkinson (2004)). Indeed, in the original CMNI development paper, Mahalik et al. (2003) originally called this construct "disdain for homosexuals" and found that this factor specifically related to anti-femininity and restrictive affectionate behavior between men. In "Homophobia Among Men," Lehne (1976) argues that homosexual activities are generally grouped with women's activities and interests, which suggests that (predominantly male) homosexuality is put in opposition to masculinity. Lehne argues that this is largely because emotion and affection between two men is seen as more feminine, since only women are allowed to be affectionate with each other in that way (platonically). Creating any sort of emotional bond with another man, then, painted a man as emotionally vulnerable and overly affectionate, and thus extremely feminine.

In this way, Lehne claims that homophobia is used to "maintain male roles" by reinforcing masculinity and appropriate types of relations between men who should be strong, invulnerable, and unemotional (Lehne, 1976). Similarly, in "Patterns of gender role conflict and strain: Sexism and fear of femininity in men's lives," O'Neil (1981) argues that the "fear of femininity is central to understanding male homophobia" and that when men fear their feminine side they are scared that people will see them as "stereotypically and negatively feminine (e.g., weak, dependent, submissive) rather than positively masculine". He argues that men and some women fear exposing their feminine side because it will result in their own devaluation.

This work allows us to begin to understand not just male homophobia but homophobia in general. While it may be intuitive that women who conform more to feminine norms would be more homophobic and more resistant to changing gender roles (especially in same-gender relationships), our results suggest that a majority of women are much less homophobic and much more comfortable with being perceived as not straight

than men. One potential explanation for this is that to be homophobic is to resist emotional and affectionate relationships with those of the same gender, per Lehne. In this sense, women who resist emotional relationships with other women, who might also fear exposing their feminine side because of the devaluation of those traits in our society, would begin to conform to more masculine norms of being less emotional, affectionate, vulnerable, etc. The ties between homophobia and emotional connections with those of the same gender also speak to the general conception that homosexuality aligns much more with femininity and feminine norms than with masculinity and masculine norms.

In conclusion, we are concerned that heterosexual self-presentation is a distant (and therefore noisy) symptom of a more fundamental set of attitudes. In future research, it may be valuable to replace it with several more granular and more precise constructs.

6.2. Difficulty in Modeling the Experiences of Non-Binary Individuals

Our results suggest that at least some of the CMNI items may capture different underlying constructs for non-binary individuals compared to individuals who identify with either of the binary genders. For example, when we compared how people of different genders scored on the 7 CMNI questions (Figure 3), we found that people who identified as men generally scored significantly higher (i.e., conformed more with masculine norms) than women. However, people who identified as non-binary sometimes scored significantly higher, sometimes scored significantly lower and sometimes scored in the same range as individuals identifying as either men or women.

Specifically, non-binary individuals reported significantly higher scores on the Self-Reliance item than either men or women, which could potentially be explained by the fact that non-binary people are subject to higher rates of physical and sexual assault as well as general harassment, loss of parental support, homelessness, unemployment, etc. (Liszewski et al., 2018). These experiences may result in non-binary people being—by necessity—more independent and self-reliant than people who have not experienced frequent discrimination. Thus, while people who identify as men may strive to be self-reliant to appear invulnerable and to limit their emotional expressiveness (Mahalik et al., 2003), non-binary individuals may be self-reliant as a result of their experience resisting discrimination and marginalization (Robinson and Schmitz, 2021). This argument may extend to people holding other marginalized identities as well.

Three other constructs (Heterosexual self-presentation, Playboy, and Winning) might also mean something entirely different for non-binary individuals compared to those who identify as either men or women. When someone identifies as outside of the gender binary, heterosexuality becomes a difficult concept to apply, because there is not just one "opposite" or "other" gender. Heterosexuality is based in the gender binary, therefore for those who identify as non-binary, it would make sense for them to have little concern about appearing queer in any way, since they most likely do not identify as heterosexual. Similarly, the queer community as a whole generally holds more open-minded views about non-monogamy (another way to "queer" the heteronormative relationship) (Carlström and Andersson, 2019), and as part of that community, it would make sense that non-binary people might be more open and supportive of that. With these two claims, it makes sense that in our results, non-binary people assigned much lower importance to Heterosexual Self-Presentation than either men and women, and that they expressed a stronger desire to have multiple relationships (reflected in the Playboy construct) than either men or women. Finally, non-binary people also scored lower than either men and women on the importance of Winning, which might tie into the construct of competition embedded into the idea of heterosexual, monogamous relationships (Lehne, 1976), which, again, may be much less valid or relevant to those who identify outside the gender binary.

These sorts of variance in the specific experiences of non-binary people compared to men and women highlight reasons why different constructs in CMNI might be more applicable or accurate in terms of capturing gender socialization for non-binary people than the constructs most significant for men and women. Additionally, because of the ties between non-binary gender identity and a non-straight sexual orientation, it is also possible that the membership in this group reflects more information than just gender identity (i.e., more specific lived experiences, sexuality, feelings about monogamy, etc.). This suggests that other groups within the categories of men and women might also have similarly different relationships to CMNI based on sexuality or other markers that suggest some sort of "queerness."

Further research would need to be done to conclude a more full explanation of these varying results, but it is important to continue to develop these sorts of questionnaires with an attempt to make questions the most universally applicable (i.e., perhaps excluding questions related to sexuality). Our results suggest that CMNI is not the universal way to capture gender socialization—as illustrated by the results in Section 4.2.5 where the regression models linking gender socialization to the performance on the Reading the Mind in the Eyes test were significantly different for non-binary individuals compared to those who self-identified as fitting within the gender binary. Instead, to fully capture gender socialization and conformity to gender norms, we need a different, more inclusive instrument that takes into account a wider variety of experiences and constructs.

6.3. Additional Observations

It is noteworthy that in all three studies, being socialized to conform to masculine norms resulted in poorer task performance. We observed this outcome even on the Mental Rotations Test where men generally perform better than women. On that test, identifying as a man was associated with better performance, but both the mediation analysis and the regression analyses showed that higher (i.e., more masculine) scores on Heterosexual Self-Presentation and on Risk-Taking resulted in *reduced* performance. While this result may appear counterintuitive, there exists prior research that found opposing effects of sex and gender (e.g., among older adults, female sex was found to be positively associated with cognitive performance while higher femininity scores were negatively associated with cognitive performance (Pohrt et al., 2022)). This finding, we believe, makes the value of explicitly modeling gender socialization even more apparent.

