Economic Behavior and Gender Typicality: The Predictive Power of Femininity and Masculinity*

Stefano Piasenti†, Müge Süer‡

January 7, 2026

Abstract

Are gender gaps in economic behavior labor market outcomes and political preferences driven by biological sex, or by heterogeneity in gender-typical traits? Using a U.S. sample and machine learning, we develop and validate a measure of gender typicality capturing masculinity and femininity as separate dimensions. We show that confidence, competitiveness, and risk preferences are associated with masculinity, whereas altruism and concerns for equality and efficiency are associated with femininity, after accounting for biological sex. Gender typicality also predicts labor market outcomes beyond biological sex: masculinity is associated with higher income, managerial positions, performance-based pay, and willingness to negotiate. Finally, we show that political preferences vary with conformity to gender-typical traits. By leveraging latent heterogeneity in gender typicality, our framework reconciles disparate findings in the gender literature and informs targeted organizational and policy interventions.

JEL-codes: C52, C91, D91, J16, J62
Keywords: Biological sex, Gender identity, Online experiment, Machine learning

*We thank Billur Aksoy, Dirk Engelmann, Tilman Fries, Lavinia Kinne, Dorothea Kübler, Nicola Lacetera, Huyen Nguyen, Regine Oexl, Davide Pace, Pascal Pillath, Giacomo Rubbini, Sebastian Schweighofer-Kodritsch, Bertil Tungodden, Roel van Veldhuizen, Roberto Weber, 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, the CRC Retreat in Schwanenwerder 2024, the Workshop on Behavioural and Experimental Economics in Florence 2025, the Workshop on Gender in Adaptive Design in Karlsruhe 2025 and the SABE Conference in Trento 2025 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`. Our study has been approved by the ethics committee of the School of Business and Economics of Humboldt-Universität zu Berlin (Ethics Approval No. 2022-08).

†University of Bologna (stefano.piasenti2@unibo.it)
‡Halle Institute for Economic Research (IWH) (muege.sueer@iwh-halle.de)

1 Introduction

Economists have extensively studied gender differences in psychological and behavioral traits to understand their implications for economic behavior, financial decisions, labor market outcomes, discrimination, and social stereotypes (Powell and Ansic, 1997; Caliendo et al., 2015; Markowsky and Beblo, 2022; Lozano et al., 2022). Yet findings are often inconsistent, frequently shifting with context and social expectations, making it difficult to derive clear implications for policy (Croson and Gneezy, 2009; Sent and van Staveren, 2019). This highlights the need to distinguish biological sex from other social factors, such as gender identity, and raises a central question to behavioral economics: are behavioral differences inherently tied to biological sex, or do broader dimensions of gender identity also shape them? In this paper, we separate the predictive power of biological sex from gender identity on behavioral traits, labor market outcomes and political preferences, focusing on one key dimension, gender typicality, which we define below.

Gender identity is a broad psychological construct encompassing multiple dimensions beyond categorical identification (e.g., man, woman, nonbinary). According to Egan and Perry (2001), gender identity includes: gender typicality (feeling similar to one's gender group), gender contentedness (satisfaction with one's gender), felt pressure to conform (perceived social pressure to behave in gender-typical ways), and intergroup attitudes (relative preference for one's own gender group). In this paper, we focus specifically on gender typicality, defined as individuals' self-perceived alignment with socially defined masculine and feminine traits, independently of biological sex. We will use this term throughout to refer to our measure. Very recent work in economics, such as, Brenøe et al. (2022) Brenøe et al. (2024), employ a single-item measure of continuous gender identity, where participants rate themselves on a scale from "very masculine" to "very feminine." While their measure captures how individuals present themselves along a gender continuum, our approach captures the extent to which individuals are categorized as typical in separate masculine and feminine traits as they are socially defined.

