Predictive Power of Biological Sex and Gender Identity on Economic Behavior: A Validated Instrument for Measuring Gender Identity*

Stefano Piasenti†, Müge Süer‡

October 11, 2024

Abstract

Behavioral differences by biological sex are still not fully understood, suggesting that studying gender differences in behavioral traits through the lenses of continuous identity might be a promising avenue to understand the remaining observed gender gaps. Using a large U.S. online sample ($N = 2017$) and machine learning, we develop and validate a new continuous gender identity measure consisting of separate femininity and masculinity scores. In a first study, we identify ninety attributes from prior research and conduct an experiment to classify them as feminine and masculine. In a subsequent study, a different group of participants completes tasks designed to elicit behavioral traits that have been previously documented in the behavioral economics literature to exhibit binary gender differences. Data for the second study are collected in two waves; the first wave serves as a training sample, allowing us to identify key attributes predicting behavioral traits, create candidate identity measures, and select the most effective one, comprising sixteen attributes, based on predictive power. Finally, we use the second wave (test sample) to validate our gender identity measure, which outperforms existing ones in explaining gender differences in economic decision-making. We show that confidence, competition, and risk are associated with masculinity, while altruism, equality, and efficiency are with femininity, providing new possibilities for targeted policymaking.

JEL-codes: D91, J16, J62, C91
Keywords: Biological sex, Gender identity, Online experiment, Machine learning

*We thank Dirk Engelmann, Lavinia Kinne, Dorothea Kübler, Davide Pace, Pascal Pillath, Giacomo Rubbini, Sebastian Schweighofer-Kodritsch, Roel van Veldhuizen, participants of the CRC Retreat in Tutzing 2022, the Economic Science Association World Meeting in Lyon 2023, the Bergen-Berlin Behavioral Economics Workshop in Bergen 2024 and the CRC Retreat in Schwanenwerder 2024 for helpful comments. Support by the Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged. This study is preregistered as Süer, M., & Piasenti, S. (2022, December 07). Predictive Power of Biological Sex and Gender. in the OSF registry, https://doi.org/10.17605/OSF.IO/TJX42.

†University of Bologna (stefano.piasenti2@unibo.it)
‡Humboldt University Berlin (sueermue@hu-berlin.de)

1 Introduction

In recent years, economists and psychologists have extensively studied gender differences in psychological and behavioral traits to understand their implications for labor market outcomes, discriminatory practices, and stereotypes (for reviews, see Markowsky and Beblo (2022) and Lozano et al. (2022)). Experimental economics has typically accounted for these differences by controlling for biological sex. However, the results concerning behavioral traits have often been inconsistent, making it difficult to derive meaningful implications for gender policy (Croson and Gneezy, 2009). This challenge highlights the need to examine whether behavioral differences are inherently related to biological sex, or whether other related social factors, such as "gender identity", also contribute to gender differences and create heterogeneity among individuals. Moreover, an increasing number of individuals define themselves as gender non-binary or genderqueer (Coffman et al., 2017; Wilson and Meyer, 2021; Coffman et al., 2024) and do not wish to fit into the traditional binary gender approach. The identity utility approach, as outlined in Akerlof and Kranton (2000), argues that individuals' alignment with their social identity affects their utility and, consequently, their behavior. This implies that any gender identity heterogeneity, as a form of social identity, should also be reflected in human behavior.

In this paper, we offer a validated gender identity measure to address this pressing need. In particular, our measure outperforms existing instruments in explaining gender differences in behavioral traits such as confidence in one's absolute and relative performance, risk, altruism, efficiency, equality, and competition preferences. This is achieved by improving model fit and accounting for a portion of the variance traditionally attributed to the biological sex dummy. Furthermore, by leveraging a two-dimensional framework to measure gender identity, we demonstrate that gender differences in confidence, risk, and competition are primarily driven by masculinity, while traits such as altruism, equality, and efficiency are associated with femininity. This approach sheds new light on the heterogeneity often obscured by conventional dichotomous categorizations of gender providing new possibilities for targeted policymaking.

A comprehensive definition of gender identity that emphasizes its social nature is given by the World Health Organization (2024), which states that gender refers to "the characteristics of women, men, girls and boys that are socially constructed".[1] However, conceptualizing and measuring gender identity has always been challenging (Muehlenhard and Peterson, 2011; Hyde et al., 2019). For this reason, the economic literature has remained predominantly within the *biological foundationalist paradigm*, equating gender with the binary variable of sex (Nicholson, 1994). More recently, with examples such as Meier-Pesti and Penz (2008), Kastlunger et al. (2010), Adamus (2018), Sent and van Staveren (2019), Brenøe et al. (2022), Dorofeev (2022), Fornwagner et al. (2022), and Sahi (2023) the discourse within the experimental economics literature has shifted towards the social constructionist perspective of Butler (1990), separating binary sex from socially perceived gender. These papers represent a departure in the economic literature from the traditional reliance on the biological sex variable.

[1] Another definition of gender by Brenøe et al. (2022) describes it as "a manifestation of individual traits and behaviors, social and personal perceptions of identity, and agreement with or divergence from societal norms".

Expanding on the existing research on gender identity, we measure individuals' femininity and masculinity scores separately, using a two-dimensional approach that operates at a multitude of subjective and social levels, as outlined in the work of Knaak (2004). To measure gender identity, our study utilizes a list of attributes on which participants rate themselves. Subsequently, we classify all attributes as feminine, masculine and neutral and finally, we combine these ratings to produce a femininity and a masculinity score for each participant. Each score is calculated by averaging the self-ratings of the corresponding attributes in the list.

The discussion about the binary (i.e., 1 if female, 0 if male) and bipolar (i.e. if one is female, they cannot be male at the same time) essence of gender has begun in the psychology literature with Constantinople (1973), who raised the question of whether masculinity and femininity should be studied on distinct scales. Bem (1974) has followed up on the Constantinople (1973) perspective providing the first two-dimensional gender identity measure in the psychological literature.

The continuous gender identity measure by Bem (1974) is still one of the most prominently used gender identity measures.[2] It is based on Bem's sex role inventory (henceforth BSRI) and made of two components called "BEMS_feminine" and "BEMS_masculine". To build Bem (1974)'s gender identity measure, respondents rate themselves, for each item in the 40-item original BSRI, on a seven-point Likert scale ranging from "never or almost never true" to "always or almost always true." Each item of the BSRI is an attribute, classified as either feminine or masculine. "BEMS_feminine" is the average of the 20 feminine attributes, and "BEMS_masculine" averages the remaining 20 masculine attributes. Due to the number of attributes necessary to create the measure, this approach is not ideal for online experiments which are typically rather short.

In the economics literature, we are not the first to propose a continuous gender identity measure. A recent attempt has been put forward by Brenøe et al. (2022). They propose a single-item, hence a bipolar, continuous gender measure based on the question "Where would you put yourself on this scale?" using a seven-point Likert scale, ranging from "very masculine" to "very feminine", and they test its predictive power for behavioral traits. Henceforth we will refer to their measure as "CGI", which stands for "Continuous Gender Identity". CGI exhibits a strong correlation with the binary sex dummy ($\rho > 0.70$). Thus, including both the binary sex dummy and the CGI variable in a regression leads to problems in interpreting the coefficients due to multicollinearity.[3] On the other hand, including CGI alone would mimic very closely the inclusion of the biological sex dummy.

In this paper, we run two studies employing online experiments with a U.S. sample ($N = 2017$) on the Prolific platform(Palan and Schitter, 2018), with the aim to create a new two-dimensional continuous gender identity measure, with separate feminine and masculine components. In our first study (Study 1), we construct an updated sex-role inventory and classify each attribute as feminine and masculine in four different ways (i.e. desirability in society at large and in the workplace, and gender norm in society at large and in the workplace) by means of

[2]The Bem Sex Role Inventory (BSRI) has garnered approximately 17,000 citations, according to Google Scholar.

[3]A correlation of even 0.70 can already affect the regression results, see e.g.: Hair et al. (2010).

an online experiment (Experiment 1, $N = 915$). In the second study (Study 2), we run another experiment (Experiment 2, $N = 1102$) to measure behavioral traits that have been previously documented to exhibit binary sex differences, namely confidence in one's absolute and relative performance on gender-congruent and incongruent tasks (female and male tasks as in Dreber et al. (2014) and Exley and Kessler (2022)), risk (Croson and Gneezy, 2009), altruism (Dreber et al., 2014), efficiency, equality (Andreoni and Vesterlund, 2001), and competition preferences (Niederle and Vesterlund, 2007). We additionally asked participants to rate themselves on all components of our contemporary sex-role inventory, which were presented in a random order at the individual level. Experiment 2 is conducted in two waves. Using the first wave of this experiment as a training sample ($N = 501$), we apply machine learning to determine the optimal combination of attributes for predicting behavioral traits and create a new gender identity measure. Finally, using the second wave ($N = 601$) as a test sample, we evaluate the predictive power of our new measure against alternative measures by conducting a comparative analysis.

Study 1. In this study, we revisit the BSRI, which has 20 feminine, 20 masculine and 20 neutral attributes previously identified by Bem (1974). We further include 30 attributes recently revealed by Eberhardt et al. (2023), creating a new inventory of ninety (90) attributes in total that we call the Contemporary Sex-Role Inventory (CSRI). We then design an online experiment to classify the attributes in this new inventory as feminine, masculine and neutral. We create four different categorizations to determine the gender of each attribute. The first categorization, desirability in society at large, is the categorization originally used by Bem (1974). We add desirability in the workplace context as a second categorization. Third, we classify attributes as feminine and masculine using the gender norms in society at large. To elicit "gender norms", we use the norm elicitation method by Krupka and Weber (2013). Hence, we refer to "injunctive" gender norms, which prescribe what one is expected to do within a specific group or context (Krupka et al., 2022). Finally, the attributes are classified based on gender norms in the workplace. To understand the difference between desirability and injunctive gender norms, consider for instance the attribute *tender*. Someone might think that being *tender* is an equally desirable trait for both men and women, but if that person believes that society is very conservative, they might think that the injunctive norm only prescribes women to be tender, making it a feminine norm.

Study 2. In the second study, we follow a four-step approach inspired by the work of Falk et al. (2022) to generate our new gender identity measure. First, we run Experiment 2 to elicit the aforementioned behavioral traits and then ask our participants to rate themselves on each of the 90 attributes. Second, using the first wave as a *training sample* for LASSO analysis, we select the attributes that are the best predictors of the behavioral traits.[4] We then classify the selected attributes as either feminine, masculine, or neutral using the four categorizations of Study 1, desirability in society at large and in the workplace and gender norms in society at large and in the workplace. This categorization process generates eleven (11) candidate gender identity measures (each measure having two components, feminine and masculine). Third, we use the second wave as *test sample* to test the predictive power of the candidate identity

[4]We use 88 attributes in the LASSO analysis. We exclude *willingness to take risk* and *competitiveness* from our analysis as they are dependent variables.

measures. By splitting our dataset into distinct training and test samples, we prevent overfitting. The training sample is solely used for attribute selection, while the test sample evaluates the selected attributes' predictive power. Eventually, we select the best measure out of the eleven (11) candidates. Our proposed measure consists of sixteen (16) attributes, nine (9) masculine and seven (7) feminine. Finally, we compare our measure with the existing measures, CGI and BEMS.

Our results deliver important arguments for treating gender as a two-dimensional continuous concept in addition to biological sex. First, we show that gender differences in confidence, risk, and competition are driven by attributes that are classified as masculine, whereas altruism, equality, and efficiency by those classified as feminine. In other words, individuals displaying more masculine traits tend to exhibit greater confidence, higher risk-taking and competitiveness, while those displaying more feminine traits tend to demonstrate higher levels of altruism, a stronger aversion to inequality, and a lesser concern for efficiency. It is important to underline that being feminine and masculine are not mutually exclusive in our measure.

Furthermore, including our measure in addition to the biological sex dummy in regressions with each of the behavioral traits as an outcome variable increases the adjusted R^2 in nine out of ten behavioral traits. Our measure also decreases the magnitude of the coefficient of the biological sex dummy significantly for five out of ten traits. From an econometric perspective, these findings show that including a continuous measure of gender identity in addition to the biological sex dummy in a regression helps advance our understanding of predicting gender differences in behavioral traits. Finally, our results show that despite the decrease in the coefficient of the biological sex dummy and the increase in the adjusted R^2, the sex dummy stays statistically significant in some traits, i.e. competitiveness and efficiency preferences. Therefore, our paper also provides insights into where more fundamental differences between sexes might lie, even after controlling for societal gender identity.

Our measure consisting of 16 attributes can be implemented at the end of the exit questionnaires by asking participants *""Indicate on a 7-point scale how well each of the following personality characteristics describes yourself."* from *""Never or almost never true (1)"* to *""Always or almost always true (7)"*. We recommend randomizing the order of the attributes in question.[5] The measure then can be included in regressions in addition to the biological sex dummy as the arithmetic average of the attribute ratings forming the femininity and the masculinity scores separately.

Our method, inspired by Bem (1974), comes with three advantages over existing measures. First, it minimizes experimenter demand effects (Zizzo, 2010) by avoiding direct disclosure of a gender identity question and highlights that it may not even be necessary for individuals to consciously identify as non-binary. They may simply describe themselves using certain attributes that society perceives as masculine or feminine. Second, being an attribute-based measure like "BEMS_masculine" and "BEMS_feminine", it has the inherent characteristic of being less strongly correlated with the biological sex dummy, thereby alleviating the concerns associated

[5]The attributes (in alphabetical order) that constitute the feminine component are: *affectionate, compassionate, feminine, flatterable, gullible, sensitive to others needs, tender.* The ones that constitute the masculine component are: *acts as a leader, analytical, assertive, athletic, broad, dominant, masculine, strong personality, willing to take a stand.*

with multicollinearity. Finally, it is designed to encompass a reduced set of attributes (with 16 attributes) in comparison to the above-mentioned BSRI measure (with 40 attributes). Remarkably, despite this reduction, the proposed measure exhibits marginally superior explanatory power with respect to "BEMS_feminine" and "BEMS_masculine" for many of the measured behavioral traits and makes it easier to implement in questionnaires.

Our work adds to the existing literature measuring gender identity (Bem, 1974, 1993; Kachel et al., 2016; Magliozzi et al., 2016; Brenøe et al., 2022; Dorofeev, 2022). We provide a validated measure that offers a more detailed understanding of gender differences, separating societal characteristics from biological sex. A better understanding of gender in turn means better-designed policies and research in the future. Moreover, our approach can be interpreted in light of recent literature that has advocated for leveraging machine learning methodologies to uncover latent structures within data (Athey and Imbens, 2019). We draw upon the principles of hidden heterogeneity to uncover attributes that are otherwise neglected by the dichotomy of biological sex.

The remainder of the paper is structured as follows. Sections 2 and 3 describe the two main studies that constitute our paper. They detail the methods we used, the experimental designs we employed and the results we obtained. Finally, Section 4 discusses the potential limitations and extensions of our work, and Section 5 concludes.

2 Study 1: A New Sex Role Inventory

The goal of our first study is to construct a new sex role inventory and determine the masculinity and femininity of each attribute in this inventory across four different categories: 1) desirability in society at large, 2) desirability in the workplace, 3) gender norm in society at large, and finally 4) gender norm in the workplace. These categorizations are later used in the process of creating our gender identity measure (see Figure 1 for the general picture connecting Study 1 and Study 2).

2.1 The Attribute Collection

A sex role inventory is a collection of attributes, i.e., personal characteristics, that are classified as feminine, masculine, and neutral to help identify a person's masculinity or femininity. One of the first attempts to group attributes as feminine and masculine in US society is the Bem Sex-Role Inventory (BSRI) which concerns 60 attributes, with 20 being masculine (e.g., "assertive"), 20 feminine (e.g., "affectionate"), and 20 neutral (e.g., "friendly") (see the full list in Table 1) (Bem, 1974). The inventory was originally constructed to measure individuals' psychological androgyny scores. Bem (1974) argued in favor of androgyny in the sense that a person could be feminine and masculine at the same time.

Meanwhile, a recent study by Eberhardt et al. (2023) has revealed another set of attributes that appear to be gender-specific. By analyzing differences in the content of recommendation letters written for male and female junior researchers, Eberhardt et al. (2023) showed that men were systematically more likely to be defined in terms of their abilities (e.g., "talented"),

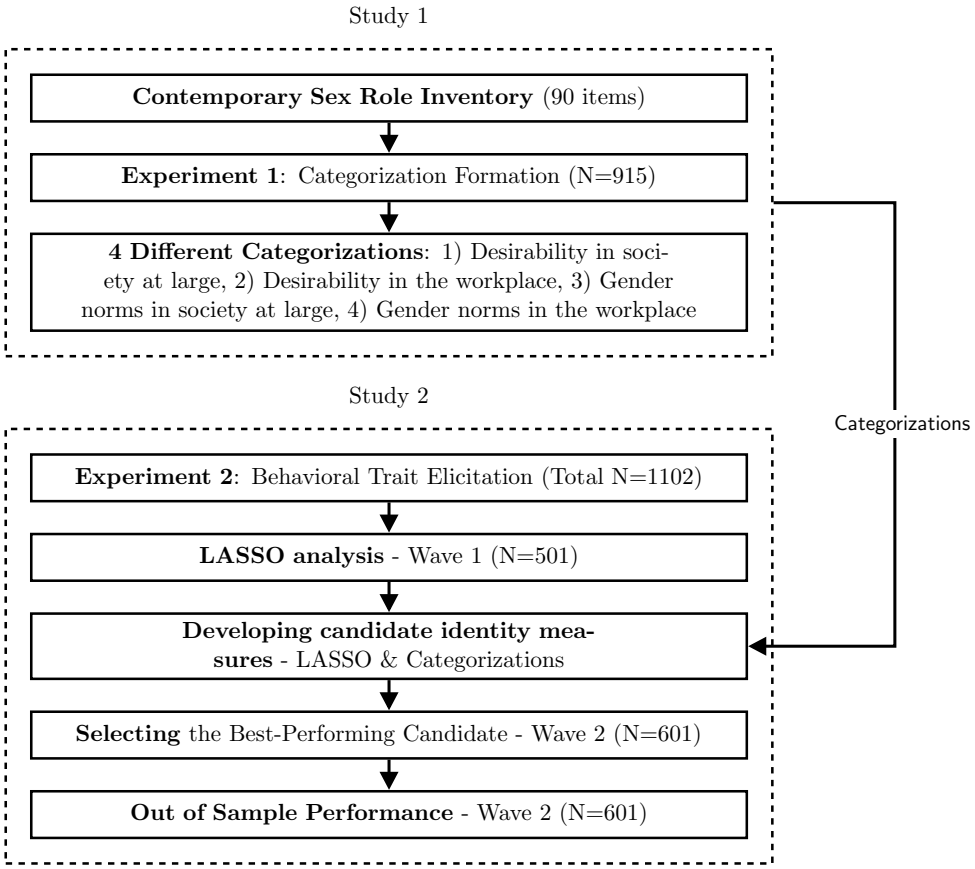


Figure 1: Flowchart depicting the design of Study 1 and Study 2.

while women were more likely to be described with grindstone attributes (e.g., "hardworking"). Bem (1974) and Eberhardt et al. (2023) have different perspectives in at least two ways. First, Eberhardt et al. (2023) only show how academics use certain attributes in a gendered way but does not provide any information about whether these attributes are seen as masculine or feminine by society at large. Second, the attributes identified by Eberhardt et al. (2023) are work-related, whereas BSRI addresses gender in society at large.