6.4. Limitations

Beyond the issues with CMNI as related to non-binary participants, the abbreviated CMNI version we used likely increased the measurement variance compared to the full CMNI-29 instrument and also potentially provided some limitations in the

level of nuance we were able to capture regarding participants' gender socialization. In our results, the gender socialization information captured by CMNI produced a much smaller effect on the score for the Mental Rotations Test than for the two other studies. While these results could point to the mental rotation requiring skills that are tied more to other factors than socialization, an alternative explanation is that the CMNI questionnaire we used was not comprehensive enough to capture all relevant aspects of gender socialization.

Our studies only captured information about a person's current gender. Thus, our results are not representative of people who transitioned genders or who are gender fluid.

Finally, while there is evidence that translated versions of CMNI tested across some cultures remained valid (e.g., French Canadian and Russian (Krivoshchekov et al., 2022; Jbilou et al., 2021)), the Chinese version did not reach the same level of reliability as the Western ones (Rochelle and Yim, 2015). Thus, it is possible that CMNI itself does not capture notions of masculinity equally well across all contemporary cultures or that it will not remain informative in the future.

7. Conclusion and Implications for Quantitative Research in HCI

Our results demonstrate that gender socialization can measurably impact human performance on tasks related to such basic cognitive skills as theory of mind and spatial working memory. Specifically, given that the effects of categorical gender in some of our studies were substantially mediated through gender socialization, it suggests that differences in task performance associated with gender categories may not be universal. People from different cultures today, or future generations, may be socialized differently and may, therefore, not exhibit the same performance differences as today's participants from one particular culture. It may be more robust, therefore, to use specific gender-related constructs—rather than broad gender categories—to construct our models and theories of how gender impacts behavior.

In our studies, we used CMNI as a set of measures of different dimensions of gender socialization. This choice was valuable in our work because it helped make apparent the link between socialization and performance on seemingly unrelated tasks. CMNI is not necessarily the best choice for all projects, however. First, our results demonstrated that CMNI may not support inclusive modeling of participants of all genders. Second, the CMNI constructs may not be scientifically the most informative for every research question. For example, in the GenderHCI work summarized in Section 2.2, other gender-related constructs such as self-efficacy, motivations, attitude toward risk, and information processing strategies were identified in the literature as relevant to people's effective use of software tools. Using these theoretically-grounded dimensions instead of gender categories is allowing the authors to construct more accurate and more inclusive theories for how to design software that effectively supports many different types of users. However, in areas where strong theoretical foundations do not yet exist, measures of gender socialization (ideally more generally applicable than CMNI) can be a good starting point.

8. Positionality Statement

The authors span a range of perspectives on sexuality and gender.

NH is a white, queer cis woman from an upper-middle class background. She has a joint major in Women, Gender, and Sexuality studies as well as Computer Science, and she is passionate about exploring the intersection between the two, as in this study. She specializes in the influence gender has on technology and technological systems, such as voice assistants, though she is new to the research field of human-computer interaction.

LWM, a white, straight cis man from an upper-middle class intellectual background whose mother came out as openly gay when he was five, was not raised or socialized as traditionally male, and has long been interested in gender equity and how gender roles play out in US society. He has predominately lived in communities with high proportions of people not identifying with the gender binary or conventional heterosexuality, giving him a deeper understanding of these different experiences than may be typical for a straight cis male.

KZG is a straight cis man and an immigrant. He is a quantitative HCI researcher based at a US university. One of his areas of focus is accessible computing, where his work is founded on the understanding that disability and handicap arise from an interaction between an individual's medical condition and the surrounding societal and environmental factors. Thus, while he is a relative newcomer to gender-related scholarship, he joined this project already committed to the idea that societal factors can shape one's identity and experiences. He is also committed to critical technical practice (conceptualized as constantly and explicitly naming and examining core assumptions of one's intellectual endeavors) and has worked on several projects that questioned, evaluated, or offered alternatives to existing research practices in HCI.

We have worked to read related literature, talk with colleagues, and generally understand the broader context of gender and sexuality research as we have dug into understanding how these features play out in the field of HCI, but we recognize our backgrounds and experiences may give us blind spots regarding how we chose to model and interpret our results. We nevertheless hope our work furthers the conversation and brings attention to these important themes in our corner of the academic research enterprise.

9. Data Availability

The data underlying this article, as well as the R code used for the analyses, are available at Harvard Dataverse, at `https://doi.org/10.7910/DVN/UJK8UU`.

10. Competing interests

No competing interest is declared.

11. Author contributions statement

N.H. conceived of the project. N.H. and K.Z.G. designed the data collection studies and drafted the initial manuscript. K.Z.G. implemented the data collection for Study 1. N.H. implemented the data collection for Study 2. L.W.M. and K.Z.G. designed the strategies for data analysis. K.Z.G. conducted the data analysis. All authors contributed to the revisions of the manuscript.

12. Acknowledgments

The authors thank Priyanka Nanayakkara, Simone Stumpf, Naveena Karusala, and the anonymous reviewers for their valuable suggestions.

A. Intersectionality Analyses for Study 1

This appendix presents the detailed intersectionality results discussed in Section 4.2.4.