In this paper, we create and validate a new two-dimensional gender typicality measure, with separate feminine and masculine components. We conduct two studies employing online experiments with a U.S. sample ($N = 2017$). In the first study, we construct an updated sex-role inventory of 90 attributes and ask participants to evaluate them in terms of desirability and social norms in the society at large and in the workplace. This allows us to classify all attributes as feminine, masculine, or neutral across different evaluative criteria and social contexts. In the second study, we measure behavioral traits that have previously shown binary sex differences and ask participants to rate *themselves* on all 90 attributes. Using machine learning, we determine the combination of attributes with the greatest predictive power for these traits, create a new measure of gender typicality, consisting of femininity and masculinity scores composed of 16 attributes in total. We then validate this measure against existing alternatives in the literature.

Our measure improves on existing approaches in explaining gender differences in behavioral traits such as confidence in one's absolute and relative performance, risk-taking, altruism, efficiency, equality, and competition preferences. Leveraging its two-dimensional structure, we show that gender differences in confidence, risk, and competition are primarily driven by masculine attributes, while traits such as altruism, equality, and efficiency are associated with feminine ones, thereby improving model fit and accounting for a portion of the variance traditionally attributed to the biological sex dummy. Additionally, we find that gender typicality predicts labor market outcomes. Masculinity is associated with higher income, higher probability of holding a managerial position, a greater likelihood of receiving performance-based pay, and a higher propensity to negotiate, whereas femininity does not exhibit systematic associations with these outcomes. Finally, we extend the analysis beyond economic behavior to political preferences. The interaction between femininity, masculinity, and biological sex helps explain ideological differences: gender-conforming individuals (masculine men and feminine women) are more likely to lean right, whereas gender-nonconforming individuals (masculine women and feminine men) are more likely to lean left. Taken together, these results shed new light on the heterogeneity among individuals often obscured by dichotomous gender classification and underscore the importance of studying gender identity alongside biological sex to better understand persistent gender gaps in economic behavior and labor market outcomes, as well as political preferences.

Our approach builds on a robust theoretical foundation provided by the identity utility framework of Akerlof and Kranton (2000). In this view, individuals' alignment with their social identity affects their utility and, consequently, their behavior. While their model focuses on discrete identity categories, it also implies that behavioral responses vary with the degree to which individuals internalize and conform to group-specific norms. In the context of gender, this suggests that individuals' self-perceived alignment with socially defined masculine and feminine traits may similarly influence economic decisions. This perspective becomes increasingly relevant in light of recent cultural and generational shifts. A recent U.S. national survey found that more than half of Generation Z and Millennial respondents consider traditional gender labels outdated and expect society to associate gender less with stereotypical traits in the coming decade Bigeye (2021). This is further supported by recent research showing that gender nonconformity and self-presentation outside of traditional roles can influence labor market behavior and perceptions (Coffman et al., 2017; Wilson and Meyer, 2021; Coffman et al., 2024). Taken together, these insights highlight the limitations of relying solely on binary sex categories and point to the need for multidimensional measures that capture the complexity of gender in ways relevant for both economic behavior and workplace outcomes.

Conceptualizing and measuring gender identity has always been challenging (Muehlenhard and Peterson, 2011; Hyde et al., 2019). Traditionally, economics has remained predominantly within what Nicholson (1994) calls *biological foundationalist paradigm*, treating gender as synonymous with biological sex and focusing on binary comparisons. In contrast, the psychological literature has long emphasized that gender identity is multidimensional, encompassing not only categorical identification but also social dimensions including typicality, contentedness, conformity, and intergroup attitudes (Egan and Perry, 2001). More recently, a growing body of work – including Meier-Pesti and Penz (2008), Kastlunger et al. (2010), Adamus (2018), Yavorsky and Buchmann (2019), Sent and van Staveren (2019), Brenøe et al. (2022), Burn and Martell (2022), Fornwagner et al. (2022), Banan et al. (2023), Sahi (2023), Brenøe et al. (2024), and Ayyar et al. (2024) – has advanced a social constructionist view of gender in economics and related disciplines, echoing the call of Butler (1990) to move beyond binary sex categories. Our contribution is to integrate the psychological concept of gender typicality, the degree to which individuals align with socially defined masculine and feminine traits, into economics, providing a validated measure and showing its predictive relevance.