To construct an updated sex role inventory, we see considerable promise in combining the 60-item BSRI with the work-related attributes by Eberhardt et al. (2023). In doing so, we create a more comprehensive list of attributes (60 from Bem's original inventory and 30 from Eberhardt et al. (2023)) that addresses not only societal but also work-related gender roles. We call this the Contemporary Sex-Role Inventory (CSRI) (see Table 1).

2.2 Experiment 1: Four Different Categorizations

Once we have collected the attributes to be used in CSRI, the second step is to classify them as feminine, masculine, or neutral. One way of doing this is the method used by Bem (1974). Masculinity, femininity and neutrality of the attributes in the original BSRI were determined by surveying two distinct samples from the US population about the desirability of each of them for women and men separately in American society at large. If an attribute was significantly more desirable for men than for women (two-sample t-test $p \leq 0.05$), it was qualified as masculine.

Table 1: Contemporary Sex-Role Inventory (CSRI).

Bem's Sex-Role Inventory∗			Eberhardt et al. 2023∗∗	
Masculine	Feminine	Neutral	Ability	Grindstone
acts as a leader	affectionate	adaptable	able	active
aggressive	cheerful	conceited	broad	challenger
ambitious	childlike	conscientious	careful	dedicated
analytical	compassionate	conventional	clear	diligent
assertive	does not use harsh language	friendly	creative	disciplined
athletic	eager to soothe hurt feelings	happy	expert	driven
competitive	feminine	helpful	insightful	endures difficult situations
defends own beliefs	flatterable	inefficient	intellectual	exerts effort
dominant	gentle	jealous	knowledgeable	hardworking
forceful	gullible	likeable	rigorous	is not afraid of difficulties
independent	loves children	moody	skillful	motivated
individualistic	loyal	reliable	smart	patient
leadership ability	sensitive to other's needs	secretive	solid	quick
makes decisions easily	shy	sincere	talented	takes on challenging tasks
masculine	soft spoken	solemn	technical	thorough
self reliant	sympathetic	tactful		
self-sufficient	tender	theatrical		
strong personality	understanding	truthful		
willing to take a stand	warm	unpredictable		
willing to take risks	yielding	unsystematic		

Notes: In total 90 items.∗ The classification of masculine, feminine and neutral attributes in the Bem´s Sex Role Inventory table is the original one from Bem (1974). ∗∗ The classification of ability and grindstone attributes in the Eberhardt et al. (2023) table is a selected list of Eberhardt et al. (2023) accommodating the most frequently used 15 masculine and 15 feminine attributes. The raw words detected by Eberhardt et al. (2023) are also transformed into personality traits, such as *hardwork* to *hardworking*, to fit in our study.

If it was significantly more desirable for women, it was qualified as feminine. If there was no significant desirability difference (two-sample t-test $p > 0.05$), the attribute was qualified as neutral.

The original Bem (1974) classification represents just one way of determining the gender of attributes. One shortcoming is that the desirability of attributes for men and women is only determined in society at large. As mentioned above, the Eberhardt et al. (2023) attributes are work-related by construction. Therefore, we also included the workplace perspective to classify the CSRI items as feminine, masculine and neutral.

Besides the desirability of attributes in the workplace context, we also chose to elicit the gender norm of each attribute for two reasons. First, the desirability of an attribute might differ from its perceived masculinity/femininity (Hoffman and Borders, 2001), which we interpret as the difference between desirability and injunctive gender norm. Second, the original desirability elicitation is not incentivized. Hence, we modified the Krupka and Weber (2013) norm elicitation technique to our context classifying attributes as feminine and masculine in terms of gender norms.

In total, we identified four possible categories to classify our 90 attributes as feminine and masculine. The first two categorizations are related to desirability in two different contexts: 1) desirability in society at large as in the original work by Bem (1974) and 2) desirability in the workplace by applying the approach of Bem (1974) in the workplace context. The last two categorizations are related to gender norms in two different contexts: 3) gender norms in society at large using a modified version of the Krupka and Weber (2013) norm elicitation and 4) gender

norms in the workplace using the same modified version of the Krupka and Weber (2013) norm elicitation in the workplace context. To classify CSRI items as feminine or masculine in these four different ways, we designed Experiment 1.

2.2.1 Procedural Details

Experiment 1 was programmed in Qualtrics and run on the platform Prolific, using a US sample in December 2022 (Palan and Schitter, 2018). It involved 915 participants. Instructions for all treatments can be found in Appendix C.1. The experiment employed six between-subject treatments. The treatment allocation of subjects was randomized and it was successful in terms of gender balance (see Table 5 in Appendix A).

In Experiment 1, participants earned a guaranteed £1.50 show-up fee for their participation upon completing the study in treatments 1, 2, 3 and 4.[6] In treatments 5 and 6, in addition to the show-up fee, participants could earn a bonus of up to £1.00 due to the Krupka and Weber (2013) incentivization. The experiment's mean payout considering all treatments was £1.70 (the median was £1.50 by design). The median completion time was 9 minutes and 31 seconds.

We used a strict exclusion criterion. Participants were asked to answer a comprehension question correctly after the initial instructions.[7] They had two chances to give the correct answer. Those who failed to answer the question correctly on both attempts could not continue in the experiment and were excluded from receiving payments. Participants also faced an attention check during the experiment, which did not exclude them from payment but is used in our analyses as a robustness check.[8] We used a captcha test to filter out non-human users and we only recruited native English speakers to ensure that the instructions were properly understood.

2.2.2 Experimental Design

Our online experiment comprised six between-subject treatments to accommodate our four different categorizations, four of which are required for the categorizations of desirability in society at large and desirability in the workplace, and the remaining two for the categorizations of gender norms in society at large and gender norms in the workplace (see Table 2).

In Bem (1974)'s original work, masculinity, femininity, and neutrality of attributes were determined based on their desirability for men and women in American society at large. Therefore, we first employed the desirability in society at large categorization of Bem (1974) in our treatment 1, GeneralMen, and treatment 2, GeneralWomen.

To elicit desirability, we followed the methodology by Bem (1974). We presented our participants with the question *"How desirable is it in American society for a man to possess each of*

[6] It may seem unusual to see payments in pounds rather than dollars, considering the US sample, but that does result from Prolific being a UK platform.

[7] The comprehension question we used can be found in Figure 4 in Appendix C.1.

[8] The attention check page was analogous to the pages where participants had to rate the attributes (see e.g. Figure 5 in Appendix C.1.) with the following differences: i) attributes were substituted with the word "check", ii) the header of the page read: "Please ignore the following question. Leave it blank and advance to the next screen by clicking the button below." Only 13 out of 915 participants did not pass the attention check. Results are robust excluding those who failed the attention check.

these attributes?" in treatment 1 and *"How desirable is it in American society for a woman to possess each of these attributes?"* in treatment 2. We asked them to rate the desirability from 1 to 7, 1 being *not at all desirable* and 7 *extremely desirable* (see e.g. Figure 5 in Appendix C.1 for details). The desirability elicitation was not incentivized as in the original work of Bem (1974).

Based on these treatments, an attribute was classified as feminine (masculine) if the difference between the average desirability for women, GeneralWomen, and the average desirability for men, GeneralMen, was significantly positive (negative) for this attribute (two-sample t-test $p < 0.05$). If the difference between the average desirability for women, GeneralWomen, and the average desirability for men, GeneralMen, was not statistically significantly different (two-sample t-test $p > 0.05$), the attribute was classified as neutral. Second, we included the desirability in the workplace with our treatment 3, WorkMen, and treatment 4, WorkWomen. In these treatments, we asked a similar desirability question but this time in the American workplace context, instead of in the American society context. *"How desirable is it in the American workplace for a man to possess each of these attributes?"* in treatment 3 and *"How desirable is it in the American workplace for a woman to possess each of these attributes?"* in the treatment 4. The rating was again from 1 to 7, 1 being *not at all desirable* and 7 *extremely desirable*. The treatments WorkMen and WorkWomen were also not incentivized.

Third, we elicited the masculinity and femininity norm for each attribute in the inventory adapting the Krupka and Weber (2013) norm elicitation technique to our setting (henceforth KW in short). This technique was developed to elicit collective norms in an incentivized manner (Krupka et al., 2022). In treatment 5, KWGeneral, we asked participants to rate the CSRI attributes on a 4-point masculinity-femininity scale based on what they believe is the most frequent answer in the experiment in the context of American society. Participants faced the following statement *"In this survey you are asked to rate the masculinity/femininity of attributes based on what you believe the most frequent answer will be in this survey".* The rating is from 1 to 4, 1 being "very masculine", 2 being "masculine", 3 being "feminine" and 4 "very feminine" (for more details see instructions in Appendix C.1). The gender norm of each attribute was then determined by taking the mode of all answers.

The KW technique was incentivized and participants were paid an additional bonus of up to £1.00. The bonus was based on their correct identification of the ten most frequently given answers by other participants. Namely, they received an extra £0.10 per correct response for 10 randomly chosen attributes out of 90, summing up to a maximum of £1.00. Finally, the last treatment KWWork addressed our fourth and final categorization, gender norms in the workplace. It therefore repeated the same procedures as in treatment 5, but in the workplace context instead.

In all treatments, the CSRI attributes were presented to participants in a random order at an individual level. After the CSRI part, participants completed an exit questionnaire collecting demographics.

It is important to underline that by moving from treatments 1, 2, 3, and 4 to treatments 5 and 6 we did not just move from measuring desirability to measuring gender norms. Between these categorizations, additional elements also changed: In treatments 5 and 6, i) femininity

and masculinity were measured on the same scale with a 4-point Likert scale, 1 being "very masculine" to 4 being "very feminine", and ii) the elicitation was incentivized. We made this decision out of our commitment to adhere closely to the original desirability measurement by Bem (1974), which did not include incentives, used a 7-point Likert scale and combined answers of two different samples to calculate the gender of each attribute. At the same time, we followed the original elicitation method of Krupka and Weber (2013), which comprised incentives, a 4-point Likert scale and a single sample.

Table 2: Four categorizations and corresponding treatments of Study 1.

Categorization	Corresponding Treatment(s)	Description	N	Incentivization
1) Desirability in society at large	(1) GeneralMen	for men	152	No
	(2) GeneralWomen	for women	151	No
2) Desirability in the workplace	(3) WorkMen	for men	150	No
	(4) WorkWomen	for women	150	No
3) Gender norms in society at large	(5) KWGeneral	for all	158	Yes
4) Gender norms in the workplace	(6) KWWork	for all	154	Yes

Notes: *Categorization* represents the four categories that we used to classify all attributes as feminine, masculine and neutral. *Corresponding Treatments* reveals which treatment of Experiment 1 is used to form this category. *Description* shows whether the question in the treatment was about men, women or for all. *N* is the sample size and *Incentivization* is whether the question in the corresponding treatment is incentivized or not.

2.3 Results

In this section, we provide a general overview of how attributes are classified based on our four categorizations displayed in Table 2. First, all 90 CSRI items are classified as feminine, masculine, or neutral based on the desirability in society at large using the GeneralMen and GeneralWomen treatments. Of these attributes, 23 stand as feminine (e.g. sensitive to others' needs), 48 as masculine (e.g. dominant), and 19 as neutral (e.g. adaptable).[9]

The same attributes are then similarly classified based on the desirability in the workplace category using the WorkMen and WorkWomen treatments. Out of the 90 attributes, 16 are feminine, 21 are masculine, and 53 are neutral.

Furthermore, we classify our list of attributes as "very masculine", "masculine", "feminine" and "very feminine" using the KWGeneral and KWWork treatments to reveal gender norms in society at large and the workplace respectively. Based on the KWGeneral treatment, 29 attributes are classified as "very masculine", 22 as "masculine", 20 as "feminine" and 19 as "very feminine". KWWork, on the other hand, makes it possible to form a gender norm categorization in the workplace context. Based on this treatment, 21 attributes are stated to be "very masculine", 30 "masculine", 23 "feminine" and 16 "very feminine". The complete list of attributes separately classified based on four categorizations can be found in Table 6 in Appendix A.[10]

[9]The higher prominence of masculine attributes arises since Eberhardt et al. (2023) attributes are mostly classified as masculine: specifically 25 as masculine, 3 as neutral, and 2 as feminine.

[10]In the desirability treatments in Table 6 of Appendix A, we additionally report the average difference in desirability for each attribute between men and women. A negative difference indicates that the attribute is more desirable for women, while a positive difference suggests it is more desirable for men. This allows us to also rank attributes from most desirable for men to least, and similarly, from most desirable for women to least.

3 Study 2: Generating a New Gender Identity Measure

We designed a second online experiment, Experiment 2, capturing behavioral traits to form a measure of gender identity. Experiment 2 was run in two waves. We followed four main steps which were inspired by the Falk et al. (2022) preference survey module: 1) capturing behavioral traits, 2) developing candidate identity measures using machine learning and Study 1 categorizations, 3) selecting the best-performing gender identity measure out of all candidates, and finally, 4) a comparison with the existing measures (see Figure 1 for the general overview of Study 1 and Study 2).

1. Capturing behavioral traits: In the first wave, we used validated elicitation methods to reveal gender differences, namely absolute and relative confidence, risk, equality and efficiency preferences, altruism, and finally competitive attitudes, for which gender differences have been previously identified in the economics literature (selective examples including Andreoni and Vesterlund (2001); Niederle and Vesterlund (2007); Croson and Gneezy (2009); Dreber et al. (2014); Exley and Kessler (2022)). Additionally, we gathered participants' self-reported personal ratings on all of the 90 CSRI attributes.

2. Developing candidate gender identity measures using machine learning and Study 1 categorizations: Using the data collected in the first wave, we ran LASSO regressions (Tibshirani, 1996) to pinpoint the attributes that have the best predictive power for each measured behavioral trait (i.e. attributes that could predict at least two different behavioral traits.). Hence, the first wave of Experiment 2 ($N = 501$) served as *training sample* for within-sample predictions. To form the candidate identity measures, we then referred back to Study 1. The four categorizations generated in Study 1 allowed us to group the attributes selected by the LASSO regressions in four different ways. Combining the four categorizations from Study 1 with the number of LASSO appearances, we created 11 candidate gender identity measures.

3. Selecting the best-performing gender identity measure out of all candidates: We then went on to compare the candidates in terms of how well they performed in absorbing the effect of the biological sex dummy in as many behavioral traits as possible when both the sex dummy and the gender identity measures were included in regressions with behavioral traits as dependent variables. The best-performing measure out of 11 candidates was selected using the second wave of Experiment 2 as the *test sample* ($N = 601$).

4. Comparison to existing measures: Finally, using again the *test sample* we compared the predictive power of our new gender identity measure on the behavioral traits against two existing measures from the literature, BEMS by (Bem, 1974) and CGI by (Brenøe et al., 2022). Details about the two waves of Experiment 2 follow.

3.1 Experiment 2: Capturing Behavioral Traits

3.1.1 Procedural Details

The experiment was programmed in Qualtrics and conducted using Prolific in two waves. The first wave was run between December 2022 and April 2023 and involved 501 participants. Par-

ticipants earned a guaranteed £2.50 show-up fee for their participation upon completing the study. In addition to that, they could earn an additional bonus of up to £3.00. The additional bonus was calculated based on one randomly selected incentivized task. The median earning (show-up fee plus earnings from the tasks) was £3.70. The median completion time for the first wave of Experiment 2 was 15 minutes and 0.5 seconds.

The second wave was run in July 2023 and involved 601 participants. As in the first wave, participants earned a guaranteed £2.50 show-up fee for their participation upon completing the study. The median earning (show-up fee plus earnings from the tasks) was £3.70. The median time to complete the second wave was 15 minutes and 19 seconds.

In both waves, earnings were calculated in points and were transformed into money at an exchange rate of 1 point = 0.02. A captcha test was used to filter out non-human users and only native English speakers were recruited. We used the same attention check as in our Experiment 1.[11]

3.1.2 Experimental Design

Experiment 2 entailed a within-subject design. In this experiment, all participants performed five incentivized tasks that have been prominently used in the literature investigating gender differences. The first two tasks, math to represent the male domain and a verbal task to represent the female domain, were taken from Dreber et al. (2014) and adapted for the online experiment setup. [12] These tasks were presented in random order. After completing the math and word tasks, participants were asked to report their beliefs about their absolute and relative performance in both tasks. Then, we elicited their risk preferences using Holt and Laury (2002), altruism as in Dreber et al. (2014) and finally, efficiency and equality preferences with the Andreoni and Vesterlund (2001) method. These tasks were selected to measure their absolute and relative confidence in a gender congruent and a gender non-congruent domain, i.e. male and female domains, altruism, risk, efficiency and equality preferences. One of the tasks was randomly selected to determine their bonus payments (see Table 3 for a full list of elicitation tasks).

After completing the tasks, participants were asked to indicate on a 7-point Likert scale how well each item of the 90-item CSRI described them (screenshots of Experiment 2 are reported in Appendix C.2). The order of CSRI items was randomized at the individual level.

Following their self-reports on each CSRI item, the experiment continued with an exit questionnaire. Risk and competition preferences following Dohmen et al. (2011) and Fallucchi et al. (2020) respectively, were further elicited as exit questionnaire survey items and they were not incentivized.

[11] We did not ask comprehension questions in this experiment before each task, since we used established tasks and thereby did not overly lengthen the experiment. Only 28 participants out of 1102, the total number of Wave 1 and 2 participants, failed the attention check. Results are robust excluding those who failed the attention check.

[12] In the gender experimental literature, math tasks are considered stereotypically male while word tasks are stereotypically female even if actual differences in performance cannot be proven (see e.g.,Kimura (2004); Günther et al. (2010))

Table 3: Elicitation tasks of Experiment 2.

Behavioral Trait	Elicitation Task	Measure
Absolute Confidence in Male Domain	Ten math questions	Belief about the number of correctly solved questions
Relative Confidence in Male Domain	Ten math questions	Belief about relative performance compared to 100 randomly chosen participants
Absolute Confidence in Female Domain	A word search matrix with ten hidden words	Belief about the number of correctly identified words
Relative Confidence in Female Domain	A word search matrix with ten hidden words	Belief about relative performance compared to 100 randomly chosen participants
Risk I	Choosing between a lottery and a safe choice (Holt and Laury, 2002)	Switching point
Risk II	Survey item (Dohmen et al., 2011)	Self-reported point in the 11-point Likert scale
Altruism	Donation to a charity (Dreber et al., 2014)	Donated portion of the endowment
Equality	Menu of 8 decisions to share systematically altered endowments (Andreoni and Vesterlund, 2001)	The average of the generated equality differences in 8 decisions
Efficiency	Menu of 8 decisions to share systematically altered endowments (Andreoni and Vesterlund, 2001)	The average of the generated efficiency ratios in 8 decisions
Competition	Survey item (Fallucchi et al., 2020)	Self-reported point in the 7-point Likert scale

Notes: Behavioral Trait is the list of all elicited traits. *Elicitation Task* is the task used to elicit the trait in question and *Measure* is how the dependent variables are later constructed. Math and word tasks are inspired by Dreber et al. (2014), but adapted to an online setting.

Absolute and Relative Confidence

To elicit absolute and relative confidence, we followed Dreber et al. (2014) and elicited them in male and female domains. To do so, we used the math and verbal tasks mentioned above. The math task was a 6-item addition and multiplication of $1s$ and $0s$. Participants were asked to complete ten problems in 1 minute. The verbal task was a 7x8 word search matrix with ten hidden 4-letter words. Participants also had 1 minute to find the words. Both tasks were presented in a randomized order. When the math (word) task was randomly selected for payment, participants earned 10 points for each question (word) they answered (identified) correctly.