	Asian or Asian American	Black or African American	Latino / Latina or Hispanic	Latino / Latina or Hispanic,Whit	White
	est [95% CI]	est [95% CI]	est [95% CI]	est [95% CI]	est [95% CI]
Total effect	-0.890 [-1.094, -0.751]	-0.377 [-0.636, -0.195]	-0.756 [-0.915, -0.579]	-0.443 [-0.699, -0.204]	-0.648 [-0.722, -0.582]
Direct effect	-0.536 [-0.727, -0.382]	0.050 [-0.209, 0.262]	-0.376 [-0.540, -0.197]	-0.083 [-0.371, 0.173]	-0.341 [-0.424, -0.268]
Total indirect effect	-0.355 [-0.428, -0.307]	-0.427 [-0.551, -0.334]	-0.380 [-0.444, -0.313]	-0.360 [-0.463, -0.242]	-0.307 [-0.333, -0.278]
Winning (27r)	-0.004 [-0.015, 0.005]	0.027 [0.007, 0.038]	0.031 [0.014, 0.042]	0.016 [-0.008, 0.040]	0.020 [0.014, 0.025]
Self-Reliance (43)	-0.004 [-0.008, 0.001]	-0.022 [-0.051, -0.009]	-0.020 [-0.031, -0.008]	0.002 [-0.009, 0.015]	-0.001 [-0.002, 0.000]
Violence (41r)	-0.006 [-0.022, 0.009]	-0.023 [-0.044, -0.004]	0.001 [-0.015, 0.021]	0.010 [-0.025, 0.048]	0.006 [-0.005, 0.020]
Heterosexual Self-presentation (24)	-0.300 [-0.358, -0.257]	-0.308 [-0.384, -0.223]	-0.216 [-0.272, -0.167]	-0.253 [-0.333, -0.171]	-0.188 [-0.206, -0.168]
Risk Taking (8)	-0.015 [-0.036, 0.005]	-0.064 [-0.104, -0.033]	-0.063 [-0.088, -0.039]	-0.030 [-0.061, -0.003]	-0.049 [-0.057, -0.039]
Emotional Control (25r)	-0.015 [-0.031, 0.001]	-0.040 [-0.066, -0.013]	-0.055 [-0.074, -0.035]	-0.073 [-0.106, -0.034]	-0.099 [-0.112, -0.088]
Playboy (36)	-0.013 [-0.037, 0.008]	-0.001 [-0.045, 0.046]	-0.052 [-0.085, -0.021]	-0.032 [-0.059, -0.001]	0.004 [-0.003, 0.012]
N	10018	5526	10071	3326	57867

Table 10. Results of mediation analyses for the Reading the Mind in the Eyes test disaggregated by race and ethnicity.

	Australia	Brazil	Canada	France	Germany	India
	est [95% CI]	est [95% CI]	est [95% CI]	est [95% CI]	est [95% CI]	est [95% CI]
Total	-0.52 [-0.695, -0.344]	-0.447 [-0.776, -0.125]	-0.536 [-0.685, -0.370]	-0.81 [-1.148, -0.505]	-0.615 [-0.856, -0.362]	-0.995 [-1.182, -0.769]
Direct	-0.109 [-0.294, 0.076]	-0.295 [-0.656, 0.051]	-0.144 [-0.307, 0.033]	-0.558 [-0.914, -0.232]	-0.247 [-0.497, 0.007]	-0.526 [-0.729, -0.286]
Indirect	-0.411 [-0.482, -0.338]	-0.152 [-0.250, -0.046]	-0.392 [-0.465, -0.316]	-0.252 [-0.397, -0.109]	-0.368 [-0.456, -0.272]	-0.469 [-0.571, -0.364]
Winning (27r)	0.013 [0.002, 0.021]	0 [-0.007, 0.007]	0.013 [0.004, 0.025]	0.005 [-0.015, 0.018]	0.007 [-0.003, 0.017]	-0.121 [-0.153, -0.085]
Self-Reliance (43)	-0.003 [-0.013, 0.004]	-0.003 [-0.017, 0.011]	-0.001 [-0.003, 0.002]	-0.001 [-0.013, 0.010]	-0.003 [-0.012, 0.004]	0 [-0.007, 0.007]
Violence (41r)	-0.054 [-0.084, -0.022]	0.009 [-0.008, 0.024]	-0.004 [-0.035, 0.022]	0.003 [-0.023, 0.032]	-0.013 [-0.039, 0.018]	-0.003 [-0.011, 0.006]
Heterosexual Self-presentation (24)	-0.234 [-0.282, -0.184]	-0.112 [-0.181, -0.036]	-0.229 [-0.282, -0.181]	-0.098 [-0.203, 0.010]	-0.232 [-0.304, -0.156]	-0.192 [-0.261, -0.124]
Risk Taking (8)	-0.035 [-0.055, -0.013]	-0.031 [-0.065, -0.003]	-0.065 [-0.088, -0.047]	-0.043 [-0.076, -0.012]	-0.067 [-0.089, -0.032]	-0.106 [-0.141, -0.067]
Emotional Control (25r)	-0.091 [-0.124, -0.060]	-0.012 [-0.048, 0.030]	-0.082 [-0.115, -0.050]	-0.088 [-0.163, -0.012]	-0.073 [-0.116, -0.030]	0.009 [-0.008, 0.024]
Playboy (36)	-0.004 [-0.027, 0.019]	-0.002 [-0.041, 0.039]	-0.019 [-0.038, 0.005]	-0.033 [-0.086, 0.025]	0.008 [-0.019, 0.032]	-0.06 [-0.108, -0.007]
N	9260	2533	9312	1992	4662	6477

	Netherlands	Philippines	Russian Federation	United Kingdom	Unied States
	est [95% CI]	est [95% CI]	est [95% CI]	est [95% CI]	est [95% CI]
Total	-0.521 [-0.757, -0.217]	-0.475 [-0.669, -0.255]	-0.72 [-1.077, -0.365]	-0.413 [-0.555, -0.282]	-0.577 [-0.632, -0.528]
Direct	-0.005 [-0.223, 0.334]	-0.116 [-0.327, 0.126]	-0.558 [-0.966, -0.146]	-0.092 [-0.235, 0.042]	-0.249 [-0.306, -0.199]
Indirect	-0.516 [-0.709, -0.376]	-0.36 [-0.443, -0.280]	-0.162 [-0.298, -0.032]	-0.321 [-0.369, -0.275]	-0.327 [-0.344, -0.310]
Winning (27r)	0.017 [-0.012, 0.048]	-0.034 [-0.060, -0.010]	-0.022 [-0.036, 0.010]	0.008 [-0.002, 0.017]	0.021 [0.017, 0.024]
Self-Reliance (43)	-0.025 [-0.050, -0.002]	-0.005 [-0.019, 0.009]	-0.001 [-0.013, 0.010]	0 [-0.003, 0.001]	-0.002 [-0.003, -0.001]
Violence (41r)	-0.006 [-0.041, 0.029]	0.005 [-0.003, 0.013]	0.007 [-0.029, 0.044]	-0.001 [-0.022, 0.022]	-0.006 [-0.014, 0.002]
Heterosexual Self-presentation (24)	-0.284 [-0.416, -0.182]	-0.235 [-0.296, -0.166]	-0.13 [-0.212, -0.035]	-0.141 [-0.175, -0.106]	-0.2 [-0.214, -0.186]
Risk Taking (8)	-0.122 [-0.191, -0.059]	-0.032 [-0.047, -0.011]	-0.036 [-0.090, 0.008]	-0.021 [-0.040, -0.005]	-0.044 [-0.050, -0.038]
Emotional Control (25r)	-0.117 [-0.178, -0.059]	0.011 [-0.002, 0.021]	-0.012 [-0.069, 0.041]	-0.171 [-0.199, -0.143]	-0.085 [-0.094, -0.077]
Playboy (36)	0.02 [-0.034, 0.066]	-0.076 [-0.104, -0.048]	0.024 [-0.037, 0.076]	0.007 [-0.014, 0.026]	-0.011 [-0.017, -0.005]
N	2713	6624	2224	17323	99158