A central contribution of our paper is to the literature on the development and validation of gender identity measures, building on the influential Bem Sex Role Inventory (BSRI) introduced in the 1970s (Bem, 1974). The BSRI conceptualizes gender identity as a continuous construct and remains one of the most widely used tools in the field.[1] It includes two components: "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." While influential, the BSRI relies on long item batteries. More recent tools include the Traditional Masculinity-Femininity Scale (Kachel et al., 2016), the two-dimensional scaling approach of Magliozzi et al. (2016) and the survey-based instrument of Fleming et al. (2017). These measures have broadened the scope of gender identity research, but they often face trade-offs: some are lengthy and thus difficult to implement in economic experiments, others are not validated against behavioral outcomes, and most remain within psychology or sociology. In the economics literature specifically, a continuous gender identity measure has recently been proposed by Brenøe et al. (2022) and Brenøe et al. (2024) and its predictive power has been tested on several outcomes. The authors propose a single-item 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". Henceforth we will refer to their measure as "CGI", which stands for "Continuous Gender Identity". Their findings suggest that CGI is statistically significantly related to preference and outcome measures often studied in gender economics, but lacks the flexibility to represent individuals who score high or low on both masculinity and femininity.

A second strand of literature links gender identity to preferences and economic outcomes. Early contributions such as Meier-Pesti and Penz (2008) and Kastlunger et al. (2010) rely on the Bem Sex Role Inventory to link masculinity and femininity to financial preferences, while Adamus (2018) adopts a similar approach in the domain of risk-taking. Others, such as Yavorsky and Buchmann (2019), Burn and Martell (2022), Banan et al. (2023), and Ayyar et al. (2024), construct gender typicality or conformity indices using detailed survey or behavioral data to explain educational or labor market outcomes. While valuable, existing contributions remain fragmented: they often focus on narrow outcomes or on specific populations such as children or students, and they frequently rely on complex data and on context-specific methodologies that are difficult to generalize. As a result, none develops a unified, validated measure of gender typicality that can be used systematically to predict a broad set of economic behaviors, labor market outcomes and political preferences. Our paper addresses this gap by introducing and validating such a measure, designed to capture the predictive power of gender typicality in a scalable and comparable way.

Finally, our findings 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

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

are otherwise neglected by the dichotomy of biological sex. To our knowledge, we are the first to develop and validate a two-dimensional gender typicality using machine learning and test its predictive power on a broad set of behavioral traits.

Our paper delivers three contributions. First, we develop and validate a short, two-dimensional measure of gender typicality that builds on the BSRI tradition while remaining concise yet scalable. Unlike unidimensional approaches, our measure allows individuals to score high or low on both masculinity and femininity, thereby capturing heterogeneity that binary sex variables or single-item measures cannot. Second, we show that this measure systematically predicts a broad set of economic behaviors and outcomes, including confidence, risk-taking, competition, altruism, and efficiency, as well as labor market outcomes such as income, managerial positions, and negotiation, as well as political preferences. This provides a unified and validated tool where previous contributions have remained fragmented across outcomes. Third, by using machine learning to identify an efficient set of predictive attributes, we demonstrate how modern methods can uncover hidden heterogeneity in gender identity that binary sex variables and unidimensional measures miss. In doing so, our framework not only advances the economics literature but also highlights promising avenues for management and organizational research, where understanding identity-related heterogeneity is key to designing inclusive and effective workplaces.

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. Sections 4 and 5 extend the analysis to labor market outcomes and political preferences. Finally, Section 6 discusses limitations and potential extensions of our work, and Section 7 concludes.