Following these two real-effort tasks, each subject was asked to report their performance-related confidence on both tasks. Confidence was elicited in two ways, first by asking how many problems they solved correctly (words they identified correctly), second by asking where they thought their performance lay within a group of 100 randomly selected subjects. Both elicitations were incentivized. For the former, if the answer was correct, they earned an additional 10 points bonus payment. We called this measure "absolute confidence". For the latter, they earned an extra 10 points bonus payment if they answered the question correctly within a 5% range. We called this measure "relative confidence". In both cases, the bonus was only earned if the related real effort task was selected to determine the final bonus payment.

Risk

In Experiment 2, risk preferences were measured in two ways, incentivized and self-reported. The incentivized risk preference task employed in our experiment was a modified version of Holt and Laury (2002), inspired by Friedrichsen et al. (2022). Participants were asked to indicate their preference between an increasingly safe amount and a fixed lottery with two equally likely outcomes (see Figure 6 in Appendix C.2). The switching point was considered to be the measure of an individual's risk preference. Only a single switching point was allowed.

Risk preferences were also elicited using a self-reported 11-item risk preference measure in the exit questionnaire (Dohmen et al., 2011). In the second wave of Experiment 2, participants were additionally asked about their risk preferences in four contexts separately, namely life-related, occupational, financial and health-related.

Altruism

To measure altruism, we followed Dreber et al. (2014). We used an incentivized task in which participants were asked to divide a given amount of money (80 points) between themselves and a charity to elicit their altruism. Dreber et al. (2014) chose the Swedish section of Save the Children in their paper. We picked the American Red Cross as the charity of choice instead as it had previously been used to elicit altruism in a US sample (Ottoni-Wilhelm et al., 2017). Hence, our participants were informed that the money they were willing to donate would be donated to the American Red Cross by the experimenters on their behalf.

Efficiency and Equality

Differences in efficiency and equality preferences between men and women have been reported by Andreoni and Vesterlund (2001). In this study, we implemented their approach by giving our participants a menu of eight decisions. In each decision, participants had a fixed amount of endowment points that they could either share with another person or keep for themselves. The recipient was another participant from Experiment 1 who was unaware of the game. This anonymity setting was inspired by Dana et al. (2006) and allowed us to eliminate any reciprocity or social image concerns.

In each decision, participants faced different relative prices of their own payoff and other person's payoff. These relative prices were called "hold value" and "pass value" respectively and were systematically varied across decisions. In some decisions giving was more efficient, while in others it was not. Similarly, the equality preference implied more giving in some decisions and less in others. In this way, we aimed to replicate the efficiency and equality gap between men and women, namely men being more efficiency-concerned and females more equality-concerned (Andreoni and Vesterlund, 2001).

The final equality preference variable was generated in the following way. For each decision, we multiplied the amount participants decided to keep for themselves by the "hold value" and the amount they decided to give away by the "pass value". We then calculated the difference between the above-mentioned quantities, ending up with eight values for each participant. The

final equality preference variable was then calculated as the average of these eight differences. As for the final efficiency preference variable, instead of taking the difference, we summed up the above amounts and obtained the total payoff generated for the pair for that particular decision. For each decision, we then divided the sum by the maximum amount that the pair could earn from that decision. We thereby ended up with eight ratios per participant. The final efficiency preference variable for each participant was the average of these eight ratios. Given the way the measure was constructed, a person who was very efficiency-concerned would have a value of 1, while one who was very equality-concerned would have a value of 0. For example, if the "hold value" was 1, the "pass value" was 3 and the initial endowment was 40 tokens and if a participant decided to keep 30 for herself and give 10 to the other person, the equality preference for that particular decision would be $30 * 1 - 10 * 3 = 0$, while the efficiency would be $(30 * 1 + 10 * 3)/(0 * 1 + 40 * 3) = 1/2$.

Competition

In Experiment 2, we captured competitiveness using a survey item recently developed by Falucchi et al. (2020). This study showed that the item, which measured participants' agreement with the statement *"Competition brings the best out of me"*, predicted individuals' willingness to compete in the laboratory, as well as the tournament entry task by Niederle and Vesterlund (2007), after controlling for their ability, beliefs, and risk attitudes. With this measure we aim at capturing the gender gap usually found in the literature, namely, men are more competitive than women.[13]

Self-Reported CSRI Items

Finally, participants were asked to report on a 7-point Likert scale how well each of the 90 CSRI items describes themselves. The scale ranges from 1 ("Never or almost never true") to 7 ("Always or almost always true"). The items were presented in random order. Participants had an attention check while answering the CSRI items. This attention check was intended to be used as a robustness check later on.

3.1.3 Replication of Gender Differences

The analysis of the pooled sample shows that we replicate previously found gender differences in absolute and relative confidence in the male domain, risk, equality, efficiency, altruism, and competitive preferences (see Table 7 and 8 in Appendix B.1). Consistent with the results of Dreber et al. (2014), we also find that women are less confident in their relative performance expectations in the female domain using a similar word search task.[14]

[13] We decided to use this approach and not the tournament entry task by Niederle and Vesterlund (2007) because it was more easily implementable in an online setting.

[14] Since in the "relative confidence" measure people were asked "What percentage of people do you think solved more questions correctly than you?" a higher value of the measure indicates lower confidence. Hence a positive sign in front of the coefficient *Female* in the relative confidence measures in all regressions indicates that women are less confident than men.

3.2 Development of the Candidate Identity Measures

This section outlines the steps taken in developing the candidate identity measures using Experiment 2 and Study 1. First, we define the process of attribute selection to be used in the candidate identity measures. Second, we explain the components of each measure. Finally, we dive into detailed explanations of how we formed the measures.

3.2.1 Attribute Selection Based on Within Sample Prediction

For each behavioral trait in Experiment 2 (i.e., measures of confidence, altruism, risk preferences, efficiency and equality concerns, and competitive attitudes), our first goal was to find a set of attributes from the self-reported CSRI items listed in Table 1 that predict behavioral traits beyond the biological sex dummy. Therefore, we performed a LASSO analysis (Tibshirani, 1996) on each choice variable.

We run the LASSO analyses using 88 CSRI attributes. We excluded the attributes *Competitive* and *Willing to take risks* which were part of the original BSRI since they were also used as dependent variables in our case (i.e. competition measured through the question *"Competition brings the best out of me"* and risk measured both through the risk preference task and the self-reported risk preference measure). Since we aimed to create a single gender identity measure to predict all behavioral traits, including *Competitive* and *Willing to take risks* in our measure would generate misleading results when predicting competition and risk preferences as dependent variables. It is important to note however that the original BEMS gender measure included these two items.

The LASSO analyses selected the attributes that were important predictors of each behavioral trait for the first wave of Experiment 2. Using each behavioral trait as the dependent variable, LASSO regressions revealed a total of 70 out of 88 attributes that were predictors of at least one trait.

Unlike OLS regressions, LASSO regressions introduce an additional penalty term into the model that shrinks coefficients for predictors that contribute less to the model's predictive accuracy and hence, due to its ability to handle high-dimensional data and perform variable selection it is a preferable method over OLS in our specific case. Specifically, in the context of LASSO, the model aims to minimize the objective function given by $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}|\hat{\beta_j}|$ where $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ is the sum of Mean Squared Residuals (MSR), $\lambda\sum_{j=1}^{p}|\hat{\beta}_j|$ the penalty term, λ the tuning parameter and $\hat{\beta}_j$ a generic estimated regression coefficient.

To ensure the robustness of variable selection, we aimed to identify a reliable approach that would retain only the variables appearing consistently in the selection process across multiple runs (see Meinshausen and Bühlmann (2010) for what it is known in the literature as "stability selection"). To achieve this, we adopted a methodology that involved running LASSO regression with different random seeds, intending to identify the variables that the LASSO algorithm selected in a significant proportion of iterations.

Random seeds played a crucial role in the initialization of the optimization process. In particular, random seeds introduced variability in the initialization, leading to different paths

of coefficient selection during the optimization process. By running the LASSO regression with different random seeds, we aimed to capture the variability in variable selection paths. Therefore, we conducted a hundred iterations of the LASSO regression, utilizing a cross-validation criterion to determine the optimal model at each iteration.[15] As a result, we generated a pool of the hundred best models based on the cross-validation results for each behavioral trait.

To establish a strict criterion for the inclusion of variables, we set a threshold of fifty out of the hundred selected models. We retained any variable that appeared in more than fifty of the one hundred models for further analysis. This filtration process guaranteed that only variables that were consistently selected by the LASSO algorithm were retained, making our variable selection procedure more reliable and stable.[16]

To ensure the robustness of our results, we repeated the analysis using an adaptive selection criterion, again retaining only those variables that surpassed the 50% threshold and were common to both cross-validation and adaptive approaches.[17]

3.2.2 Structure of Each Measure

Not all of the 70 attributes were selected as important predictors by the LASSO analysis for each behavioral trait. Some attributes were only associated with one trait, while others predicted multiple traits. To ensure that we focused on attributes with broader predictive relevance, we retained only those that emerged as important predictors for at least two behavioral traits. As a result, we refined our list of attributes to 41, from which we began to create our candidate measures.

To form our candidate identity measures, we first needed to classify the 41 selected attributes as feminine, masculine and neutral based on the four categorizations of Study 1 namely, desirability in society at large, desirability in the workplace, gender norm in society at large, and gender norm in the workplace.

Each gender identity measure consists of two components, feminine and masculine. These components were calculated as arithmetic averages of the feminine and masculine attributes respectively. In building the gender identity measures, we followed the two-dimensional ap-

[15] Cross-validation involves splitting the dataset into training and test sets multiple times, fitting the LASSO model with different values of λ on the training set, and evaluating the model's performance on the test set. This tuning parameter technique has the advantage of minimizing overfitting without relying on strong assumptions about the underlying data distribution. A large theoretical literature recommends using a cross-validation criterion to select the value of the tuning parameter (see e.g. Chatterjee and Jafarov (2015) for a literature review or see Bühlmann and Van De Geer (2011); Giraud (2015); Hastie et al. (2015) for a textbook-level discussion.)

[16] In this regard, the literature does not recommend any specific threshold. We started from a threshold of 50% and gradually increased it as a robustness check. Reassuringly, when the threshold is increased to 60% or even 70%, the results remain quite robust, with only a few attributes being dropped in all the measured dimensions. When pushing the threshold to 90% or above though, only very few attributes survive.

[17] There are different ways of fitting a LASSO model. One method of selection is cross-validation (CV), and it is more suitable for predictions. The criterion is the CV function f (λ), an estimate of the out-of-sample prediction error, which we minimize. The model for the λ that minimizes the CV function is the selected model. Another method, adaptive lasso, is more suitable when the goal is to find parsimonious models (i.e. models with fewer variables in them) that might better reflect the true model. Adaptive LASSO starts by finding the CV solution and then, using weights on the coefficients in the penalty function does another LASSO and selects a model that has fewer variables. As a robustness check, we performed both and kept the variables common to both approaches. Reassuringly, variables selected by the two approaches were almost always identical.

proach as Bem (1974). Thus, we did not use the attributes that were classified as neutral in constructing our candidate measures.

We believe that including femininity and masculinity scales separately allows us to identify which gender identity correlates with which behavioral trait separately, providing a richer analysis with respect to the unidimensional approach. Moreover, it allows participants to rate themselves both very low or very high in both masculinity and femininity which would not be possible on a unidimensional scale. Indeed, around 10% of our participants scored 1 and 2 (very low) or 6 and 7 (very high) on both of our final feminine and masculine measures.

3.2.3 Candidate Identity Measures

The first candidate measure which we named "GSfeminine2" and "GSmasculine2" encompassed all the attributes selected by LASSO as important predictors of at least two behavioral traits, and classified as feminine and masculine based on Study 1. Feminine and masculine components of each measure were obtained as the arithmetic mean of feminine and masculine attributes. To obtain "GSfeminine2" ("GSmasculine2"), we, therefore, took the arithmetic average of all attributes that have been selected by the LASSO analyses as important predictors for at least two behavioral traits and that have been classified as feminine based on the desirability in society at large (see Table 9 in Appendix B for a complete list of LASSO selections).

Subsequently, we generated two condensed versions of the above by selecting attributes that appeared in at least three or four behavioral traits and named them "GSfeminine3" and "GSmasculine3", and "GSfeminine4" and "GSmasculine4", respectively. As a result, three candidate identity measures were generated alone for the first categorization, namely desirability in society at large.

The second categorization, desirability in the workplace, resulted in two different candidate identity measures based on the frequency of appearance of the attributes in our LASSO analyses, "WPfeminine2" and "WPmasculine2", and "WPfeminine3" and "WPmasculine3". There were no attributes selected more than three times and classified as masculine based on desirability in the workplace.

Attributes were then classified by the third categorization as very masculine/very feminine based on the gender norm in society at large. As for the first categorization, this resulted in three candidate identity measures based on the frequency of appearance in LASSO analyses, each comprising of two components, "KWGSfeminine2" and "KWGSmasculine2", "KWGSfeminine3" and "KWGSmasculine3", and "KWGSfeminine4" and "KWGSmasculine4". Finally, the fourth categorization, gender norms in the workplace, also yielded three candidate identity measures, "KWWPfeminine2" and "KWWPmasculine2", "KWWPfeminine3" and "KWWPmasculine3", and "KWWPfeminine4" and "KWWPmasculine4".

These steps resulted in eleven candidate identity measures of gender identity. Table 10 in Appendix B provides a summary of their names and the number of attributes that belong to each measure.

3.3 New Gender Identity Measure

Following the creation of our candidate identity measures, the next step was to pick the best one. In this regard, the second wave of Experiment 2 served to test them on a fresh *test sample* to select the best-performing alternative in terms of predictive power.

Heilman (2012) claims that what is considered typical for women differs from what is considered necessary in the workplace. They suggest that this discrepancy is due to the masculinity of the workplace context. More specifically, women are expected to be more masculine in the workplace than in society at large.

Following this argument, we expect that if an attribute is persistently feminine in the workplace, where women are expected to behave more masculine, then it is one of the most feminine persistent traits. For example, we found that the attribute *friendly* was more desirable for women than for men in society at large, whereas this difference was no longer significant in the workplace context. On the other hand, the attribute *soft-spoken* appears to be more desirable for women only, both in society at large and in the workplace. This shows that being *soft-spoken* is a more feminine trait than being *friendly*, as it remains feminine also in the workplace. The same is true of masculinity. If a trait is seen as masculine (e.g., *dominant*) in both the society at large and the workplace context, it is likely to be one of the more prominent masculine traits. Therefore, we argue that persistent attributes in the workplace context are stronger identifiers of gender than those in the society at large, and thus the measure of gender identity created by workplace categorizations would be a better tool for predicting gender differences in behavior.

Hypothesis 1. *Work-based gender identity measures perform better than societal ones in predicting confidence, altruism, risk, competition, efficiency and equality preferences.*

We tested the candidate identity measures to make out-of-sample predictions on the *test sample*.[18] We set out a hierarchy of four criteria to identify our preferred measure. If after the first step, no measure was strictly preferred (i.e. in the case of ties) then the second step was applied, and so on. The first step consisted of checking the inclusion of which candidate measure absorbed the effect of the *Female* variable the most (i.e. biological sex dummy) in terms of the statistical significance of the coefficient and in how many traits. The second step consisted of ranking the magnitude of the coefficient *Female* after the inclusion of the gender identity measure. The third step consisted of measuring the significance of the reduction of the coefficient *Female* from the model with only itself and the models including the variable *Female* and the new measure. The fourth step consisted of comparing the R^2 of the models including the measure and looking at the number of attributes included in each one. Fortunately, we could stop already at the first step. The measure consisting of "WPfeminine2" and "WPmasculine2" absorbed the significance of *Female* dummy the most in four out of ten behavioral traits, while the competing measures "KWGSmasculine2-KWGSfeminine2", "KWWPmasculine2-KWWPmasculine2",

[18] For each behavioral trait we have 11 models each including one of the 11 measures and the biological sex dummy, *Female*, plus a model including only the biological sex dummy. Since we measured 14 traits we have in total 12*14 = 168 models. The models are reported in tables 11 to 24 in Appendix B.3.1

in only three out of ten (see Tables 11 to 24 in Appendix B.3.1 for detailed regressions used in the model selection process).[19]

Therefore, we find support for Hypothesis 1. We chose the workplace desirability measure, including all attributes that appeared in at least two regressions, as our continuous gender measure (previously "WPfeminine2" and "WPmasculine2" hereafter called *WP_feminine* and *WP_masculine* for simplicity). *WP_feminine* constitutes 7 attributes and *WP_masculine* 9. The attributes (in alphabetical order) that constitute *WP_feminine* are: *affectionate, compassionate, feminine, flatterable, gullible, sensitive to others needs, tender.* The ones that constitute *WP_masculine*: *acts as a leader, analytical, assertive, athletic, broad, dominant, masculine, strong personality, willing to take a stand.*

Result 1. *Among our four categorizations, workplace desirability performed better than society at large desirability while society at large norms performed better than workplace norms in predicting confidence, altruism, risk, competition, efficiency, and equality preferences according to our selection mechanism.*

3.4 Out-of-Sample Performance of the New Gender Identity Measure

Our ultimate goal is to determine whether our selected measure, *WP_feminine* and *WP_masculine*, can more accurately explain gender differences in the behavioral traits compared to the binary sex, the CGI of Brenøe et al. (2022) and the BEMS measure of masculinity and femininity of Bem (1974).

Hypothesis 2. *The new gender identity measure predicts confidence, altruism, risk, competition, efficiency, and equality preferences better than the binary sex, the single-item masculinity/femininity measure CGI and BEMS gender identity.*

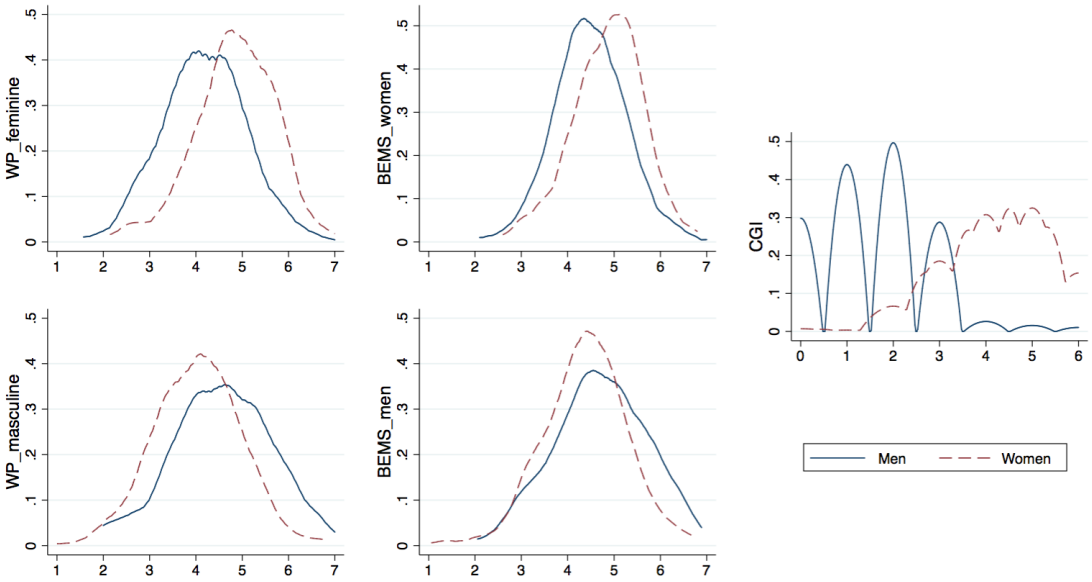
To start, we first show the heterogeneity captured by the three continuous gender identity measures in question. We use kernel density estimation and plot the probability density functions of femininity and masculinity for men and women separately. Figure 2 shows the probability densities of all three continuous gender identity measures. The two-dimensional measures, including our proposed approach and BEMS, capture notable variations in masculinity and femininity within individuals of both biological sexes. These measures reveal distributions where masculine traits among men and feminine traits among women are not as polarized. The figure shows that Bems_women and Bems_men perform very similarly to *WP_feminine* and *WP_masculine*. On the other hand, the self-rating of men and women on the CGI appears to be closer to the binary sex representation of the two groups compared to both two-dimensional measures.