Table 11. Results of mediation analyses for the Reading the Mind in the Eyes test disaggregated by country.

		Asian or Asian American	Black or African American	Latino / Latina or Hispanic	Latino / Latina or Hispanic and White	White
R^2	Step 1 (Covariates only)	0.077	0.074	0.076	0.068	0.061
R^2	Step 2a (Covariates and gender)	0.088	0.076	0.084	0.072	0.068
R^2	Step 2b (Covariates and CMNI)	0.099	0.093	0.101	0.093	0.082
R^2	Step 3a (Covariates, Gender and CMNI)	0.103	0.093	0.103	0.093	0.083
ΔR^2	Step 2a - Step 1	0.011	0.002	0.008	0.003	0.007
ΔR^2	Step 2b - Step 1	0.023	0.019	0.025	0.024	0.020
ΔR^2	Step 3a - Step 2a	0.015	0.018	0.019	0.021	0.015
	N	9593	5275	9690	3267	56841

Table 12. Results of regression analyses for the Reading the Mind in the Eyes test disaggregated by race and ethnicity.

		Australia	Brazil	Canada	France	Germany	India	Netherlands	Philippines	Russian Federation	United Kingdom	United States
R^2	Step 1 (Covariates only)	0.124	0.043	0.098	0.026	0.087	0.068	0.049	0.027	0.059	0.052	0.075
R^2	Step 2a (Covariates and gender)	0.127	0.046	0.102	0.037	0.093	0.078	0.053	0.030	0.066	0.054	0.080
R^2	Step 2b (Covariates and CMNI)	0.146	0.051	0.118	0.045	0.107	0.101	0.076	0.050	0.068	0.072	0.095
R^2	Step 3a (Covariates, Gender and CMNI)	0.146	0.052	0.118	0.048	0.108	0.103	0.076	0.050	0.072	0.072	0.095
ΔR^2	Step 2a - Step 1	0.003	0.003	0.004	0.011	0.005	0.011	0.003	0.003	0.008	0.003	0.005
ΔR^2	Step 2b - Step 1	0.022	0.008	0.020	0.018	0.020	0.033	0.027	0.022	0.010	0.020	0.020
ΔR^2	Step 3a - Step 2a	0.019	0.006	0.016	0.011	0.015	0.025	0.023	0.020	0.006	0.017	0.016
	N	9260	2533	9312	1992	4662	6477	2713	6624	2224	17323	99158

Table 13. Results of regression analyses for the Reading the Mind in the Eyes test disaggregated by country.

References

G. M. Alexander, M. G. Packard, and B. S. Peterson. Sex and spatial position effects on object location memory following intentional learning of object identities. *Neuropsychologia*, 40(8):1516–1522, 2002. URL https://doi.org/10.1016/S0028-3932(01)00215-9.

T. August and K. Reinecke. Pay attention, please: Formal language improves attention in volunteer and paid online experiments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–11, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300478. URL https://doi.org/10.1145/3290605.3300478.

C. J. Auster and S. C. Ohm. Masculinity and femininity in contemporary american society: A reevaluation using the bem sex-role inventory. *Sex roles*, 43:499–528, 2000.

A. M. Barnfield. Development of sex differences in spatial memory. *Perceptual and Motor Skills*, 89(1):339–350, 1999. URL https://doi.org/10.2466/pms.1999.89.1.339.

S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry, and allied disciplines*, 42 (2):241–251, Feb. 2001.

L. Beckwith, M. Burnett, S. Wiedenbeck, C. Cook, S. Sorte, and M. Hastings. Effectiveness of end-user debugging software features: are there gender issues? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 869–878, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1581139985. doi: 10.1145/1054972.1055094. URL https://doi.org/10.1145/1054972.1055094.

L. Beckwith, M. Burnett, V. Grigoreanu, and S. Wiedenbeck. Gender hci: What about the software? *Computer*, 39(11): 97–101, 2006.

S. L. Bem. Bem sex role inventory. *Journal of Personality and Social Psychology*, 1981.

K. E. Boerner, C. T. Chambers, J. Gahagan, E. Keogh, R. B. Fillingim, and J. S. Mogil. Conceptual complexity of gender and its relevance to pain. *Pain*, 159(11):2137–2141, 2018.

M. A. Borkin, C. S. Yeh, M. Boyd, P. Macko, K. Gajos, M. Seltzer, and H. Pfister. Evaluation of filesystem provenance visualization tools. *IEEE transactions on visualization and computer graphics*, 19(12):2476–2485, 2013.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan. Gendermag: A method for evaluating software's gender inclusiveness. *Interacting with computers*, 28(6):760–787, 2016.

J. Butler. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre Journal*, 40 (4):519–531, 1988. ISSN 01922882, 1086332X. URL http://www.jstor.org/stable/3207893.

Canadian Institutes of Health Research. What is gender? what is sex?, 2023. URL https://cihr-irsc.gc.ca/e/48642.html.

C. Carlström and C. Andersson. Living outside protocol: Polyamorous orientations, bodies, and queer temporalities. *Sexuality & Culture*, 23(4):1315–1331, 2019.

B. J. Carothers and H. T. Reis. Men and women are from earth: examining the latent structure of gender. *Journal of personality and social psychology*, 104(2):385, 2013.

E. Chapman, S. Baron-Cohen, B. Auyeung, R. Knickmeyer, K. Taylor, and G. Hackett. Fetal testosterone and empathy: evidence from the empathy quotient (eq) and the "reading the mind in the eyes" test. *Social Neuroscience,*, 1(2):135–148, 2006.

M. Czerwinski, D. S. Tan, and G. G. Robertson. Women take a wider view. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 195–202, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134533. doi: 10.1145/503376.503412. URL https://doi.org/10.1145/503376.503412.

G. Domes, M. Heinrichs, A. Michel, C. Berger, and S. C. Herpertz. Oxytocin improves "mind-reading" in humans. *Biological psychiatry*, 61(6):731–733, 2007.

S. J. Duff and E. Hampson. A sex difference on a novel spatial working memory task in humans. *Brain and Cognition*, 47 (3):470–493, 2001. URL https://doi.org/10.1006/brcg.2001.1326.

A. H. Eagly and W. Wood. The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, 8(3):340–357, 2013.

N. M. Else-Quest and S. Grabe. The Political Is Personal: Measurement and Application of Nation-Level Indicators of Gender Equity in Psychological Research. *Psychology of Women Quarterly*, 36(2):131–144, 2012. ISSN 03616843. doi: 10.1177/0361684312441592.

Z. Estes and S. Felker. Confidence mediates the sex difference in mental rotation performance. *Archives of sexual behavior*, 41:557–570, 2012. doi: 10.1007/s10508-011-9875-5.

A. Fausto-Sterling. *Sexing the Body: Gender Politics and the Construction of Sexuality*, chapter Of gender and genitals: The use and abuse of the modern intersexual, pages 45–77. Basic Books, 2000.

K. Z. Gajos and K. Chauncey. The influence of personality traits and cognitive load on the use of adaptive user interfaces. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, pages 301–306, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4348-0. doi: 10.1145/3025171.3025192. URL http://doi.acm.org/10.1145/3025171.3025192.

K. Z. Gajos, K. Reinecke, M. Donovan, C. D. Stephen, A. Y. Hung, J. D. Schmahmann, and A. S. Gupta. Computer mouse use captures ataxia and parkinsonism, enabling accurate measurement and detection. *Movement Disorders*, 35:354–358, February 2020. URL https://doi.org/10.1002/mds.27915.

D. C. Geary, S. J. Saults, F. Liu, and M. K. Hoard. Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, 77(4):337–353, 2000. URL https://doi.org/10.1006/jecp.2000.2594.

C. Geiser, W. Lehmann, and M. Eid. A note on sex differences in mental rotation in different age groups. *Intelligence*, 36 (6):556–563, 2008. URL https://doi.org/10.1016/j.intell.2007.12.003.

L. T. Germine, B. Duchaine, and K. Nakayama. Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, 118(2):201–210, Feb. 2011.

D. Gold and D. Andres. Comparisons of adolescent children with employed and nonemployed mothers. *Merrill-Palmer*

Quarterly of Behavior and Development, 24(4):243–254, 1978.

D. M. Greenberg, V. Warrier, A. Abu-Akel, C. Allison, K. Z. Gajos, K. Reinecke, P. J. Rentfrow, M. A. Radecki, and S. Baron-Cohen. Sex and age differences in "theory of mind" across 57 countries using the english version of the "reading the mind in the eyes" test. *Proceedings of the National Academy of Sciences*, 120(1):e2022385119, 2023.

V. Grigoreanu, J. Cao, T. Kulesza, C. Bogart, K. Rector, M. Burnett, and S. Wiedenbeck. Can feature design reduce the gender gap in end-user software development environments? In *2008 IEEE symposium on visual languages and human-centric computing*, pages 149–156. IEEE, 2008.

M. Guizani, I. Steinmacher, J. Emard, A. Fallatah, M. Burnett, and A. Sarma. How to debug inclusivity bugs? a debugging process with information architecture. In *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society*, ICSE-SEIS '22, pages 90–101, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392273. doi: 10.1145/3510458.3513009. URL https://doi.org/10.1145/3510458.3513009.

J. H. Hammer, P. J. Heath, and D. L. Vogel. Fate of the total score: Dimensionality of the conformity to masculine norms inventory-46 (cmni-46). *Psychology of Men & Masculinity*, 19(4):645, 2018.

B. Hassan and Q. Rahman. Selective sexual orientation-related differences in object location memory. *Behavioral Neuroscience*, 121(3):625–633, 2007. URL https://doi.org/10.1037/0735-7044.121.3.625.

L. W. Hoffman. Effects of maternal employment on the child: A review of the research. *Developmental Psychology*, 10(2): 204–228, 1974.

R. M. Hoffman and L. D. Borders. Twenty-five years after the bem sex-role inventory: A reassessment and new issues regarding classification variability. *Measurement and evaluation in counseling and development*, 34(1):39–55, 2001.

S. Horstmann, C. Schmechel, K. Palm, S. Oertelt-Prigione, and G. Bolte. The operationalisation of sex and gender in quantitative health–related research: a scoping review. *International Journal of Environmental Research and Public Health*, 19(12):7493, 2022.

K. Hsu and D. K. Iwamoto. Testing for measurement invariance in the conformity to masculine norms-46 across white and asian american college men: Development and validity of the cmni-29. *Psychology of men & masculinity*, 15(4):397, 2014. doi: 10.1037/a0034548.

L. Hu and I. Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, jan 2020. doi: 10.1145/3351095.3375674. URL https://doi.org/10.1145%2F3351095.3375674.

J. Huang, S. Kumar, and C. Hu. Gender differences in motivations for identity reconstruction on social network sites. *International Journal of Human–Computer Interaction*, 34 (7):591–602, 2018.

B. Huber and K. Z. Gajos. Conducting online virtual environment experiments with uncompensated, unsupervised samples. *Plos one*, 15(1):e0227629, 2020.

A. H.-C. Hwang and A. S. Won. The sound of support: Gendered voice agent as support to minority teammates in gender-imbalanced team. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642202. URL https://doi.org/10.1145/3613904.3642202.