Moving forward to our model comparisons, we regress our new gender identity measure, the comparable BEMS classification (hereafter called "BEMS_men" and "BEMS_women")[20] and the continuous gender identity measure (CGI) on each of the ten behavioral traits: absolute

[19] We are referring to ten and not fourteen traits here because we consider risk (life), risk (occupation), risk (finance) and risk (health) as variants of the main questionnaire risk elicitation question.

[20] "BEMS_men" does not contain the attributes *Competitive* and *Willing to take risks.*

Figure 2: Heterogeneity in femininity and masculinity between sexes.



Notes: The probability density functions (PDFs) of femininity and masculinity for men and women for three distinct gender identity measures. The densities are estimated with univariate Epanechnikov kernel density estimation (KDE), providing a smooth representation of the distribution of these traits within the sample. For WP_feminine, WP_masculine, BEMS_women and BEMS_men the femininity and masculinity scores are represented in two different graphs by their two-dimensional nature. The x-axes represent increasing scores of femininity for WP_feminine and BEMS_women, while the x-axes represent increasing scores of masculinity for WP_masculine and BEMS_men. For the CGI measure, one axis ranges from very masculine (0) to very feminine (6).

confidence in math, relative confidence in math, absolute confidence in word, relative confidence in word, risk (Holt and Laury), risk survey question, altruism, equality, efficiency, and competition. We created four models for each trait: 1. biological sex dummy only, 2. our gender identity measure with the biological sex dummy, 3. BEMS with the biological sex dummy, and 4. CGI with the biological sex dummy.

When comparing the four models, we focus on three key aspects: the reduction in the *Female* coefficient, improvements in model selection criteria (such as adjusted R^2 and Root Mean Square Error), and the significance of the gender identity measure. It is important to clarify that we do not have a predetermined hierarchy among these criteria; instead, we conduct a comprehensive evaluation that considers all relevant metrics equally.

The second model, which includes *WP_feminine* and *WP_masculine*, often shows a substantial reduction in the *Female* coefficient across traits like risk, competition, and efficiency, reflecting the ability of these gender identity components to better capture variations previously attributed to biological sex alone. Additionally, this model frequently outperforms the BEMS and CGI models in terms of adjusted R^2, particularly in traits such as absolute confidence and risk, indicating an improved model fit and stronger explanatory power. The significance of the *WP_feminine* and *WP_masculine* coefficients in these models highlights their relevance in explaining behavior. Table 4 presents all 40 regressions across the 10 behavioral traits, while Figure 3 visualizes the impact of each gender identity measure on the *Female* coefficient. Fur-

thermore, results from the Wald Test, shown in Table 29 in Appendix B.3.3, provide a detailed comparison of the significance of the *Female* coefficients across models.

Before moving on to examine each behavioral trait in more detail, it is useful to recall that all three continuous gender identity models include the biological sex dummy, *Female*, alongside the gender identity measures. If the inclusion of a continuous gender identity measure significantly increases the model fit and reduces the magnitude and significance of the biological sex dummy, *Female*, compared to a model that includes only the binary sex variable, It suggests that the binary variable was previously capturing some effects of gender identity traits. In other words, the coefficient of the biological sex dummy partially absorbs the effect of different attributes correlated with being biologically male or female that explain the dependent variable in question, resulting in an upward bias in the coefficient of *Female* when interpreted as a biological sex difference. However, it is important to note that if we are interested in the average gap between men and women—rather than disentangling biological influences from societal ones—the biological sex dummy *Female* remains useful. In this context, not including the gender identity measure is not a bias, as we are satisfied with a coarser measure that encompasses both biological and social traits.

Absolute and Relative Confidence. Considering absolute and relative confidence in a male task (Columns 1 and 2 of Table 4), specifically solving math problems, the model including our candidate gender measure achieves the best model fit, as indicated by the highest adjusted R^2. In terms of absorbing the significance of *Female*, the model with CGI performs better than our measure. Additionally, our measure reveals that confidence in a male task is positively correlated with masculinity attributes, while the feminine component shows no significant correlation.

In terms of absolute confidence in a female task (Column 3 of Table 4), such as the word task used in this experiment, none of the continuous gender measures provides extra fit in terms of statistical significance or absorbs the effect of *Female* coefficient. While our measure offers a slight improvement in adjusted R^2, it is not substantial. In this task, 55% of participants correctly guessed their score, resulting in a steeper distribution around zero. In contrast, the absolute confidence in the math task has a greater variance (Variance Ratio Test, p-value $<$ 0.0001). Therefore, our study might be underpowered to test any gender difference for this specific task. Regarding relative confidence (Column 4 of Table 4), our measure absorbs the *Female* coefficient the most. In terms of model fit, our model performs the best, but again the difference is not substantial.

It is also worth noting that, in the female task, we find that women exhibit lower confidence levels than men both in absolute and relative terms. This finding aligns with the results of Dreber et al. (2014) but contrasts with (Exley and Kessler, 2022). The difference may be due to the task itself, as we use a similar task to Dreber et al. (2014) in Experiment 2.

Risk. When comparing the models based on their ability to absorb the *Female* coefficient, our gender identity measure stands out, absorbing more of the *Female* coefficient than both the BEMS and CGI models (see Figure 3). Additionally, the difference in the *Female* coefficient between the model using our measure and the BEMS model is statistically significant ($p < 0.001$ in a Wald test), underscoring the stronger explanatory power of our measure (see Table 29 in

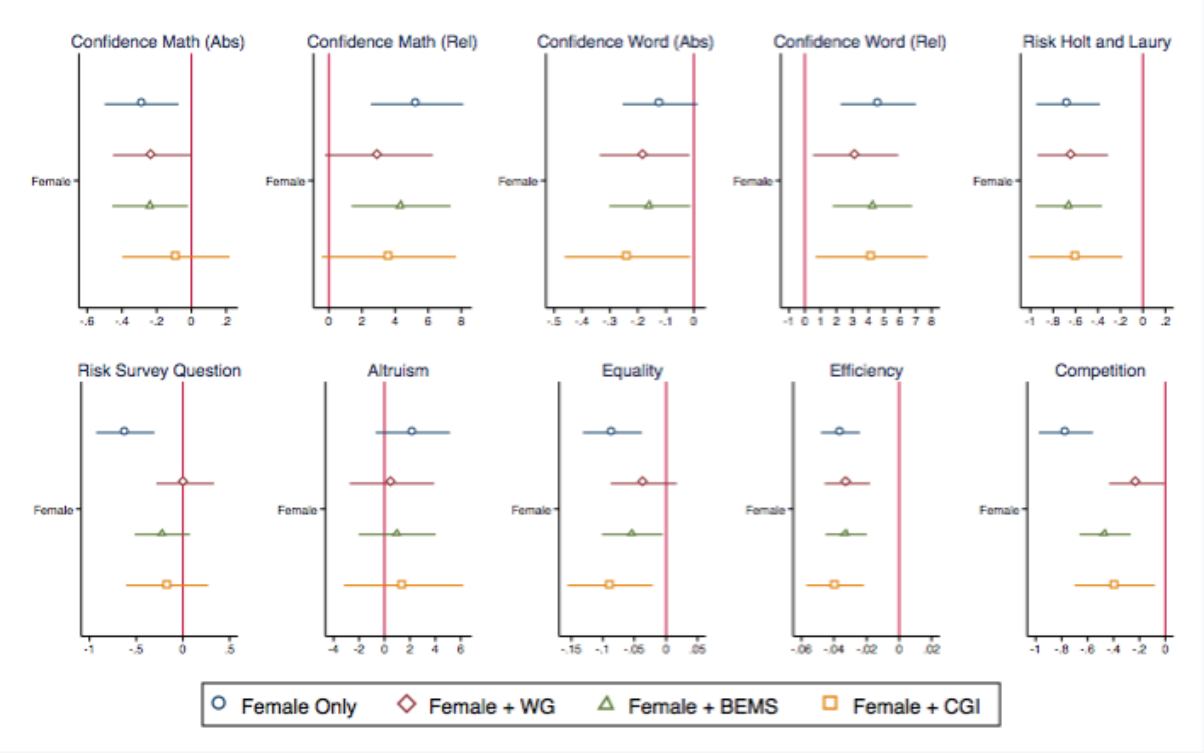
Table 4: Predictive power of gender identity measures on behavioral traits.

Dependent Variable	Confidence Math (Absolute)	Confidence Math (Relative)	Confidence Word (Absolute)	Confidence Word (Relative)	Risk (Holt and Laury)	Risk (Survey Question)	Altruism	Equality	Efficiency	Competition
Female	-0.4706**	5.5759**	-0.0055	2.4287	-0.7860***	-0.6623**	1.0190	-0.1113***	-0.0405***	-1.0165***
	(0.1557)	(1.9133)	(0.0948)	(1.6738)	(0.1954)	(0.2189)	(2.0278)	(0.0334)	(0.0086)	(0.1449)
adj. R^2:	0.5526	0.2180	0.5879	0.0913	0.0574	0.0585	0.0008	0.0301	0.0645	0.1114
RMSE:	1.8007	22.8726	1.1377	20.1099	2.2413	2.5618	23.9955	0.3863	0.0997	1.6785
Female	-0.3755*	4.1174	-0.0512	1.8906	-0.8782***	-0.0115	-1.4697	-0.0538	-0.0313**	-0.5539***
	(0.1665)	(2.2508)	(0.1099)	(1.8598)	(0.2167)	(0.2193)	(2.3053)	(0.0376)	(0.0099)	(0.1471)
WP_feminine	0.0159	0.1720	0.0828	-0.0973	0.1077	0.1123	3.6613**	-0.0861***	-0.0144**	0.0501
	(0.0844)	(1.1567)	(0.0556)	(0.9318)	(0.1068)	(0.1104)	(1.1442)	(0.0182)	(0.0047)	(0.0725)
WP_masculine	0.1884*	-2.4079*	0.0247	-1.2301	-0.0378	1.3431***	-0.0873	0.0006	-0.0008	0.9180***
	(0.0805)	(1.0746)	(0.0609)	(0.9464)	(0.1018)	(0.0973)	(1.0059)	(0.0178)	(0.0044)	(0.0627)
adj. R^2:	0.5557	0.2229	0.5887	0.0914	0.0557	0.3041	0.0155	0.0648	0.0774	0.3594
RMSE:	1.7944	22.8001	1.1366	20.1092	2.2432	2.2025	23.8182	0.3794	0.0990	1.4251
Female	-0.4201**	5.0149*	-0.0354	2.5568	-0.8639***	-0.2300	-0.8924	-0.0747*	-0.0342***	-0.7371***
	(0.1592)	(2.0622)	(0.1001)	(1.7388)	(0.2027)	(0.2103)	(2.1177)	(0.0344)	(0.0090)	(0.1389)
BEMS_women	0.0058	-0.1731	0.0973	-0.7093	0.1168	0.1115	4.6000***	-0.1026***	-0.0183***	0.1100
	(0.1013)	(1.2959)	(0.0649)	(1.0599)	(0.1239)	(0.1314)	(1.2929)	(0.0206)	(0.0055)	(0.0890)
BEMS_men	0.1433	-1.6898	0.0211	-0.4417	-0.1070	1.3537***	-0.8894	0.0031	-0.0000	0.9128***
	(0.0840)	(1.1168)	(0.0574)	(0.9887)	(0.1078)	(0.1069)	(1.0768)	(0.0192)	(0.0049)	(0.0698)
adj. R^2:	0.5533	0.2188	0.5886	0.0894	0.0565	0.2724	0.0177	0.0667	0.0800	0.3300
RMSE:	1.7993	22.8604	1.1367	20.1310	2.2422	2.2522	23.7909	0.3790	0.0988	1.4575
Female	-0.2908	3.5980	-0.1216	2.8977	-0.7463*	-0.1613	-1.8630	-0.0854	-0.0328*	-0.6662**
	(0.2152)	(3.0084)	(0.1429)	(2.5910)	(0.3032)	(0.3152)	(3.2858)	(0.0481)	(0.0133)	(0.2222)
CGI	-0.0672	0.7386	0.0435	-0.1759	-0.0148	-0.1864*	1.0724	-0.0096	-0.0029	-0.1304*
	(0.0646)	(0.8623)	(0.0452)	(0.7420)	(0.0847)	(0.0929)	(0.9231)	(0.0138)	(0.0036)	(0.0627)
adj. R^2:	0.5527	0.2178	0.5880	0.0898	0.0558	0.0637	0.0018	0.0293	0.0640	0.1172
RMSE:	1.8006	22.8760	1.1375	20.1263	2.2431	2.5548	23.9835	0.3865	0.0997	1.6730

OLS with robust standard errors. On the rows: 4 different models per behavioral trait: 1. *Female* alone, 2. *Female* + our new gender identity measure (*WP_feminine* and *WP_masculine*), 3. *Female* + BEMS without attributes *willingness to take risk* and *competitiveness* (*BEMS_women* and *BEMS_men*), 4. *Female* + *CGI*. On the columns, ten different behavioral traits as dependent variables: Absolute confidence – higher is more confidence, relative confidence – lower is more confidence, *Risk (Holt and Laury)* and *Risk (Survey Question)* – higher is more risk taking, *Altruism* – higher is more altruism, *Equality* – lower is more inequality aversion, *Efficiency* – higher is more efficiency preference, *Competition* – higher is more competitiveness. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix B.3.3). In terms of model fit, our measure also demonstrates the best performance, as indicated by the highest adjusted R^2 in the risk survey question (Column 6 of Table 4). We further extend our analysis to different contexts of risk preference – life, financial, occupational, and health-related – in Appendix B.3.2, and we observe the same pattern in terms of adjusted R^2 and the significance of the coefficients across all these domains.

Figure 3: Comparison of the *Female* coefficient among four models for each behavioral trait.



Notes: Coefficient plot of the coefficients of *Female* in 40 regressions of Table 4 grouped by behavioral trait. Each behavioral trait shows the coefficient of *Female* in the following order: 1. *Female* alone, 2. *Female* + *WP_feminine* and *WP_masculine*, 3. *Female* + *BEMS_women* and *BEMS_men* (BEMS without Risk and Competitiveness), 4. *Female* + CGI

Turning to risk elicitation based on Holt and Laury (2002), this dependent variable shows that none of the three models provided a better fit or additional explanatory power (Column 5 of Table 4). In other words, none of the continuous gender identity measures seem to capture anything different than what the coefficient of the binary sex dummy, *Female*, is already capturing. In this regard, we see the value in raising the previously detected external validity issues of this task since its low-stakes version, as the one we have implemented, was not found to correlate with self-reported risk preferences (Andreoni and Kuhn, 2019; Galizzi et al., 2016). Nonetheless, it is still a fact that there is a biological sex difference in this type of risk preference, which remains unexplained by the additional gender identity measures.

Altruism, Equality, Efficiency, and Competition. Across these four traits (Columns 7, 8, 9, and 10 of Table 4), our model shows distinct strengths when compared to the other three models. In terms of absorbing the *Female* coefficient, our model performs better than the BEMS and CGI models, particularly for equality, efficiency, and competition, where it captures more of the variation attributed to biological sex (see Figure 3). For equality and competition,

the reduction in the *Female* coefficient between our model and the BEMS model is statistically significant ($p < 0.001$ in a Wald test) (see Table 29 in Appendix B.3.3).

Regarding model fit, our model closely matches the BEMS model in terms of adjusted R^2 for altruism, equality, and efficiency. However, in the case of competition, our model outperforms the others, achieving the highest adjusted R^2, even with only 16 attributes compared to BEMS's 38. This suggests that our model offers a more parsimonious solution while maintaining or even improving fit, particularly in competitive settings.

In terms of the significance of the gender identity components, our measure shows that altruism, equality, and efficiency are primarily correlated with femininity, while competition preferences are more strongly associated with masculinity. These findings further highlight the model's ability to distinguish between gendered traits across different domains of behavior.

Summing up, in risk (survey question), competition, altruism, equality and efficiency our measure reduces the coefficient of the binary sex dummy, *Female*, significantly with respect to the model only including *Female*. Furthermore, the increase in adjusted R^2 between the model including the BEMS measure and the one including ours, is very small in confidence, while is substantially higher for the risk survey question and the competition measure. Our adjusted R^2 is slightly lower for the risky choice measure, altruism, equality and efficiency, which might be seen as a compromise considering the fewer attributes that generate our measure. For all the traits where the "Female" coefficient is significant in the starting model with "Female" alone, our measure reduces the significance of the *Female* variable more than the BEMS measure (Table 4).

All in all, the gender identity measure that we are proposing in this study explains behavioral traits as well as or better than the previously suggested measures with the exception of risk, altruism, equality, and efficiency, where the adjusted R^2 is marginally higher for the BEMS measure. Importantly, our measure significantly reduces the reliance on the *Female* coefficient in traits where gender differences were previously significant, outperforming BEMS in this regard. Finally, it provides an easier-to-implement tool to measure continuous gender identity including only 16 attributes instead of 38.[21]

Result 2. *WP_feminine and WP_masculine predict behavioral traits better than earlier measures in terms of R^2 in math and word confidence, risk (questionnaire) and competition. WP_feminine and WP_masculine reduce the significance of the Female coefficient in all measured traits more than the BEMS measure.*

Our proposed measure of gender identity, *WP_feminine* and *WP_masculine*, offers a substantial improvement over traditional binary and continuous gender measures in explaining behavioral traits. By reducing the significance of the *Female* coefficient across nearly all measured traits, our measure reveals a more nuanced understanding of gender identity that exceeds the explanatory power of a binary sex dummy and a single-item gender identity measure. Because of its two-dimensional nature, it reveals whether differences lie in femininity or masculinity. Furthermore, despite incorporating fewer attributes than the BEMS, our measure achieves competitive or superior model fit, particularly in domains such as confidence, risk, and competition.

[21] The comparison between our model and CGI can also be found in Figure 3 and Table 4.

This makes our approach not only more parsimonious but also more effective at capturing the complex relationship between gender identity and behavior, thus contributing valuable insights to the ongoing discourse on gender differences in economic behavior.