A. B. Jannsen and C. Geiser. Cross-cultural differences in spatial abilities and solution strategies - an investigation in cambodia and germany. *Journal of Cross-Cultural Psychology*, 43(4):553–557, 2011. URL https://doi.org/10.1177/0022022111399646.

S. Jaroszewski, D. Lottridge, O. L. Haimson, and K. Quehl. " genderfluid" or" attack helicopter" responsible hci research practice with non-binary gender variation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2018.

J. Jbilou, N. Levesque, R.-P. Sonier, P. J. Tully, I. Pinette-Drapeau, V. Sonier, A. Charbonneau, P. S. Greenman, J. Grenier, and M.-H. Chomienne. Canadian french translation and preliminary validation of the conformity to masculine norms inventory: A pilot study. *American Journal of Men's Health*, 15(6):15579883211057391, 2021.

Y. Jing, S. Jing, C. Huajian, S. Chuangang, and L. Yan. The gender difference in distraction of background music and noise on the cognitive task performance. In *2012 8th International Conference on Natural Computation*, pages 584–587. IEEE, 2012.

S. Kachel, M. C. Steffens, and C. Niedlich. Traditional masculinity and femininity: Validation of a new scale assessing gender roles. *Frontiers in Psychology*, 7:956, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.00956.

O. Keyes, C. May, and A. Carrell. You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW1):1–23, 2021.

D. C. Kidd and E. Castano. Reading literary fiction improves theory of mind. *Science*, 342(6156):377–380, 2013.

A. M. Kimbrough, R. E. Guadagno, N. L. Muscanell, and J. Dill. Gender differences in mediated communication: Women connect more than do men. *Computers in Human Behavior*, 29(3):896–900, 2013.

R. A. Kirkland, E. Peterson, C. A. Baker, S. Miller, and S. Pulos. Meta-analysis reveals adult female superiority in 'reading the mind in the eyes' test. *North American Journal of Psychology*, 15(1):121–146, 2013.

T. M. Kivisalu, C. King, C. E. Phillips, and S. K. O'Toole. Reliability generalization of the conformity to masculine norms inventory (cmni). *Race, Gender, and Class*, 22(1):173–188, 2015.

D. Kloo and J. Perner. Training theory of mind and executive control: A tool for improving school achievement? *Mind, Brain, and Education*, 2(3):122–127, 2008. ISSN 17512271. doi: 10.1111/j.1751-228X.2008.00042.x.

V. Krivoshchekov, O. Gulevich, and M. Ostroverkhova. The conformity to masculine norms inventory-30: Validity and measurement invariance of a russian-language version. *Psychology of Men & Masculinities*, 23(1):59, 2022.

G. K. Lehne. Homophobia among men. *The forty-nine percent majority: The male sex role*, pages 66–88, 1976.

L. Lejbak, M. Vrbancic, and M. Crossley. The female advantage in object location memory is robust to verbalizability and mode of presentation of test stimuli. *Brain and Cognition*, 69 (1):148–153, 2009. URL https://doi.org/10.1016/j.bandc.2008.06.006.

L. J. Levy, R. S. Astur, and K. M. Frick. Men and women differ in object memory but not performance of virtual radial maze. *Behavioral Neuroscience*, 119(4):853–862, 2005. URL

https://doi.org/10.1037/0735-7044.119.4.853.

Q. Li, K. Z. Gajos, and K. Reinecke. Volunteer-based online studies with older adults and people with disabilities. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, pages 229–241, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5650-3. doi: 10.1145/3234695.3236360.

Q. Li, S. J. Joo, J. D. Yeatman, and K. Reinecke. Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. *Scientific Reports*, 10(1):1–11, 2020.

W. Liszewski, J. K. Peebles, H. Yeung, and S. Arron. Persons of nonbinary gender—awareness, visibility, and health disparities. *The New England journal of medicine*, 379(25): 2391, 2018.

H. Lytton and D. M. Romney. Parents' differential socialization of boys and girls: A meta-analysis. *Psychological bulletin*, 109(2):267–296, 1991.

D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz. Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614, 2007.

J. R. Mahalik, B. D. Locke, L. H. Ludlow, and M. A. Diemer. Development of the conformity to masculine norms inventory. *Psychology of Men & Masculinity*, 4(1):3–25, 2003. URL https://doi.org/10.1037/1524-9220.4.1.3.

A. Mahmood and C.-M. Huang. Gender biases in error mitigation by voice assistants. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), Apr. 2024. doi: 10.1145/3637337. URL https://doi.org/10.1145/3637337.

M. S. Masters. The gender difference on the mental rotations test is not due to performance factors. *Memory & Cognition*, 26(3):444–448, 1998. URL https://doi.org/10.3758/BF03201154.

D. H. McBurney, S. J. Gaulin, T. Devineni, and C. Adams. Superior spatial memory of women: Stronger evidence for the gathering hypothesis. *Evolution and Human Behavior*, 18(3):165–174, 1997. URL https://doi.org/10.1016/S1090-5138(97)00001-9.

L. McCall. The complexity of intersectionality. *Signs: Journal of women in culture and society*, 30(3):1771–1800, 2005.

S. M. McHale, K. A. Updegraff, H. Helms-Erikson, and A. C. Crouter. Sibling influences on gender development in middle childhood and early adolescence: A longitudinal study. *Developmental Psychology*, 37(1):115–125, 2001.

S. M. McHale, A. C. Crouter, and S. D. Whiteman. The family contexts of gender development in childhood and adolescence. *Social development*, 12(1):125–148, 2003.

W. Mischel. A social-learning view of sex differences in behavior. In E. E. Maccoby, editor, *The development of sex differences*, pages 57–81. Stanford University Press, Stanford, CA, 1966.

B. Montagne, R. P. Kessels, E. Frigerio, E. H. de Haan, and D. I. Perrett. Sex differences in the perception of affective facial expressions: Do men really lack emotional sensitivity? *Cognitive Processing*, 6(2):136–141, 2005. URL https://doi.org/10.1007/s10339-005-0050-6.

J. Morawski. The Troubled Quest for Masculinity, Femininity, and Androgyny. In P. R. Shaver and C. Hendrick, editors, *Sex and gender*, pages 44–69. Sage, Newbury, CA, US, 1987.

M. Morgan. Television and adolescents' sex role stereotypes: A longitudinal study. *Journal of Personality and Social Psychology*, 43(5):947, 1982.

M. Morgan. Television, sex-role attitudes, and sex-role behavior. *The Journal of Early Adolescence*, 7(3):269–282, 1987.

T. Morgenroth and M. K. Ryan. The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. *Perspectives on Psychological Science*, 16(6):1113–1142, 2021.

A. Nazareth, A. Herrera, and S. M. Pruden. Explaining sex differences in mental rotation: role of spatial activity experience. *Cognitive Process*, 14(2):201–204, 2013. URL https://doi.org/10.1007/s10339-013-0542-8.

N. Neave, C. Hamilton, L. Hutton, N. Tildesley, and A. T. Pickering. Some evidence of a female advantage in object location memory using ecologically valid stimuli. *Human Nature*, 16(2):146–163, 2005. URL https://doi.org/10.1007/s12110-005-1001-8.

N. Newcombe, M. M. Bandura, and D. G. Taylor. Sex differences in spatial ability and spatial activities. *Sex roles*, 9: 377–386, 1983.

M. W. Nielsen, M. L. Stefanick, D. Peragine, T. B. Neilands, J. P. Ioannidis, L. Pilote, J. J. Prochaska, M. R. Cullen, G. Einstein, I. Klinge, et al. Gender-related variables for health research. *Biology of Sex Differences*, 12:1–16, 2021.

J. M. O'Neil. Patterns of gender role conflict and strain: Sexism and fear of femininity in men's lives. *The personnel and guidance journal*, 60(4):203–210, 1981.

M. C. Parent and B. Moradi. Confirmatory factor analysis of the conformity to masculine norms inventory and development of the conformity to masculine norms inventory-46. *Psychology of Men & Masculinity*, 10(3):175–189, 2009. URL https://doi.org/10.1037/a0015481.

M. C. Parent and A. P. Smiler. Metric invariance of the conformity to masculine norms inventory-46 among women and men. *Psychology of Men & Masculinity*, 14(3):324–328, 2012. URL https://doi.org/10.1037/a0027642.

E. J. Pedhazur and T. J. Tetenbaum. Bem sex role inventory: A theoretical and methodological critique. *Journal of Personality and Social psychology*, 37(6):996, 1979.

R. Pelletier, N. A. Khan, J. Cox, S. S. Daskalopoulou, M. J. Eisenberg, S. L. Bacon, K. L. Lavoie, K. Daskupta, D. Rabi, K. H. Humphries, C. M. Norris, G. Thanassoulis, H. Behlouli, L. Pilote, and null null. Sex versus gender-related characteristics. *Journal of the American College of Cardiology*, 67(2):127–135, 2016. doi: 10.1016/j.jacc.2015.10.067. URL https://www.jacc.org/doi/abs/10.1016/j.jacc.2015.10.067.

M. Peters. Sex differences and the factor of time in solving vandenberg and kuse mental rotation problems. *Brain and Cognition*, 57(2):176–184, 2005. URL https://doi.org/10.1016/j.bandc.2004.08.052.

M. Peters, B. Laeng, K. Latham, M. Jackson, R. Zaiyouna, and C. Richardson. A redrawn vandenberg and kuse mental rotations test: Different versions and factors that affect performance. *Brain and Cognition*, 28(1):39–58, 1995. URL https://doi.org/10.1006/brcg.1995.1032.

A. Pohrt, F. Kendel, I. Demuth, J. Drewelies, T. Nauman, H. Behlouli, G. Stadler, L. Pilote, V. Regitz-Zagrosek, and D. Gerstorf. Differentiating sex and gender among older men and women. *Psychosomatic Medicine*, 84(3):339–346, 2022.

Y. A. Rankin and J. O. Thomas. Straighten up and fly right: rethinking intersectionality in hci research. *Interactions*, 26 (6):64–68, Oct. 2019. ISSN 1072-5520. doi: 10.1145/3363033. URL https://doi.org/10.1145/3363033.

K. Reinecke and K. Z. Gajos. LabintheWild: conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1364–1378, New York, NY, USA, 2015. ACM.

ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675246.

B. A. Robinson and R. M. Schmitz. Beyond resilience: Resistance in the lives of lgbtq youth. *Sociology compass*, 15(12): e12947, 2021.

T. L. Rochelle and K. Yim. Assessing the factor structure of the chinese conformity to masculine norms inventory. *The Journal of psychology*, 149(1):29–41, 2015.

J. A. Rode. A theoretical agenda for feminist HCI. *Interacting with Computers*, 23(5):393–400, 2011.

S. P. Ross, R. W. Skelton, and S. C. Mueller. Gender differences in spatial navigation in virtual space: implications when using virtual environments in instruction and assessment. *Virtual Reality*, 10(3-4):175–184, 2006.

C. D. Russell and J. B. Ellis. Sex-role development in single parent households. *Social Behavior and Personality: an international journal*, 19(1):5–9, 1991.

M. Sakamoto and M. V. Spiers. Sex and cultural differences in spatial performance between japanese and north americans. *Archives of Sexual Behavior*, 43(3):483–491, 2014. URL `https://doi.org/10.1007/s10508-013-0232-8`.

D. M. Saucier, D. R. McCreary, and J. K. Saxberg. Does gender role socialization mediate sex differences in mental rotations? *Personality and Individual Differences*, 32(6): 1101–1111, 2002.

N. A. Scott and J. Siltanen. Intersectionality and quantitative methods: Assessing regression from a feminist perspective. *International Journal of Social Research Methodology*, 20 (4):373–385, 2017.

N. Signorielli. Television's gender role images and contribution to stereotyping: Past, present, future. In D. G. Singer and J. L. Singer, editors, *Handbook of children and the media*, pages 341–358. Sage, Thousand Oaks, CA, 2001.

A. P. Smiler and M. Epstein. Measuring gender: Options and issues. In *Handbook of gender research in psychology*, pages 133–157. Springer, 2010.

J. T. Spence, R. L. Helmreich, and J. Stapp. *The Personal Attributes Questionnaire: A measure of sex role stereotypes and masculinity-femininity*. University of Texas, 1974.