3.5 Robustness Checks

As a first robustness check, we conducted the same analyses including only continuous gender measures in the regressions, excluding the variable *Female*. We compare the goodness of fit of a model including our measure alone to models that include only the BEMS measure and only CGI, and to a model that includes only the variable *Female*. This test gives us a clear comparison of the three gender identity measures, eliminating any concern regarding multicollinearity (for correlations see Table 30 in Appendix B.3.4, for regressions see Tables 31 to 40 in Appendix B.3.5). Concerning the confidence in a male task (both absolute and relative), the adjusted R^2 from the model with our measure is higher than the one from the model with the BEMS measure but lower than the one from the model including CGI. The highest adjusted R^2 is the one from the model including only the variable *Female*. The same patterns apply to Holt and Laury (2002) risk elicitation. For confidence in the female task (absolute and relative), the highest adjusted R^2 comes from the model including only CGI. In risk (questionnaire measure), competition and equality the highest adjusted R^2 comes from the model including our measure. For altruism, the BEMS measure has the highest adjusted R^2, closely followed by our model and finally regarding efficiency, the biological sex dummy *Female* has the highest adjusted R^2 closely followed by our measure. Briefly, our model has the highest adjusted R^2 in three out of ten behavioral traits, whereas the biological sex dummy in four out of ten, CGI in two and the BEMS measure in one.

A second robustness check is to run the regressions in Table 4 without controls to demonstrate that our results are not affected by the demographic variables, namely age, education, employment status and ethnicity. For both absolute and relative confidence in a math task, our model remains the one with the highest adjusted R^2. On absolute confidence in a female task, our measure and the BEMS measure perform very similarly, but on relative confidence, our measure outperforms all others in both the model selection criterion and the coefficient on the biological sex dummy. Concerning preferences for risk, altruism, equality, efficiency, and competition, our results are robust to the exclusion of controls.

As a third robustness check, we address the potential issue of a single one of our attributes in both components of our two-dimensional measure driving our results. To address this issue, we perform an iterative exclusion analysis. This means that we systematically exclude one attribute at a time from the feminine (masculine) component of our measure while leaving the masculine (feminine) component untouched. We run 7 different versions of *WP_feminine* keeping *WP_masculine* constant and 9 different versions of *WP_masculine* keeping *WP_feminine* constant, excluding one attribute at the time. In Appendix B.3.5 we report the median coefficients of *Female*, *WP_feminine*, and *WP_masculine* in these 7 (9) versions as, well as the lower (min) and the higher (max) coefficients that appear in the reported versions. We then compare the coefficients of the binary sex dummy *Female*, *WP_feminine*, and *WP_masculine*, and finally

the adjusted R^2. As can be seen from Tables 41 and 42 in Appendix B.3.5, this exercise minimally affects the significance of the coefficients and the model fit. Thus, we conclude that our results are not driven by a spurious correlation between a single attribute in our measure and the dependent variables.

The fourth robustness check is a split-sample analysis. This entails running the regressions presented in Table 4 including *WP_feminine*, and *WP_masculine* on female and male samples separately (see Tables 43 and 44 in Appendix B.3.5). This analysis allows us to observe that masculinity-related behavioral traits, namely confidence, risk, and competition preferences, appear to be related to masculinity for both samples, i.e., confidence, risk-taking, and competitiveness increase with masculinity for both men and women. However, there are points worth noting. First, it can be observed that absolute confidence in gendered domains appears to increase with gender congruency for the opposite sex. This implies that women (men) with higher masculinity (femininity) are more confident in the male (female) domain. Second, the relation between masculinity and relative confidence appears to be more prominent for men. Conversely, the relation of altruism, equality, and efficiency to femininity seems to be driven by men. Nevertheless, the direction of the effects still holds for women.

Finally, to illustrate the usefulness and value of a two-dimensional measure of gender identity, we perform an analysis that collapses our own two-dimensional gender identity measure into a unidimensional one. We define our collapsed measure as *WP_collapsed*, computed as *WP_collapsed=(WP_feminine-WP_masculine)*/2. In the regressions of Table 45 in Appendix B.3.5, we see that the collapsed version does not reflect the statistical significance of our findings for confidence and efficiency, and reduces the adjusted R^2 in all regressions, except Holt and Laury risk elicitation, compared to our main model specification in Table 4. Thus, in addition to losing valuable information in the female and male dimensions, we show that the one-dimensional version of our gender identity measure has less predictive power than its two-dimensional one.

4 Discussion

An important question to ask is whether the femininity/masculinity of attributes changes over time. Holt et al. (1998) perform a validity check on the original BEMS attributes and their desirability in society at large. They find that only two out of forty, namely *loyal* and *childlike*, change their desirability from more desirable for women than men to neutral. Using the data from the first two treatments of Experiment 1 and the femininity/masculinity scores of the original Bem (1974), we follow the steps of their validity check. Our analysis reveals that the only differences between our study and Bem (1974) are observed in the same attributes. This suggests that the gender associations of these attributes remain stable over short periods. However, we still recognize the importance of a timely validity check.

It is also important to recognize that alternative frameworks for defining gender identity could be developed through different categorization methods. We acknowledge that our methodology may not encompass all possible predictive approaches, as there are various ways to organize and classify attributes that could yield different results. Given the vast array of potential

categorization methods, we had to make specific choices in our approach. We opted to focus on desirability in society at large to align with Bem's original work. Additionally, we extended our analysis to the workplace context, given the well-documented differences in perceptions of femininity and masculinity between societal and workplace settings (Heilman, 2012). Furthermore, we incorporated an examination of social norms, as they play a crucial role in explaining behavior (see e.g. Bursztyn et al. (2020)). With this in mind, we strongly support expanded research efforts in this domain.

An additional criticism of the BEMS gender identity measure was that its main data were developed using a student sample (Ballard-Reisch and Elton, 1992; Hoffman and Borders, 2001). Our study, which recruited a large sample of the US general population from Prolific, bypasses these claims.

Another noteworthy aspect is the potential to expand existing continuous gender measures to include a more cultural dimension. In this paper, we focus on the gender of attributes specifically within American society and the American workplace, as has been done in the entirety of previous literature on continuous gender identity. While we acknowledge the value of future research that extends to other cultural contexts, it is important to note that studies such as (Löckenhoff et al., 2014) have demonstrated that gender stereotype differences remain consistent across cultures. This suggests that our findings in the US context may have broader relevance beyond this specific setting.

Finally, it is also critical to emphasize that the analyses of gender identity in this paper are correlational rather than causal. It is difficult to determine whether a certain behavioral trait is a result of being more feminine or masculine, or vice versa. However, demonstrating the correlation between the two is a necessary first step in exploring this relationship.

5 Conclusion

This paper develops a gender identity measure with a feminine and a masculine component to avoid multicollinearity problems due to unidimensionality. Each component is made of different attributes and they form a gender identity measure minimizing the influence of demand effects. Furthermore, it aims to improve the model fit in predicting behavioral traits and to reduce potential misinterpretation of the biological sex dummy alone. With the new gender identity measure we provide, we successfully meet our initial objectives for many behavioral traits and significantly reduce the number of attributes compared to the BEMS gender identity measure, resulting in a more streamlined and practical tool for surveys. Our studies further reveal new insights into previously observed gender disparities in economic decision-making. In particular, our findings demonstrate a correlation between confidence, risk-taking, competitiveness and masculine traits. Whereas altruism, efficiency, and equality concerns are rather explained by feminine ones.

The main contributions of our paper can be summarized in three points. First, we present the Contemporary Sex Role Inventory (CSRI), a novel inventory that incorporates work-related attributes and is organized based on four distinct categorizations. Second, we offer new insights for future research in explaining the gender gap by identifying whether the gap is related

to masculinity or femininity traits and providing new possibilities for targeted policymaking. Third, and most importantly, our study provides a validated parsimonious tool that can be used in addition to the biological sex dummy to better disentangle gender differences beyond the binary classification.

To address the gender gap traditionally defined by biological sex, various institutional and policy changes have been proposed in the behavioral economics literature. Our study suggests that these strategies, initially designed to reduce the male-female gap, can be adapted to promote greater inclusion by acknowledging a broader spectrum of gender traits. We show that not every man and woman are the same; their gender identity can significantly differ. In this light, for instance, encouraging team-based competitions (Dargnies, 2012) could help balance individuals with lower masculine traits by pairing them with those who exhibit higher levels of these traits. This approach could promote learning and confidence-building within teams. Implementing role modeling programs can also specifically target individuals with low masculine traits. In fact, guidance and role modeling from successful individuals with similar characteristics could help build confidence and competitiveness (Porter and Serra, 2020; Schier, 2020), which are likely to be lower in the aforementioned group compared to the one with high masculine traits. The same group can be targeted using priming techniques (Balafoutas et al., 2018) and providing personalized performance feedback (Berlin and Dargnies, 2016) encouraging risk-taking and competitiveness. Finally, individuals with low masculine traits can also benefit from changing the nature of competition from being against others to being against one's own performance (Apicella et al., 2017; Carpenter et al., 2018; Apicella et al., 2020) making competitive environments more approachable for them. Specific interventions that promote altruistic behavior and fairness can instead target individuals with low feminine traits. Creating reward structures that recognize collaborative efforts and equitable outcomes can encourage these individuals to develop stronger altruistic behaviors (Fehr and Gächter, 2000). By adapting strategies previously used to reduce the male-female gap to address the masculine-feminine gap, it might be possible to create policies that not only tackle traditional biological sex disparities but also foster an inclusive environment that values and encourages a diverse set of traits. This approach could help individuals with low masculine traits to develop greater competitiveness, confidence, and willingness to take risks, and individuals with low feminine traits to develop greater altruism and equality concerns, ultimately promoting a more balanced and diverse economic decision-making landscape.

Ultimately, further understanding the nature of the gender gaps previously identified in the literature is an important step toward more accurate policy-making and better-targeted research. We strongly believe that interventions aimed at closing gender gaps will benefit from our contribution as they can be more efficiently designed to target either masculine or feminine traits depending on the gap in question. Such targeted approaches can increase the efficiency and effectiveness of efforts to close gender gaps. Therefore, policies should consider not only binary sex but also the attributes that our study has pinpointed as crucial in explaining behavioral traits.

References

Adamus, M. (2018). Who doesn't take a risk, never gets to drink champagne: women, risk and economics. *Individual & Society/Clovek a Spolocnost 21*(2).

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *The Quarterly Journal of Economics 115*(3), 715–753.

Andreoni, J. and M. A. Kuhn (2019). Is it safe to measure risk preferences? Assessing the completeness, predictive validity, and measurement error of various techniques. *Unpublished Manuscript. https://www. makuhn. net/s/mCRB WP. pdf*.

Andreoni, J. and L. Vesterlund (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics 116*(1), 293–312.

Apicella, C. L., E. E. Demiral, and J. Mollerstrom (2017). No gender difference in willingness to compete when competing against self. *American Economic Review 107*(5), 136–140.

Apicella, C. L., E. E. Demiral, and J. Mollerstrom (2020). Compete with others? no, thanks. with myself? yes, please! *Economics Letters 187*, 108878.

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*(1), 685–725.

Balafoutas, L., H. Fornwagner, and M. Sutter (2018). Closing the gender gap in competitiveness through priming. *Nature Communications 9*(1), 4359.

Ballard-Reisch, D. and M. Elton (1992). Gender orientation and the bem sex role inventory: A psychological construct revisited. *Sex Roles 27*, 291–306.

Bem, S. L. (1974). Sex role inventory. *Journal of Personality and Social Psychology 42*, 122–162.

Bem, S. L. (1993). *The lenses of gender: Transforming the debate on sexual inequality*. Yale University Press.

Berlin, N. and M.-P. Dargnies (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization 130*, 320–336.

Brenøe, A. A., L. Heursen, E. Ranehill, and R. A. Weber (2022). Continuous gender identity and economics. In *AEA Papers and Proceedings*, Volume 112, pp. 573–77.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American economic review 110*(10), 2997–3029.

Butler, J. (1990). *Gender Trouble: Feminism and the subversion of indentity*. routledge New York.

Carpenter, J., R. Frank, and E. Huet-Vaughn (2018). Gender differences in interpersonal and intrapersonal competitive behavior. *Journal of Behavioral and Experimental Economics 77*, 170–176.

Chatterjee, S. and J. Jafarov (2015). Prediction error of cross-validated lasso. *arXiv preprint arXiv:1502.06291*.

Coffman, K. B., L. C. Coffman, and K. M. Ericson (2024). Non-binary gender economics. Technical report, National Bureau of Economic Research.

Coffman, K. B., L. C. Coffman, and K. M. M. Ericson (2017). The size of the lgbt population and the magnitude of anti-gay sentiment are substantially underestimated. *Management Science 63*(10), 3168–3186.

Constantinople, A. (1973). Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin 80*(5), 389.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic Literature 47*(2), 448–74.

Dana, J., D. M. Cain, and R. M. Dawes (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes 100*(2), 193–201.

Dargnies, M.-P. (2012). Men too sometimes shy away from competition: The case of team competition. *Management Science 58*(11), 1982–2000.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*(3), 522–550.

Dorofeev, S. (2022). On the difficulty of distinguishing gender from sex in economics.

Dreber, A., E. Von Essen, and E. Ranehill (2014). Gender and competition in adolescence: Task matters. *Experimental Economics 17*(1), 154–172.

Eberhardt, M., G. Facchini, and V. Rueda (2023). Gender differences in reference letters: Evidence from the economics job market. *The Economic Journal 133*(655), 2676–2708.

Exley, C. L. and J. B. Kessler (2022). The gender gap in self-promotion. *The Quarterly Journal of Economics 137*(3), 1345–1381.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2022). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*.

Fallucchi, F., D. Nosenzo, and E. Reuben (2020). Measuring preferences for competition with experimentally-validated survey questions. *Journal of Economic Behavior & Organization 178*, 402–423.

Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review 90*(4), 980–994.

Fornwagner, H., B. Grosskopf, A. Lauf, V. Schöller, and S. Städter (2022). On the robustness of gender differences in economic behavior. *Scientific Reports 12*(1), 21549.

Friedrichsen, J., K. Momsen, and S. Piasenti (2022). Ignorance, intention and stochastic outcomes. *Journal of Behavioral and Experimental Economics 100*, 101913.

Galizzi, M. M., S. R. Machado, and R. Miniaci (2016). Temporal stability, cross-validity, and external validity of risk preferences measures: Experimental evidence from a uk representative sample. *SSRN 10.2139/ssrn.2822613*.

Giraud, C. (2015). *Introduction to high-dimensional statistics*, Volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Günther, C., N. A. Ekinci, C. Schwieren, and M. Strobel (2010). Women can't jump?—an experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization 75*(3), 395–401.

Hair, J. F., W. C. Black, B. J. Babin, R. E. Anderson, and R. Tatham (2010). Multivariate data analysis: Pearson education. *Upper Saddle River, New Jersey*.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity*, Volume 143 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. The lasso and generalizations.

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior 32*, 113–135.

Hoffman, R. M. and L. D. Borders (2001). 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.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

Holt, C. L., J. Ellis, et al. (1998). Assessing the current validity of the bem sex-role inventory. *Sex Roles 39*(11), 929–941.

Hyde, J. S., R. S. Bigler, D. Joel, C. C. Tate, and S. M. van Anders (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist 74*(2), 171.

Kachel, S., M. C. Steffens, and C. Niedlich (2016). Traditional masculinity and femininity: Validation of a new scale assessing gender roles. *Frontiers in Psychology 7*, 956.

Kastlunger, B., S. G. Dressler, E. Kirchler, L. Mittone, and M. Voracek (2010). Sex differences in tax compliance: Differentiating between demographic sex, gender-role orientation, and prenatal masculinization (2d: 4d). *Journal of Economic Psychology 31*(4), 542–552.

Kimura, D. (2004). Human sex differences in cognition, fact, not predicament. *Sexualities, Evolution & Gender 6*(1), 45–53.

Knaak, S. (2004). On the reconceptualizing of gender: Implications for research design. *Sociological Inquiry 74*(3), 302–317.

Krupka, E. L., R. Weber, R. T. Crosno, and H. Hoover (2022). "when in rome": Identifying social norms using coordination games. *Judgment and Decision Making 17*(2), 263–283.

Krupka, E. L. and R. A. Weber (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association 11*(3), 495–524.

Löckenhoff, C. E., W. Chan, R. R. McCrae, F. De Fruyt, L. Jussim, M. De Bolle, P. T. Costa Jr, A. R. Sutin, A. Realo, J. Allik, et al. (2014). Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology 45*(5), 675–694.

Lozano, L., E. Ranehill, and E. Reuben (2022). Gender and preferences in the labor market: Insights from experiments. *Handbook of Labor, Human Resources and Population Economics*, 1–34.

Magliozzi, D., A. Saperstein, and L. Westbrook (2016). Scaling up: Representing gender diversity in survey research. *Socius 2*, 2378023116664352.

Markowsky, E. and M. Beblo (2022). When do we observe a gender gap in competition entry? a meta-analysis of the experimental literature. *Journal of Economic Behavior & Organization 198*, 139–163.

Meier-Pesti, K. and E. Penz (2008). Sex or gender? expanding the sex-based view by introducing masculinity and femininity as predictors of financial risk taking. *Journal of Economic Psychology 29*(2), 180–196.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology 72*(4), 417–473.

Muehlenhard, C. L. and Z. D. Peterson (2011). Distinguishing between sex and gender: History, current conceptualizations, and implications. *Sex Roles 64*(11), 791–803.

Nicholson, L. (1994). Interpreting gender. *Signs: Journal of Women in Culture and Society 20*(1), 79–105.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

Ottoni-Wilhelm, M., L. Vesterlund, and H. Xie (2017). Why do people give? testing pure and impure altruism. *American Economic Review 107*(11), 3617–3633.

Palan, S. and C. Schitter (2018). Prolific.ac-A subject pool for online experiments. *Journal of Behavioral and Experimental Finance 17*, 22–27.

Porter, C. and D. Serra (2020). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics 12*(3), 226–254.

Sahi, S. K. (2023). Understanding gender differences in money attitudes: biological and psychological gender perspective. *International Journal of Bank Marketing 41*(3), 619–640.

Schier, U. K. (2020). Female and male role models and competitiveness. *Journal of Economic Behavior & Organization 173*, 55–67.

Sent, E.-M. and I. van Staveren (2019). A feminist review of behavioral economic research on gender differences. *Feminist Economics 25*(2), 1–35.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Wilson, B. D. and I. H. Meyer (2021). Nonbinary lgbtq adults in the united states.

World Health Organization (2024). Gender. Accessed: April 9, 2024.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics 13*, 75–98.

APPENDIX

A Study 1

Table 5: Gender distribution by treatment in Experiment 1.

Treatment	Men	Women	**Total**
GeneralMen	86	66	152
GeneralWomen	77	74	151
WorkMen	80	70	150
WorkWomen	75	75	150
KWGeneral	80	78	158
KWWork	71	83	154
Total	469	446	915

Notes: P-value = 0.587 of a χ^2 test indicates that there is no significant difference in the gender distribution by treatment.

Table 6: List of all attributes grouped according to the categorization in Experiment 1.

Categorization	**Gender**	**Attributes**	**# of attributes**
Desirability (society at large)	feminine	Feminine (-3.55), Sensitive to other's needs (-0.89), Shy (-0.89), Tender (-0.89), Gullible (-0.89), Soft-spoken (-0.83), Gentle (-0.79), Affectionate (-0.77), Cheerful (-0.67), Warm (-0.67), Flatterable (-0.66), Eager to soothe hurt feelings (-0.66), Sympathetic (-0.66), Compassionate (-0.61), Yielding (-0.50), Friendly (-0.49), Happy (-0.45), Does not use harsh language (-0.45), Patient (-0.42), Creative (-0.35), Understanding (-0.34), Loves children (-0.33), Inefficient (-0.24)	23
	masculine	Masculine (3.41), Dominant (1.61), Assertive (1.31), Acts as a leader (1.22), Strong personality (1.20), Leadership ability (1.19), Competitive (1.03), Driven (0.93), Willing to take a stand (0.93), Ambitious (0.87), Hardwork (0.86), Takes on challenging tasks (0.86), Self-sufficient (0.85), Technical (0.82), Athletic (0.80), Aggressive (0.77), Broad (0.76), Independent (0.76), Self-reliant (0.73), Willing to take risks (0.68) Analytical (0.65), Individualistic (0.65), Forceful (0.63), Expert (0.61), Motivated (0.57), Skillful (0.54), Effort (0.51), Is not afraid of difficulties (0.50), Intellectual (0.48), Rigorous (0.46), Challenging (0.44), Defends own beliefs (0.42), Solid (0.42), Active (0.41), Endures difficult situations (0.40), Quick (0.40), Thorough (0.40), Conceited (0.36), Diligent (0.36), Disciplined (0.36), Knowledgeable (0.35), Smart (0.32), Makes decisions easily (0.29), Jealous (0.28), Reliable (0.25), Clear (0.25), Able (0.24), Dedicated (0.23)	48
Desirability (workplace)	feminine	Feminine (-1.70), Soft-spoken (-0.77), Shy (-0.69), Gentle (-0.67), Compassionate (-0.61), Sympathetic (-0.59), Sensitive to other's needs (-0.57), Affectionate (-0.55), Warm (-0.53), Tender (-0.52), Eager to soothe hurt feelings (-0.51), Flatterable (-0.46), Cheerful (-0.44), Gullible (-0.37), Creative (-0.33), Happy (-0.30)	16
	masculine	Masculine (1.73), Diligent (1.39), Insightful (1.34), Willing to take a stand (1.23), Dominant (0.81), Strong personality (0.64), Acts as a leader (0.58), Assertive (0.53), Forceful (0.50), Aggressive (0.49), Competitive (0.49), Leadership ability (0.49), Driven (0.48), Ambitious (0.47), Willing to take risks (0.47), Broad (0.45), Athletic (0.43), Takes on challenging tasks (0.36), Analytical (0.33), Rigorous (0.32), Dedicated (0.23)	21
Gender Norm (society at large)	very feminine	Affectionate, Compassionate, Does not use harsh language, Feminine, Flatterable, Gentle, Jealous, Loves children, Moody, Patient, Sensitive to other's needs, Soft-spoken, Sympathetic, Tender, Theatrical, Understanding, Yielding	17
	very masculine	Active, Acts as a leader, Aggressive, Ambitious, Analytical, Assertive, Athletic, Challenging, Competitive, Conceited, Defends own beliefs, Dominant, Driven, Endures difficult situations, Forceful, Individualistic, Is not afraid of difficulties, Leadership ability, Masculine, Rigorous, Self-reliant, Strong personality, Takes on challenging tasks, Technical, Willing to take a stand, Willing to take risks	26
Gender Norm (workplace)	very feminine	Affectionate, Compassionate, Does not use harsh language, Feminine, Flatterable, Gentle, Loves children, Moody, Sensitive to other's needs, Soft-spoken, Sympathetic, Tender, Theatrical, Understanding	14
	very masculine	Acts as a leader, Aggressive, Ambitious, Analytical, Assertive, Athletic, Challenging, Competitive, Conceited, Defends own beliefs, Dominant, Driven, Endures difficult situations, Forceful, Leadership ability, Masculine, Rigorous, Strong personality, Technical, Willing to take a stand, Willing to take risks	21

Notes: Femininity and masculinity rates of attributes for desirability categorization are provided in parentheses.

B Study 2

B.1 Pooled Sample

Table 7: Gender differences with controls.

	Confidence Math (Absolute)	Confidence Math (Relative)	Confidence Word (Absolute)	Confidence Word (Relative)	Risk (Holt and Laury)	Risk Survey Question	Competition	Equality	Efficiency	Altruism
Female	-0.287**	5.327***	-0.121#	4.639***	-0.666***	-0.617***	-0.768***	-0.0848***	-0.0361***	2.225
	(0.108)	(1.421)	(0.0681)	(1.208)	(0.144)	(0.159)	(0.106)	(0.0239)	(0.00604)	(1.490)
adj. R^2	0.5974	0.2189	0.6086	0.1222	0.0442	0.0488	0.0780	0.0269	0.0428	0.0269
N	1102	1102	1102	1102	1102	1102	1102	1101	1101	1102

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 8: Gender differences without controls.

	Confidence Math (Absolute)	Confidence Math (Relative)	Confidence Word (Absolute)	Confidence Word (Relative)	Risk (Holt and Laury)	Risk Survey Question	Competition	Equality	Efficiency	Altruism
Female	-0.264*	5.501***	-0.0958	4.749***	-0.732***	-0.718***	-0.810***	-0.107***	-0.0388***	3.707*
	(0.106)	(1.390)	(0.0651)	(1.215)	(0.140)	(0.157)	(0.103)	(0.0229)	(0.00595)	(1.454)
adj. R^2	0.5939	0.2081	0.6104	0.0989	0.0232	0.0178	0.0521	0.0185	0.0364	0.0050
N	1102	1102	1102	1102	1102	1102	1102	1101	1101	1102

Robust standard errors in parentheses.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

B.2 Wave 1 - Attributes Selection

Table 9: List of attributes selected by LASSO.

Categorization	Exact frequency in LASSO	Attributes	# of Attributes
Desirability (society at large) feminine	2	Affectionate, Compassionate, Flatterable	3
	3	Sensitive to others needs, Tender, Yielding	3
	4	Feminine, Gullible, Loves children	3
Desirability (society at large) masculine	2	Able, Active, Acts as a leader, Athletic, Broad, Conceited, Defends own beliefs, Dominant, Expert, Is not afraid of difficulties, Masculine, Strong personality, Technical, Thorough, Willing to take a stand	15
	3	Analytical, Assertive, Endures difficult situations, Intellectual, Makes decisions easily, Self sufficient	6
	4	Challenging, Smart	2
Desirability (workplace) feminine	2	Affectionate, Compassionate, Flatterable	3
	3	Sensitive to others needs, Tender	2
	4	Feminine, Gullible	2
Desirability (workplace) masculine	2	Acts as a leader, Athletic, Broad, Dominant, Masculine, Strong personality, Willing to take a stand	7
	3	Analytical, Assertive	2
	4	-	0
Gender Norm (society at large) very feminine	2	Affectionate, Compassionate, Flatterable	3
	3	Sensitive to others needs, Tender, Theatrical, Yielding	4
	4	Feminine, Moody	2
	5	Loves children	1
Gender Norm (society at large) very masculine	2	Active, Acts as a leader, Athletic, Conceited, Defends own beliefs, Dominant, Is not afraid of difficulties, Masculine, Strong personality, Technical, Willing to take a stand	11
	3	Analytical, Assertive, Endures difficult situations	3
	4	Challenging	1
Gender Norm (workplace) very feminine	2	Affectionate, Compassionate, Flatterable	3
	3	Sensitive to others needs, Tender, Theatrical	3
	4	Feminine, Moody	2
	5	Loves children	1
Gender Norm (workplace) very masculine	2	Acts as a leader, Athletic, Conceited, Defends own beliefs, Dominant, Masculine, Strong personality, Technical, Willing to take a stand	9
	3	Analytical, Assertive, Endures difficult situations	3
	4	Challenging	1

Notes: Attributes are classified according to the categorization in Experiment 1 and to the frequency of appearance. "Exact Frequency in LASSO" means that the reported attributes are selected by LASSO in exactly two/three/four behavioral traits.
Interpreting the Table: Looking at the first row, we can see for instance that the words "Affectionate", "Compassionate", "Flatterable" are classified as feminine based on the society at large desirability category of Experiment 1 and that they are selected by LASSO in exactly two behavioral traits. All other attributes can be interpreted the same way.

Table 10: Final list of 11 candidate identity measures used in final analyses.

Variable Name	**# of Attributes**	**Inclusion Criteria of Attributes**	**Category**
GSfeminine2	9	chosen by LASSO in the prediction of **at least 2 behavioral traits**	**Desirability** (society at large)
GSmasculine2	23		
GSfeminine3	6	...**at least 3**	
GSmasculine3	8		
GSfeminine4	3	...**at least 4**	
GSmasculine4	2		
WPfeminine2	7	chosen by LASSO in the prediction of **at least 2 behavioral traits**	**Desirability** (workplace)[22]
WPmasculine2	9		
WPfeminine3	4	...**at least 3**	
WPmasculine3	2		
KWGSfeminine2	10	chosen by LASSO in the prediction of **at least 2 behavioral traits**	**Gender Norm** (society at large)
KWGSmasculine2	15		
KWGSfeminine3	7	...**at least 3**	
KWGSmasculine3	4		
KWGSfeminine4	3	...**at least 4**	
KWGSmasculine4	1		
KWWPfeminine2	9	chosen by LASSO in the prediction of **at least 2 behavioral traits**	**Gender Norm** (workplace)
KWWPmasculine2	13		
KWWPfeminine3	6	...**at least 3**	
KWWPmasculine3	4		
KWWPfeminine4	3	...**at least 4**	
KWWPmasculine4	1		

Notes: Name of the candidate variable (Variable Name), number of attributes in each variable (# of Attributes), categorization based on the minimum number of appearances in LASSO selection (Inclusion Criteria of Attributes), and categorization based on Experiment 1 (Category). Each candidate measure is obtained as the arithmetic mean of all corresponding attributes.

[22]In the workplace desirability context, there are no masculine attributes that are chosen by LASSO in the prediction of 4 behavioral traits. Therefore, workplace desirability has only 2 candidate identity measures.

B.3 Wave 2

B.3.1 Comparing 11 Candidate Identity Measures for Each Dependent Variable

Table 11

Confidence Math (Absolute)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.471**	-0.381*	-0.455**	-0.473**	-0.376*	-0.526**	-0.331*	-0.399*	-0.325#	-0.372*	-0.448**	-0.494**
	(0.156)	(0.161)	(0.163)	(0.172)	(0.166)	(0.170)	(0.164)	(0.163)	(0.186)	(0.161)	(0.166)	(0.156)
GSfemininе2		-0.0318										
		(0.0916)										
GSmasculine2		0.215*										
		(0.0988)										
GSfemininе3			0.0110									
			(0.0929)									
GSmasculine3			0.144									
			(0.0904)									
GSfemininе4				0.0173								
				(0.0803)								
GSmasculine4				0.144#								
				(0.0745)								
WPfemininе2					0.0159							
					(0.0844)							
WPmasculine2					0.188*							
					(0.0805)							
WPfemininе3						0.0928						
						(0.0846)						
WPmasculine3						0.0834						
						(0.0633)						
KWGSfemininе2							-0.0881					
							(0.0998)					
KWGSmasculine2							0.213*					
							(0.0911)					
KWGSfemininе3								-0.0712				
								(0.100)				
KWGSmasculine3								0.138#				
								(0.0814)				
KWGSfemininе4									-0.120			
									(0.0866)			
KWGSmasculine4									0.0915			
									(0.0567)			
KWWPfemininе2										-0.0597		
										(0.0992)		
KWWPmasculine2										0.173*		
										(0.0790)		
KWWPfemininе3											-0.0250	
											(0.0946)	
KWWPmasculine3											0.0323	
											(0.0463)	
KWWPfemininе4												0.120
												(0.0858)
KWWPmasculine4												0.0179
												(0.0441)
adj. R^2	0.5526	0.5548	0.5534	0.5542	0.5557	0.5535	0.5551	0.5535	0.5542	0.5545	0.5514	0.5531
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 12

Confidence Math (Relative)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	5.576**	4.286*	5.010*	4.705*	4.117#	5.197*	3.717#	4.227*	3.750#	4.164#	4.752*	5.285**
	(1.913)	(2.161)	(2.096)	(2.182)	(2.251)	(2.204)	(2.211)	(2.091)	(2.241)	(2.127)	(2.059)	(1.959)
GSfeminine2		0.588										
		(1.237)										
GSmasculine2		-2.893*										
		(1.278)										
GSfeminine3			0.240									
			(1.161)									
GSmasculine3			-2.575*									
			(1.119)									
GSfeminine4				0.651								
				(1.009)								
GSmasculine4				-2.591**								
				(0.952)								
WPfeminine2					0.172							
					(1.157)							
WPmasculine2					-2.408*							
					(1.075)							
WPfeminine3						-0.435						
						(1.075)						
WPmasculine3						-2.102**						
						(0.802)						
KWGSfeminine2							1.276					
							(1.302)					
KWGSmasculine2							-2.745*					
							(1.211)					
KWGSmfeminine3								1.374				
								(1.238)				
KWGSmasculine3								-2.524*				
								(1.032)				
KWGSfeminine4									1.321			
									(1.020)			
KWGSmasculine4									-1.752*			
									(0.724)			
KWWPfeminine2										0.842		
										(1.300)		
KWWPmasculine2										-2.500*		
										(1.072)		
KWWPfeminine3											0.909	
											(1.157)	
KWWPmasculine3											-1.201#	
											(0.624)	
KWWPfeminine4												-0.0735
												(1.084)
KWWPmasculine4												-1.071#
												(0.608)
adj. R^2	0.2180	0.2224	0.2227	0.2254	0.2229	0.2248	0.2225	0.2243	0.2246	0.2231	0.2205	0.2198
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 13

Confidence Word (Absolute)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.00551	-0.0506	-0.0960	-0.0761	-0.0512	-0.126	-0.0594	-0.102	-0.0608	-0.0538	-0.0919	-0.0382
	(0.0948)	(0.103)	(0.102)	(0.100)	(0.110)	(0.112)	(0.105)	(0.102)	(0.103)	(0.103)	(0.101)	(0.0983)
GSfeminine2		0.0990										
		(0.0603)										
GSmasculine2		0.0404										
		(0.0699)										
GSfeminine3			0.137*									
			(0.0630)									
GSmasculine3			0.0374									
			(0.0609)									
GSfeminine4				0.0746								
				(0.0505)								
GSmasculine4				0.0742								
				(0.0474)								
WPfeminine2					0.0828							
					(0.0556)							
WPmasculine2					0.0247							
					(0.0609)							
WPfeminine3						0.121*						
						(0.0592)						
WPmasculine3						-0.00863						
						(0.0471)						
KWGSfeminine2							0.124#					
							(0.0656)					
KWGSmasculine2							0.0230					
							(0.0658)					
KWGSfeminine3								0.169*				
								(0.0695)				
KWGSmasculine3								0.0119				
								(0.0556)				
KWGSfeminine4									0.0651			
									(0.0574)			
KWGSmasculine4									0.0538			
									(0.0367)			
KWWPfeminine2										0.115#		
										(0.0609)		
KWWPmasculine2										0.0184		
										(0.0601)		
KWWPfeminine3											0.148*	
											(0.0645)	
KWWPmasculine3											0.00127	
											(0.0330)	
KWWPfeminine4												0.118#
												(0.0623)
KWWPmasculine4												0.0110
												(0.0325)
adj. R^2	0.5879	0.5897	0.5915	0.5907	0.5887	0.5902	0.5901	0.5922	0.5900	0.5896	0.5913	0.5907
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 14

Confidence Word (Relative)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	2.429	1.583	2.061	1.469	1.891	2.377	1.439	1.701	1.336	1.528	2.206	2.332
	(1.674)	(1.794)	(1.751)	(1.870)	(1.860)	(1.828)	(1.830)	(1.759)	(1.882)	(1.783)	(1.729)	(1.687)
GSfeminine2		0.334										
		(1.018)										
GSmasculine2		-2.199*										
		(1.086)										
GSfeminine3			0.171									
			(1.033)									
GSmasculine3			-1.951*									
			(0.945)									
GSfeminine4				0.809								
				(0.908)								
GSmasculine4				-2.285**								
				(0.798)								
WPfeminine2					-0.0973							
					(0.932)							
WPmasculine2					-1.230							
					(0.946)							
WPfeminine3						-0.255						
						(0.945)						
WPmasculine3						-0.932						
						(0.752)						
KWGSfeminine2							0.572					
							(1.096)					
KWGSmasculine2							-1.709					
							(1.042)					
KWGSfeminine3								0.723				
								(1.117)				
KWGSmasculine3								-1.590#				
								(0.904)				
KWGSfeminine4									0.850			
									(0.910)			
KWGSmasculine4									-1.070#			
									(0.621)			
KWWPfeminine2										0.629		
										(1.048)		
KWWPmasculine2										-1.666#		
										(0.929)		
KWWPfeminine3											0.310	
											(1.030)	
KWWPmasculine3											-0.254	
											(0.544)	
KWWPfeminine4												0.205
												(0.977)
KWWPmasculine4												-0.230
												(0.541)
adj. R^2	0.0913	0.0944	0.0946	0.1003	0.0914	0.0910	0.0923	0.0934	0.0935	0.0934	0.0886	0.0885
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p< 0.01, *** p<0.001

Table 15

Risk (Holt and Laury)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.786***	-0.836***	-0.829***	-0.835***	-0.878***	-0.912***	-0.818***	-0.827***	-0.823***	-0.851***	-0.809***	-0.824***
	(0.195)	(0.209)	(0.206)	(0.212)	(0.217)	(0.219)	(0.210)	(0.203)	(0.217)	(0.206)	(0.202)	(0.195)
GSfeminine2		0.0577										
		(0.116)										
GSmasculine2		-0.0578										
		(0.126)										
GSfeminine3			0.0322									
			(0.116)									
GSmasculine3			-0.142									
			(0.114)									
GSfeminine4				0.0444								
				(0.101)								
GSmasculine4				-0.0695								
				(0.0971)								
WPfeminine2					0.108							
					(0.107)							
WPmasculine2					-0.0378							
					(0.102)							
WPfeminine3						0.0977						
						(0.106)						
WPmasculine3						-0.0920						
						(0.0835)						
KWGSfeminine2							0.0368					
							(0.124)					
KWGSmasculine2							-0.0334					
							(0.115)					
KWGSfeminine3								0.0190				
								(0.124)				
KWGSmasculine3								-0.132				
								(0.0987)				
KWGSfeminine4									0.0150			
									(0.101)			
KWGSmasculine4									-0.0847			
									(0.0732)			
KWWPfeminine2										0.0879		
										(0.119)		
KWWPmasculine2										-0.0650		
										(0.0991)		
KWWPfeminine3											0.0299	
											(0.115)	
KWWPmasculine3											-0.0275	
											(0.0613)	
KWWPfeminine4												0.114
												(0.0997)
KWWPmasculine4												-0.0342
												(0.0603)
adj. R^2	0.0574	0.0547	0.0569	0.0552	0.0557	0.0575	0.0543	0.0572	0.0566	0.0553	0.0545	0.0566
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 16

Risk (Survey Question)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.662**	-0.219	-0.671**	-0.822***	-0.0115	-0.657**	-0.0374	-0.647**	-0.611**	-0.246	-0.684**	-0.544**
	(0.219)	(0.220)	(0.229)	(0.227)	(0.219)	(0.247)	(0.214)	(0.220)	(0.228)	(0.213)	(0.216)	(0.205)
GSfeminine2		0.0803										
		(0.126)										
GSmasculine2		1.500***										
		(0.120)										
GSfeminine3			0.239#									
			(0.130)									
GSmasculine3			1.017***									
			(0.113)									
GSfeminine4				0.240*								
				(0.106)								
GSmasculine4				0.935***								
				(0.0938)								
WPfeminine2					0.112							
					(0.110)							
WPmasculine2					1.343***							
					(0.0973)							
WPfeminine3						0.261*						
						(0.123)						
WPmasculine3						0.689***						
						(0.0943)						
KWGSfeminine2							0.131					
							(0.127)					
KWGSmasculine2							1.537***					
							(0.107)					
KWGSfeminine3								0.363**				
								(0.134)				
KWGSmasculine3								0.949***				
								(0.106)				
KWGSfeminine4									0.143			
									(0.109)			
KWGSmasculine4									0.757***			
									(0.0732)			
KWWPfeminine2										0.189		
										(0.128)		
KWWPmasculine2										1.237***		
										(0.0999)		
KWWPfeminine3											0.339**	
											(0.123)	
KWWPmasculine3											0.652***	
											(0.0640)	
KWWPfeminine4												0.213#
												(0.110)
KWWPmasculine4												0.680***
												(0.0616)
adj. R^2	0.0585	0.2697	0.1868	0.2051	0.3041	0.1604	0.3192	0.2112	0.2246	0.2805	0.2483	0.2434
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 17