K. Spiel, O. L. Haimson, and D. Lottridge. How to do better with gender on surveys: A guide for hci researchers. *Interactions*, 26(4):62–65, June 2019. ISSN 1072-5520. doi: 10.1145/3338283. URL `https://doi.org/10.1145/3338283`.

M. V. Spiers, M. Sakamoto, R. J. Elliott, and S. Baumann. Sex differences in spatial object-location memory in a virtual grocery store. *CyberPsychology & Behavior*, 11(4):471–473, 2008. URL `https://doi.org/10.1089/cpb.2007.0058`.

T. D. Steensma, B. P. Kreukels, A. L. de Vries, and P. T. Cohen-Kettenis. Gender identity development in adolescence. *Hormones and behavior*, 64(2):288–297, 2013.

M. R. Stevenson and K. N. Black. Paternal absence and sex-role development: A meta-analysis. *Child Development*, pages 793–814, 1988.

S. L. Stewart and J. A. Kirkham. Predictors of individual differences in emerging adult theory of mind. *Emerging Adulthood*, 2020.

S. Stumpf, A. Peters, S. Bardzell, M. Burnett, D. Busse, J. Cauchard, and E. Churchill. Gender-inclusive hci research and design: A conceptual review. *Foundations and Trends in Human–Computer Interaction*, 13(1):1–69, 2020.

M. Sánchez-López and I. Cuéllar-Flores. Comparison of feminine gender norms among spanish and american college women. *Psychology*, 2(4):300–306, 2011. URL `https://doi.org/10.4236/psych.2011.24047`.

D. Tager and G. Good. Italian and american masculinities: A comparison of masculine gender role norms. *Psychology of Men & Masculinity*, 6(4):264–274, 2005. URL `https://doi.org/10.1037/1524-9220.6.4.264`.

D. S. Tan, M. Czerwinski, and G. Robertson. Women go with the (optical) flow. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 209–215, 2003.

C. Tannenbaum, R. P. Ellis, F. Eyssel, J. Zou, and L. Schiebinger. Sex and gender analysis improves science and engineering. *Nature*, 575(7781):137–146, 2019.

B. Thorne and Z. Luria. Sexuality and gender in children's daily worlds. *Social problems*, 33(3):176–190, 1986.

L. S. Tottenham, D. Saucier, L. Elias, and C. Gutwin. Female advantage for spatial location memory in both static and dynamic environments. *Brain and Cognition*, 53(2):381–383, 2003. URL `https://doi.org/10.1016/S0278-2626(03)00149-0`.

R. Umbach, N. Henry, G. F. Beard, and C. M. Berryessa. Nonconsensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642382. URL `https://doi.org/10.1145/3613904.3642382`.

F. Uzefovsky, R. A. Bethlehem, S. Shamay-Tsoory, A. Ruigrok, R. Holt, M. Spencer, L. Chura, V. Warrier, B. Chakrabarti, E. Bullmore, et al. The oxytocin receptor gene predicts brain activity during an emotion recognition task in autism. *Molecular autism*, 10(1):12, 2019.

S. M. Van Anders, J. Steiger, and K. L. Goldey. Effects of gendered behavior on testosterone in women and men. *Proceedings of the National Academy of Sciences*, 112(45): 13805–13810, 2015.

S. G. Vandenberg and A. R. Kuse. Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, 47(2):599–604, 1978.

D. Voyer, A. Postma, B. Brake, and J. Imperato-McGinley. Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin & Review*, 14(1):23–38, 2007. URL `https://doi.org/10.3758/BF03194024`.

M. Wang, M. M. Bhuiyan, E. H. R. Rho, K. Luther, and S. W. Lee. Understanding the relationship between social identity and self-expression through animated gifs on social media. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), Apr. 2024a. doi: 10.1145/3641031. URL `https://doi.org/10.1145/3641031`.

X. Wang, Z. Wang, M. Zhang, K. Yu, P. Hui, and M. Fan. Avatar appearance and behavior of potential harassers affect users' perceptions and response strategies in social virtual reality (vr): A mixed-methods study. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), Nov. 2024b. doi: 10.1145/3686934. URL `https://doi.org/10.1145/3686934`.

V. Warrier, V. Chee, P. Smith, B. Chakrabarti, and S. Baron-Cohen. A comprehensive meta-analysis of common genetic variants in autism spectrum conditions. *Molecular autism*, 6(1):49, 2015.

S. Wijenayake, N. van Berkel, V. Kostakos, and J. Goncalves. Measuring the effects of gender on online social conformity. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

W. W. Wilkinson. Authoritarian hegemony, dimensions of masculinity, and male antigay attitudes. *Psychology of Men & Masculinity*, 5(2):121, 2004.

W. Wood and A. H. Eagly. Biosocial construction of sex differences and similarities in behavior. In *Advances in experimental social psychology*, volume 46, pages 55–123. Elsevier, 2012.

World Economic Forum. Global gender gap report 2020, 2020. URL http://www3.weforum.org/docs/WEF_GGGR_2020.pdf.

T. Yamauchi, J. H. Seo, N. Jett, G. Parks, and C. Bowman. Gender differences in mouse and cursor movements. *International Journal of Human-Computer Interaction*, 31(12): 911–921, 2015.

T. Ye, K. Reinecke, and L. P. Robert. Personalized feedback versus money: The effect on reliability of subjective data in online experimental platforms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 Companion, pages 343–346, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346887. doi: 10.1145/3022198.3026339. URL https://doi.org/10.1145/3022198.3026339.

Q. Yu and B. Li. mma: an r package for mediation analysis with multiple mediators. *Journal of Open Research Software*, 5 (1), 2017.

Q. Yu, X. Wu, B. Li, and R. A. Scribner. Multiple mediation analysis with survival outcomes: with an application to explore racial disparity in breast cancer survival. *Statistics in medicine*, 38(3):398–412, 2019.

M. Zentner and K. Mitura. Stepping Out of the Caveman's Shadow: Nations' Gender Gap Predicts Degree of Sex Differentiation in Mate Preferences. *Psychological Science*, 23(10):1176–1185, 2012. ISSN 14679280. doi: 10.1177/0956797612441004.