Risk (Life)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.743***	-0.318	-0.707**	-0.834***	-0.109	-0.668**	-0.136	-0.655**	-0.543*	-0.312	-0.704**	-0.627**
	(0.215)	(0.220)	(0.224)	(0.229)	(0.217)	(0.240)	(0.215)	(0.218)	(0.235)	(0.212)	(0.216)	(0.204)
GSfeminine2		0.0500										
		(0.119)										
GSmasculine2		1.390***										
		(0.120)										
GSfeminine3			0.166									
			(0.124)									
GSmasculine3			0.990***									
			(0.111)									
GSfeminine4				0.166								
				(0.105)								
GSmasculine4				0.888***								
				(0.0921)								
WPfeminine2					0.0696							
					(0.104)							
WPmasculine2					1.260***							
					(0.0970)							
WPfeminine3						0.177						
						(0.116)						
WPmasculine3						0.664***						
						(0.0920)						
KWGSfeminine2							0.0665					
							(0.124)					
KWGSmasculine2							1.428***					
							(0.108)					
KWGSfeminine3								0.230#				
								(0.134)				
KWGSmasculine3								0.941***				
								(0.104)				
KWGSfeminine4									-0.00977			
									(0.113)			
KWGSmasculine4									0.732***			
									(0.0729)			
KWWPfeminine2										0.0980		
										(0.123)		
KWWPmasculine2										1.173***		
										(0.0976)		
KWWPfeminine3											0.218#	
											(0.123)	
KWWPmasculine3											0.625***	
											(0.0638)	
KWWPfeminine4												0.183#
												(0.107)
KWWPmasculine4												0.638***
												(0.0614)
adj. R^2	0.0726	0.2577	0.1929	0.2031	0.2926	0.1660	0.2991	0.2144	0.2225	0.2701	0.2401	0.2399
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 18

Risk (Occupation)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.632**	-0.237	-0.621**	-0.720**	-0.106	-0.616*	-0.115	-0.624**	-0.535*	-0.312	-0.683**	-0.528*
	(0.221)	(0.218)	(0.225)	(0.233)	(0.221)	(0.240)	(0.212)	(0.218)	(0.237)	(0.214)	(0.214)	(0.209)
GSfeminine2		0.107										
		(0.124)										
GSmasculine2		1.396***										
		(0.126)										
GSfeminine3			0.212#									
			(0.124)									
GSmasculine3			1.026***									
			(0.113)									
GSfeminine4				0.163								
				(0.103)								
GSmasculine4				0.888***								
				(0.0977)								
WPfeminine2					0.157							
					(0.112)							
WPmasculine2					1.167***							
					(0.107)							
WPfeminine3						0.232*						
						(0.118)						
WPmasculine3						0.645***						
						(0.0965)						
KWGSfeminine2							0.190					
							(0.127)					
KWGSmasculine2							1.358***					
							(0.116)					
KWGSfeminine3								0.352**				
								(0.129)				
KWGSmasculine3								0.893***				
								(0.106)				
KWGSfeminine4									0.0759			
									(0.113)			
KWGSmasculine4									0.669***			
									(0.0756)			
KWWPfeminine2										0.266*		
										(0.128)		
KWWPmasculine2										1.083***		
										(0.105)		
KWWPfeminine3											0.353**	
											(0.121)	
KWWPmasculine3											0.570***	
											(0.0676)	
KWWPfeminine4												0.190#
												(0.112)
KWWPmasculine4												0.601***
												(0.0660)
adj. R^2	0.0629	0.2438	0.1870	0.1846	0.2481	0.1485	0.2689	0.1948	0.1842	0.2386	0.2096	0.2029
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 19

Risk (Finance)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-1.163***	-0.927***	-1.312***	-1.425***	-0.686**	-1.209***	-0.763***	-1.277***	-1.216***	-0.862***	-1.247***	-1.089***
	(0.222)	(0.226)	(0.230)	(0.238)	(0.228)	(0.242)	(0.224)	(0.224)	(0.244)	(0.220)	(0.218)	(0.209)
GSfeminine2		0.242#										
		(0.124)										
GSmasculine2		1.131***										
		(0.141)										
GSfeminine3			0.391**									
			(0.125)									
GSmasculine3			0.754***									
			(0.125)									
GSfeminine4				0.327**								
				(0.108)								
GSmasculine4				0.714***								
				(0.106)								
WPfeminine2					0.163							
					(0.111)							
WPmasculine2					1.083***							
					(0.115)							
WPfeminine3						0.270*						
						(0.118)						
WPmasculine3						0.572***						
						(0.103)						
KWGSfeminine2							0.277*					
							(0.130)					
KWGSmasculine2							1.190***					
							(0.128)					
KWGSfeminine3								0.494***				
								(0.135)				
KWGSmasculine3								0.703***				
								(0.115)				
KWGSfeminine4									0.211#			
									(0.119)			
KWGSmasculine4									0.617***			
									(0.0790)			
KWWPfeminine2										0.263*		
										(0.128)		
KWWPmasculine2										1.034***		
										(0.111)		
KWWPfeminine3											0.396**	
											(0.126)	
KWWPmasculine3											0.532***	
											(0.0697)	
KWWPfeminine4												0.273*
												(0.113)
KWWPmasculine4												0.562***
												(0.0687)
adj. R^2	0.0757	0.2069	0.1596	0.1701	0.2323	0.1444	0.2419	0.1755	0.1866	0.2324	0.2079	0.2031
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 20

Risk (Health)

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.769***	-0.788**	-0.980***	-1.084***	-0.550*	-0.884***	-0.713**	-0.968***	-0.931***	-0.720**	-0.938***	-0.796***
	(0.223)	(0.243)	(0.236)	(0.237)	(0.251)	(0.253)	(0.244)	(0.231)	(0.247)	(0.239)	(0.230)	(0.217)
GSfeminine2		0.300*										
		(0.136)										
GSmasculine2		0.445**										
		(0.150)										
GSfeminine3			0.385**									
			(0.134)									
GSmasculine3			0.317*									
			(0.128)									
GSfeminine4				0.359**								
				(0.113)								
GSmasculine4				0.447***								
				(0.108)								
WPfeminine2					0.167							
					(0.126)							
WPmasculine2					0.612***							
					(0.120)							
WPfeminine3						0.265*						
						(0.128)						
WPmasculine3						0.372***						
						(0.0933)						
KWGSfeminine2							0.384**					
							(0.144)					
KWGSmasculine2							0.536***					
							(0.139)					
KWGSfeminine3								0.526***				
								(0.145)				
KWGSmasculine3								0.411***				
								(0.114)				
KWGSfeminine4									0.280*			
									(0.120)			
KWGSmasculine4									0.466***			
									(0.0806)			
KWWPfeminine2										0.333*		
										(0.141)		
KWWPmasculine2										0.488***		
										(0.122)		
KWWPfeminine3											0.453**	
											(0.139)	
KWWPmasculine3											0.319***	
											(0.0679)	
KWWPfeminine4												0.446***
												(0.119)
KWWPmasculine4												0.341***
												(0.0658)
adj. R^2	0.0093	0.0445	0.0393	0.0649	0.0681	0.0450	0.0670	0.0682	0.0898	0.0624	0.0832	0.0898
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 21

Competition

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-1.016***	-0.728***	-1.014***	-1.118***	-0.554***	-0.943***	-0.592***	-0.924***	-0.875***	-0.664***	-1.010***	-0.938***
	(0.145)	(0.144)	(0.141)	(0.138)	(0.147)	(0.153)	(0.144)	(0.137)	(0.140)	(0.138)	(0.137)	(0.133)
GSfeminine2		0.0669										
		(0.0805)										
GSmasculine2		1.000***										
		(0.0808)										
GSfeminine3			0.169*									
			(0.0760)									
GSmasculine3			0.780***									
			(0.0694)									
GSfeminine4				0.164**								
				(0.0616)								
GSmasculine4				0.733***								
				(0.0621)								
WPfeminine2					0.0501							
					(0.0725)							
WPmasculine2					0.918***							
					(0.0627)							
WPfeminine3						0.148*						
						(0.0746)						
WPmasculine3						0.585***						
						(0.0529)						
KWGSfeminine2							0.0524					
							(0.0846)					
KWGSmasculine2							1.005***					
							(0.0756)					
KWGSfeminine3								0.149#				
								(0.0805)				
KWGSmasculine3								0.764***				
								(0.0612)				
KWGSfeminine4									0.0116			
									(0.0694)			
KWGSmasculine4									0.591***			
									(0.0457)			
KWWPfeminine2										0.0371		
										(0.0794)		
KWWPmasculine2										0.913***		
										(0.0646)		
KWWPfeminine3											0.184*	
											(0.0805)	
KWWPmasculine3											0.424***	
											(0.0407)	
KWWPfeminine4												0.127#
												(0.0697)
KWWPmasculine4												0.438***
												(0.0397)
adj. R^2	0.1114	0.3205	0.2762	0.3063	0.3594	0.2665	0.3504	0.3053	0.3221	0.3598	0.2807	0.2784
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 22

Equality

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.111***	-0.0566	-0.0552	-0.0638#	-0.0538	-0.0377	-0.0607#	-0.0601#	-0.0629	-0.0635#	-0.0657#	-0.0942**
	(0.0334)	(0.0362)	(0.0353)	(0.0372)	(0.0376)	(0.0371)	(0.0364)	(0.0348)	(0.0383)	(0.0353)	(0.0348)	(0.0332)
GSfeminine2		-0.0924***										
		(0.0194)										
GSmasculine2		0.0146										
		(0.0228)										
GSfeminine3			-0.0836***									
			(0.0188)									
GSmasculine3			0.00394									
			(0.0203)									
GSfeminine4				-0.0494**								
				(0.0162)								
GSmasculine4				-0.00737								
				(0.0168)								
WPfeminine2					-0.0861***							
					(0.0182)							
WPmasculine2					0.000627							
					(0.0178)							
WPfeminine3						-0.0804***						
						(0.0171)						
WPmasculine3						-0.00598						
						(0.0140)						
KWGSfeminine2							-0.0979***					
							(0.0212)					
KWGSmasculine2							0.00977					
							(0.0206)					
KWGSfeminine3								-0.0913***				
								(0.0210)				
KWGSmasculine3								0.00478				
								(0.0176)				
KWGSfeminine4									-0.0507**			
									(0.0173)			
KWGSmasculine4									-0.00878			
									(0.0118)			
KWWPfeminine2										-0.105***		
										(0.0207)		
KWWPmasculine2										0.00449		
										(0.0181)		
KWWPfeminine3											-0.0831***	
											(0.0204)	
KWWPmasculine3											0.00143	
											(0.0106)	
KWWPfeminine4												-0.0716***
												(0.0163)
KWWPmasculine4												-0.00337
												(0.0101)
adj. R^2	0.0301	0.0638	0.0585	0.0436	0.0648	0.0607	0.0638	0.0585	0.0439	0.0711	0.0573	0.0580
N	600	600	600	600	600	600	600	600	600	600	600	600

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
Due to a technical problem having to do with saving data, we miss one observation for this behavioral trait.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 23

Efficiency

	1	2	3	4	5	6	7	8	9	10	11	12
Female	-0.0405***	-0.0328***	-0.0330***	-0.0345***	-0.0313**	-0.0314**	-0.0346***	-0.0358***	-0.0373***	-0.0340***	-0.0351***	-0.0379***
	(0.00856)	(0.00956)	(0.00943)	(0.00978)	(0.00988)	(0.00992)	(0.00973)	(0.00950)	(0.0102)	(0.00942)	(0.00941)	(0.00858)
GSfeminine2		-0.0148**										
		(0.00519)										
GSmasculine2		-0.000861										
		(0.00587)										
GSfeminine3			-0.0123*									
			(0.00500)									
GSmasculine3			-0.00462									
			(0.00532)									
GSfeminine4				-0.00646								
				(0.00425)								
GSmasculine4				-0.00372								
				(0.00457)								
WPfeminine2					-0.0144**							
					(0.00475)							
WPmasculine2					-0.000755							
					(0.00442)							
WPfeminine3						-0.0120**						
						(0.00447)						
WPmasculine3						-0.00584						
						(0.00363)						
KWGSfeminine2							-0.0137*					
							(0.00579)					
KWGSmasculine2							-0.00124					
							(0.00540)					
KWGSfeminine3								-0.0103#				
								(0.00578)				
KWGSmasculine3								-0.00378				
								(0.00469)				
KWGSfeminine4									-0.00400			
									(0.00464)			
KWGSmasculine4									-0.00309			
									(0.00317)			
KWWPfeminine2										-0.0166**		
										(0.00564)		
KWWPmasculine2										-0.00181		
										(0.00480)		
KWWPfeminine3											-0.0104#	
											(0.00534)	
KWWPmasculine3											-0.000900	
											(0.00276)	
KWWPfeminine4												-0.0116**
												(0.00416)
KWWPmasculine4												-0.00126
												(0.00267)
adj. R^2	0.0645	0.0770	0.0747	0.0673	0.0774	0.0772	0.0735	0.0700	0.0650	0.0796	0.0692	0.0739
N	600	600	600	600	600	600	600	600	600	600	600	600

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
Due to a technical problem having to do with saving data, we miss one observation for this behavioral trait.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 24

Altruism

	1	2	3	4	5	6	7	8	9	10	11	12
Female	1.019	-1.669	-2.376	-2.789	-1.470	-3.213	-1.251	-2.154	-2.098	-1.338	-1.456	-0.230
	(2.028)	(2.225)	(2.196)	(2.271)	(2.305)	(2.372)	(2.259)	(2.233)	(2.346)	(2.220)	(2.215)	(2.046)
GSfeminine2		4.196***										
		(1.203)										
GSmasculine2		-1.253										
		(1.262)										
GSfeminine3			4.602***									
			(1.225)									
GSmasculine3			-2.153#									
			(1.109)									
GSfeminine4				3.755***								
				(1.065)								
GSmasculine4				-1.728#								
				(0.996)								
WPfeminine2					3.661**							
					(1.144)							
WPmasculine2					-0.0873							
					(1.006)							
WPfeminine3						4.171***						
						(1.186)						
WPmasculine3						-0.778						
						(0.808)						
KWGSfeminine2							4.277**					
							(1.352)					
KWGSmasculine2							-0.509					
							(1.155)					
KWGSfeminine3								5.000***				
								(1.423)				
KWGSmasculine3								-1.716#				
								(0.959)				
KWGSfeminine4									2.928*			
									(1.149)			
KWGSmasculine4									-0.624			
									(0.715)			
KWWPfeminine2										4.436**		
										(1.341)		
KWWPmasculine2										-0.959		
										(0.978)		
KWWPfeminine3											4.242**	
											(1.334)	
KWWPmasculine3											-0.520	
											(0.593)	
KWWPfeminine4												4.576***
												(1.024)
KWWPmasculine4												-0.358
												(0.574)
adj. R^2	0.0008	0.0169	0.0243	0.0226	0.0155	0.0220	0.0160	0.0228	0.0100	0.0170	0.0174	0.0295
N	601	601	601	601	601	601	601	601	601	601	601	601

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

B.3.2 Additional Risk Tables

Table 25: Risk (Life)

	(1)	(2)	(3)	(4)
Female	-0.7429***	-0.1828	-0.3178	-0.1086
	(0.2147)	(0.3194)	(0.2076)	(0.2173)
CGI		-0.2084*		
		(0.0912)		
BEMS_women			0.0613	
			(0.1227)	
BEMS_men			1.2827***	
			(0.1057)	
WP_feminine				0.0696
				(0.1035)
WP_masculine				1.2598***
				(0.0970)
N	601	601	601	601
R^2	0.1020	0.1105	0.2937	0.3174
$adj.R^2$	0.0726	0.0798	0.2680	0.2926
rmse	2.4964	2.4866	2.2178	2.1802

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 26: Risk (Occupation)

	(1)	(2)	(3)	(4)
Female	-0.6322**	-0.2804	-0.2780	-0.1064
	(0.2209)	(0.3276)	(0.2096)	(0.2212)
CGI		-0.1309		
		(0.0953)		
BEMS_women			0.2023	
			(0.1286)	
BEMS_men			1.2199***	
			(0.1124)	
WP_feminine				0.1568
				(0.1119)
WP_masculine				1.1666***
				(0.1073)
N	601	601	601	601
R^2	0.0926	0.0957	0.2660	0.2745
$adj.R^2$	0.0629	0.0645	0.2394	0.2481
rmse	2.5933	2.5910	2.3364	2.3229

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 27: Risk (Finance)

	(1)	(2)	(3)	(4)
Female	-1.1629***	-0.7241*	-0.8793***	-0.6863**
	(0.2218)	(0.3393)	(0.2191)	(0.2280)
CGI		-0.1633		
		(0.0969)		
BEMS_women			0.2454	
			(0.1334)	
BEMS_men			1.0605***	
			(0.1235)	
WP_feminine				0.1625
				(0.1113)
WP_masculine				1.0827***
				(0.1150)
N	601	601	601	601
R^2	0.1050	0.1098	0.2382	0.2592
$adj.R^2$	0.0757	0.0791	0.2105	0.2323
rmse	2.6103	2.6056	2.4125	2.3790

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 28: Risk (Health)

	(1)	(2)	(3)	(4)
Female	-0.7693***	-0.3104	-0.6607**	-0.5498*
	(0.2230)	(0.3429)	(0.2330)	(0.2506)
CGI		-0.1708		
		(0.1025)		
BEMS_women			0.1958	
			(0.1429)	
BEMS_men			0.5081***	
			(0.1286)	
WP_feminine				0.1668
				(0.1259)
WP_masculine				0.6120***
				(0.1200)
N	601	601	601	601
R^2	0.0406	0.0465	0.0787	0.1008
$adj.R^2$	0.0093	0.0136	0.0453	0.0681
rmse	2.5636	2.5581	2.5166	2.4863

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B.3.3 Female Coefficient Comparison Between Models

Table 29: Wald test p-values.

Confidence Math (Absolute)

	M1	M2	M3
M2	.21811242	.	.
M3	.29916949	.28138175	.
M4	.29031392	.57820947	.42230492

Confidence Math (Relative)

	M1	M2	M3
M2	.19385934	.	.
M3	.40840465	.11263997	.
M4	.38334763	.80598335	.51567975

Confidence Word (Absolute)

	M1	M2	M3
M2	.34083769	.	.
M3	.29731702	.55449876	.
M4	.32842008	.52322486	.45993966

Confidence Word (Relative)

	M1	M2	M3
M2	.5217239	.	.
M3	.80784648	.13334107	.
M4	.80928583	.58103054	.8557661

Risk (Holt & Laury)

	M1	M2	M3
M2	.34557869	.	.
M3	.20399762	.78087328	.
M4	.85930617	.51527294	.57923397

Risk (Survey Question)

	M1	M2	M3
M2	6.566e-06	.	.
M3	.00028919	.00041161	.
M4	.04388212	.55173682	.78978975

Risk (Life)

	M1	M2	M3
M2	3.134e-06	.	.
M3	.00019494	.00040496	.
M4	.02180961	.76372408	.59744465

Risk (Occupation)

	M1	M2	M3
M2	.00013791	.	.
M3	.00170014	.00553131	.
M4	.1644947	.49081667	.99278113

Risk (Finance)

	M1	M2	M3
M2	.00035595	.	.
M3	.00703778	.00188455	.
M4	.09059715	.88164392	.55257804

Risk (Health)

	M1	M2	M3
M2	.07966508	.	.
M3	.18738084	.0818471	.
M4	.09419795	.34830683	.18754152

Altruism

	M1	M2	M3
M2	.01424877	.	.
M3	.00316289	.32365794	.
M4	.23706143	.8600421	.68467898

Equality

	M1	M2	M3
M2	.00133839	.	.
M3	.00204257	.02254021	.
M4	.47757377	.34130094	.75853921

Efficiency

	M1	M2	M3
M2	.04149171	.	.
M3	.03024296	.21037473	.
M4	.41615688	.86645512	.88168787

Competition

	M1	M2	M3
M2	5.631e-06	.	.
M3	.00117254	.00001187	.
M4	.03729592	.51581402	.69145759

Notes: M1: Female, M2: Female and WP_feminine & WP_masculine, M3: Female and BEMS_women & BEMS_men, M4: Female and CGI. All models use the same specifications as in Table 4. The coefficient comparisons are executed with the Wald Test following Seemingly Unrelated Estimations of all four models.

B.3.4 Correlations Between Continuous Measures and the Binary Sex Dummy

Table 30: Correlation between different gender identity measures.

	Female	WP_feminine	WP_masculine	BEMS_women	BEMS_men	CGI
Female	1.0000					
WP_feminine	0.3512	1.0000				
WP_masculine	-0.2576	0.1469	1.0000			
BEMS_women	0.2399	0.9085	0.1775	1.0000		
BEMS_men	-0.1755	0.1589	0.9445	0.1909	1.0000	
CGI	0.7501	0.4115	-0.3640	0.2613	-0.2800	1.0000

B.3.5 Robustness Checks

Comparing *WP_feminine* and *WP_masculine*, CGI and the BEMS Measures Excluding the Variable *Female*

Table 31: Confidence Math (Absolute)

	(1)	(2)	(3)	(4)
Female	-0.471**			
	(0.156)			
CGI		-0.127**		
		(0.0462)		
BEMS_women			-0.0600	
			(0.0994)	
BEMS_men			0.199*	
			(0.0847)	
WP_feminine				-0.0588
				(0.0801)
WP_masculine				0.249**
				(0.0795)
adj. R^2	0.0568	0.0554	0.0381	0.0428
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 32: Confidence Math (Relative)

	(1)	(2)	(3)	(4)
Female	5.576**			
	(1.913)			
CGI		1.482**		
		(0.550)		
BEMS_women			0.613	
			(1.263)	
BEMS_men			-2.349*	
			(1.073)	
WP_feminine				0.990
				(1.046)
WP_masculine				-3.077**
				(0.979)
adj. R^2	0.0579	0.0562	0.0406	0.0494
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 33: Confidence Word (Absolute)

	(1)	(2)	(3)	(4)
Female	-0.00551			
	(0.0948)			
CGI		0.0187		
		(0.0300)		
BEMS_women			0.0915	
			(0.0639)	
BEMS_men			0.0253	
			(0.0556)	
WP_feminine				0.0724
				(0.0519)
WP_masculine				0.0324
				(0.0571)
adj. R^2	0.0153	0.0174	0.0013	0.0072
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 34: Confidence Word (Relative)

	(1)	(2)	(3)	(4)
Female	2.429			
	(1.674)			
CGI		0.417		
		(0.480)		
BEMS_women			-0.286	
			(1.061)	
BEMS_men			-0.742	
			(0.953)	
WP_feminine				0.287
				(0.896)
WP_masculine				-1.512#
				(0.883)
adj. R^2	0.0217	0.0218	0.0200	0.0207
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 35: Risk (Holt and Laury)

	(1)	(2)	(3)	(4)
Female	-0.786***			
	(0.195)			
CGI		-0.170**		
		(0.0549)		
BEMS_women			-0.0242	
			(0.125)	
BEMS_men			0.00462	
			(0.107)	
WP_feminine				-0.0712
				(0.103)
WP_masculine				0.104
				(0.0978)
adj. R^2	0.0574	0.0464	0.0282	0.0306
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 36: Risk (Survey Question)

	(1)	(2)	(3)	(4)
Female	-0.662**			
	(0.219)			
CGI		-0.220***		
		(0.0642)		
BEMS_women			0.0740	
			(0.123)	
BEMS_men			1.383***	
			(0.102)	
WP_feminine				0.110
				(0.0978)
WP_masculine				1.345***
				(0.0912)
adj. R^2	0.0585	0.0650	0.2720	0.3053
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 37: Altruism

	(1)	(2)	(3)	(4)
Female	1.019			
	(2.028)			
CGI		0.686		
		(0.570)		
BEMS_women			4.454***	
			(1.240)	
BEMS_men			-0.774	
			(1.069)	
WP_feminine				3.362**
				(1.025)
WP_masculine				0.149
				(0.976)
adj. R^2	0.0008	0.0029	0.0191	0.0165
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 38: Equality

	(1)	(2)	(3)	(4)
Female	-0.111***			
	(0.0334)			
CGI		-0.0273**		
		(0.00957)		
BEMS_women			-0.115***	
			(0.0198)	
BEMS_men			0.0127	
			(0.0189)	
WP_feminine				-0.0970***
				(0.0163)
WP_masculine				0.00930
				(0.0169)
adj. R^2	0.0301	0.0260	0.0606	0.0629
N	600	600	600	600

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 39: Efficiency

	(1)	(2)	(3)	(4)
Female	-0.0405***			
	(0.00856)			
CGI		-0.00966***		
		(0.00232)		
BEMS_women			-0.0238***	
			(0.00518)	
BEMS_men			0.00440	
			(0.00486)	
WP_feminine				-0.0208***
				(0.00420)
WP_masculine				0.00430
				(0.00420)
adj. R^2	0.0645	0.0549	0.0580	0.0619
N	600	600	600	600

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 40: Competition

	(1)	(2)	(3)	(4)
Female	-1.016***			
	(0.145)			
CGI		-0.268***		
		(0.0412)		
BEMS_women			-0.0103	
			(0.0864)	
BEMS_men			1.008***	
			(0.0672)	
WP_feminine				-0.0627
				(0.0644)
WP_masculine				1.007***
				(0.0552)
adj. R^2	0.1114	0.1040	0.2945	0.3426
N	601	601	601	601

Notes: BEMS men measure excludes *Competitive* and *Willing to take risks*. Controls include age, education, employment status, ethnicity and session fixed effects. Robust standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Iterative Exclusion of Attributes

Table 41: on WP_feminine

Confidence Math (Absolute)

	M2	Median	Min	Max
Female	-0.3755*	-0.3817*	-0.4154*	-0.3376*
WP_feminine	0.0159	0.0228	-0.0278	0.0652
WP_masculine	0.1884*	0.1875*	0.1772*	0.2006*
adj. R^2	0.5557	0.5558	0.5557	0.5562

Confidence Math (Relative)

	M2	Median	Min	Max
Female	4.1174	4.1613	3.9207	4.3567
WP_feminine	0.172	0.1827	-0.1296	0.3989
WP_masculine	-2.4079*	-2.4163*	-2.4608*	-2.3352*
adj. R^2	0.2229	0.2229	0.2229	0.2231

Confidence Word (Absolute)

	M2	Median	Min	Max
Female	-0.0512	-0.0448	-0.0767	-0.0207
WP_feminine	0.0828	0.0735	0.0582	0.1113
WP_masculine	0.0247	0.0245	0.0195	0.0316
adj. R^2	0.5887	0.5885	0.588	0.59

Confidence Word (Relative)

	M2	Median	Min	Max
Female	1.8906	1.8507	1.6958	2.2193
WP_feminine	-0.0973	-0.0475	-0.4337	0.1389
WP_masculine	-1.2301	-1.2432	-1.2885	-1.1246
adj. R^2	0.0914	0.0914	0.0914	0.0917

Risk (Holt & Laury)

	M2	Median	Min	Max
Female	-0.8782***	-0.88***	-0.9115***	-0.8378***
WP_feminine	0.1077	0.1059	0.0734	0.1411
WP_masculine	-0.0378	-0.0371	-0.045	-0.0307
adj. R^2	0.0557	0.0558	0.055	0.0567

Risk (Survey Question)

	M2	Median	Min	Max
Female	-0.0115	-0.0039	-0.0433	0.0435
WP_feminine	0.1123	0.099	0.0649	0.1511
WP_masculine	1.3431***	1.3503***	1.3329***	1.352***
adj. R^2	0.3041	0.3038	0.3033	0.3053

Altruism

	M2	Median	Min	Max
Female	-1.4697	-1.5515	-1.9998	-0.2625
WP_feminine	3.6613**	3.5497**	2.3762*	4.0876***
WP_masculine	-0.0873	-0.0318	-0.2071	0.1228
adj. R^2	0.0155	0.0159	0.0069	0.019

Equality

	M2	Median	Min	Max
Female	-0.0538	-0.0539	-0.0823*	-0.0473
WP_feminine	-0.0861***	-0.0831***	-0.0888***	-0.0743***
WP_masculine	0.0006	-0.0006	-0.0022	0.0033
adj. R^2	0.0648	0.0646	0.0605	0.0662

Efficiency

	M2	Median	Min	Max
Female	-0.0313**	-0.0311**	-0.0359***	-0.0302**
WP_feminine	-0.0144**	-0.0141**	-0.0151**	-0.013**
WP_masculine	-0.0008	-0.0011	-0.0014	-0.0002
adj. R^2	0.0774	0.0772	0.075	0.0787

Competition

	M2	Median	Min	Max
Female	-0.5539***	-0.5597***	-0.5726***	-0.5279***
WP_feminine	0.0501	0.0543	0.0292	0.0731
WP_masculine	0.918***	0.9178***	0.912***	0.9235***
adj. R^2	0.3594	0.3595	0.3591	0.3601

Table 42: on WP_masculine

Confidence Math (Absolute)

	M2	Median	Min	Max
Female	-0.3755*	-0.3694*	-0.4189*	-0.3451*
WP_feminine	0.0159	0.0141	0.0083	0.0306
WP_masculine	0.1884*	0.1845*	0.1466*	0.2205*
adj. R^2	0.5557	0.5557	0.5545	0.5568

Confidence Math (Relative)

	M2	Median	Min	Max
Female	4.1174	4.1422	3.5896	4.6683*
WP_feminine	0.172	0.1827	0.0189	0.2938
WP_masculine	-2.4079*	-2.347*	-2.9667*	-2.0048*
adj. R^2	0.2229	0.2228	0.2213	0.2260

Confidence Word (Absolute)

	M2	Median	Min	Max
Female	-0.0512	-0.0522	-0.0612	-0.04
WP_feminine	0.0828	0.0828	0.0791	0.0871
WP_masculine	0.0247	0.0241	0.0101	0.0402
adj. R^2	0.5887	0.5887	0.5885	0.5889

Confidence Word (Relative)

	M2	Median	Min	Max
Female	1.8906	1.8772	1.6201	2.1896
WP_feminine	-0.0973	-0.1227	-0.2234	0.0247
WP_masculine	-1.2301	-1.2723	-1.5196	-0.8405
adj. R^2	0.0914	0.0914	0.0899	0.0926

Risk (Holt & Laury)

	M2	Median	Min	Max
Female	-0.8782***	-0.8748***	-0.9064***	-0.8581***
WP_feminine	0.1077	0.1057	0.0993	0.1186
WP_masculine	-0.0378	-0.0319	-0.0796	-0.0097
adj. R^2	0.0557	0.0557	0.0555	0.0566

Risk (Survey Question)

	M2	Median	Min	Max
Female	-0.0115	0.0185	-0.3161	0.0773
WP_feminine	0.1123	0.1182	0.0658	0.1701
WP_masculine	1.3431***	1.3283***	1.2204***	1.3904***
adj. R^2	0.3041	0.2980	0.2757	0.3264

Altruism

	M2	Median	Min	Max
Female	-1.4697	-1.4789	-1.636	-1.2743
WP_feminine	3.6613**	3.6645**	3.5821**	3.7486**
WP_masculine	-0.0873	-0.0934	-0.3274	0.1649
adj. R^2	0.0155	0.0155	0.0155	0.0157

Equality

	M2	Median	Min	Max
Female	-0.0538	-0.0539	-0.056	-0.0508
WP_feminine	-0.0861***	-0.0861***	-0.0881***	-0.0853***
WP_masculine	0.0006	0.0004	-0.0022	0.0063
adj. R^2	0.0648	0.0648	0.0648	0.0650

Efficiency

	M2	Median	Min	Max
Female	-0.0313**	-0.0315**	-0.0323**	-0.0302**
WP_feminine	-0.0144**	-0.0144**	-0.0149**	-0.0141**
WP_masculine	-0.0008	-0.0009	-0.0019	0.0007
adj. R^2	0.0774	0.0774	0.0773	0.0776

Competition

	M2	Median	Min	Max
Female	-0.5539***	-0.5478***	-0.7549***	-0.4938***
WP_feminine	0.0501	0.0543	0.0137	0.0894
WP_masculine	0.918***	0.9056***	0.8349***	0.9496***
adj. R^2	0.3594	0.3554	0.3306	0.3677

Notes: M2: Main model in Table 4, Median: Median coefficient of iterations, Min: Minimum coefficient of iterations, Max: Maximum coefficient of iterations. All models use the same specifications as in Table 4.

Split Sample

Table 43: Female sample.

	Confidence Math (Absolute)	Confidence Math (Relative)	Confidence Word (Absolute)	Confidence Word (Relative)	Risk (Holt and Laury)	Risk (Survey Question)	Altruism	Equality	Efficiency	Competition
WP_feminine	0.0279	-2.1164	-0.0547	0.3275	-0.1364	0.2152	2.6725	-0.0520	-0.0100	0.2408*
	(0.1225)	(1.6166)	(0.0703)	(1.3684)	(0.1452)	(0.1648)	(1.7520)	(0.0277)	(0.0068)	(0.1219)
WP_masculine	0.3418***	-1.7987	0.0491	1.1167	-0.2038	1.2905***	0.4453	0.0084	-0.0016	0.8945***
	(0.0924)	(1.5197)	(0.0635)	(1.4422)	(0.1669)	(0.1495)	(1.6866)	(0.0289)	(0.0061)	(0.0982)
adj. R^2	0.4816	0.1443	0.6073	0.0679	-0.0095	0.2592	0.0195	0.0241	0.0321	0.2695
N	300	300	300	300	300	300	300	299	299	300

Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 44: Male sample.

	Confidence Math (Absolute)	Confidence Math (Relative)	Confidence Word (Absolute)	Confidence Word (Relative)	Risk (Holt and Laury)	Risk (Survey Question)	Altruism	Equality	Efficiency	Competition
WP_feminine	-0.0030	2.5813	0.2177**	-0.6796	0.2888	-0.0155	4.1055**	-0.1298***	-0.0239**	-0.1574
	(0.1268)	(1.6139)	(0.0785)	(1.3796)	(0.1674)	(0.1586)	(1.5725)	(0.0260)	(0.0072)	(0.0915)
WP_masculine	0.0958	-3.4762*	-0.0367	-2.7449*	0.1004	1.3898***	-0.2483	0.0104	0.0039	1.0017***
	(0.1318)	(1.5439)	(0.1062)	(1.3408)	(0.1379)	(0.1349)	(1.2809)	(0.0244)	(0.0069)	(0.0824)
adj. R^2	0.5778	0.2473	0.5733	0.1049	0.0943	0.3107	0.0061	0.0645	0.0541	0.3926
N	301	301	301	301	301	301	301	301	301	301

Robust standard errors in parentheses.
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Collapsed Gender Identity Measure

Table 45: Collapsed version of WP_feminine and WP_masculine.

	Confidence Math (Absolute)	Confidence Math (Relative)	Confidence Word (Absolute)	Confidence Word (Relative)	Risk (Holt and Laury)	Risk (Survey Question)	Competition	Equality	Efficiency	Altruism
Female	-0.3541*	3.8824	-0.0340	1.6788	-0.8695***	0.1695	-1.0255	-0.0641	-0.0332***	-0.4336**
	(0.1665)	(2.2572)	(0.1108)	(1.8840)	(0.2161)	(0.2450)	(2.2987)	(0.0382)	(0.0099)	(0.1636)
WP_collapsed	-0.1932	2.8065	0.0479	1.2584	0.1384	-1.3780***	3.3871*	-0.0781**	-0.0122	-0.9657***
	(0.1265)	(1.8473)	(0.0811)	(1.4344)	(0.1628)	(0.1808)	(1.6273)	(0.0291)	(0.0074)	(0.1234)
$adj.R^2$	0.5534	0.2203	0.5874	0.0908	0.0569	0.1413	0.0052	0.0408	0.0672	0.2010
N	601	601	601	601	601	601	601	600	600	601

Notes: Robust standard errors in parentheses. Controls include age, education, employment status, ethnicity and session fixed effects.
$W\_collapsed = (W\_feminine + W\_masculine)/2$
# p<0.10, * p<0.05, ** p<0.01, *** p<0.001

C Instructions

The full instructions for the experiments are available upon request.

C.1 Experiment 1

Figure 4: Comprehension question common to all treatments

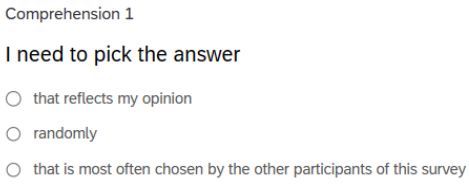
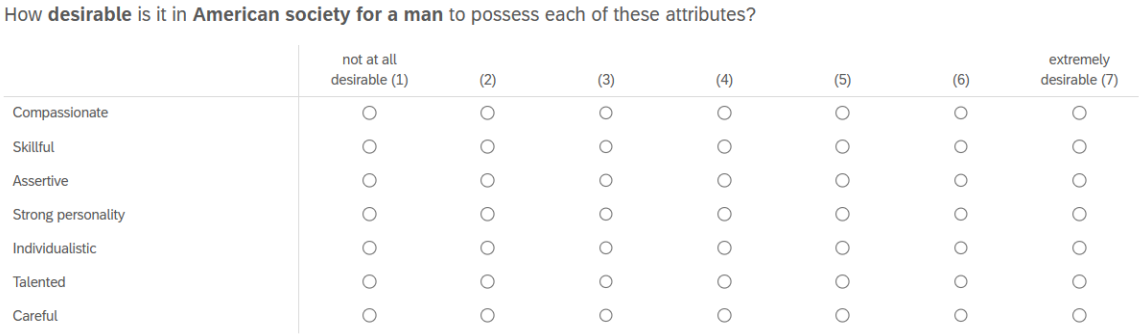


Figure 5: Example question: general desirability men (treatment 1) first page of a list of 14 pages.



C.2 Experiment 2

Figure 6: Risk elicitation task: Holt and Laury.