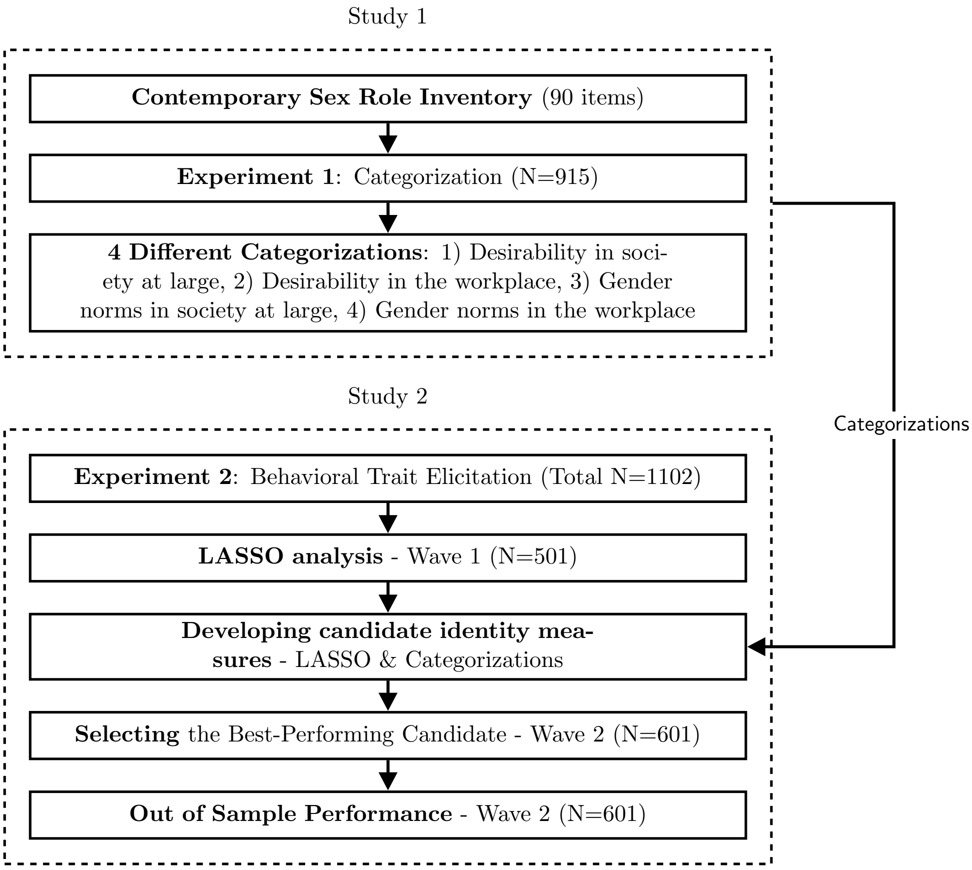
2 Study 1: A New Sex Role Inventory

The goal of our first study is to construct a new sex role inventory and determine the masculinity and femininity of each attribute in 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 I 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 in its short form, 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 I) (Bem, 1974, 1993). 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, Eber-

Figure I: Flowchart Depicting the Design of Study 1 and Study 2



hardt et al. (2023) showed that men were systematically more likely to be defined in terms of their abilities (e.g., "talented"), 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).

2.2 Experiment 1: Four Different Categorizations

After creating CSRI, the second step is to classify the attributes as feminine, masculine, or neutral. One way of doing this is the method used by Bem (1974), where masculinity, femininity and neutrality of the attributes are 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. If it was significantly more desirable for

Table I: 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.

women than men, 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. 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.[2] In total, we identified four possible categories to classify our 90 attributes as feminine and masculine: 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, 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

[2]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.

the same modified version of the Krupka and Weber (2013) norm elicitation in the workplace context.

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 Online Appendix 3.1. 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 1 in Online Appendix 1). 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.[3] 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.[4] 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.[5] 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 II).

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 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

[3]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.

[4]The comprehension question we used can be found in Figure 4 in Online Appendix 3.1.

[5]The attention check page was analogous to the pages where participants had to rate the attributes (see e.g. Figure 5 in Online Appendix 3.1.) 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.

Table II: 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.

from 1 to 7, 1 being *not at all desirable* and 7 *extremely desirable* (see e.g. Figure 5 in Online Appendix 3.1 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 Online Appendix 3.1). 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.[6]

2.3 Results

In this section, we provide a general overview of how attributes are classified based on our four categorizations displayed in Table II. 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).[7] 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 2 in Online Appendix 1.[8]

3 Study 2: Generating the 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: 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

[6]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.

[7]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.

[8]In the desirability treatments in Table 2 of Online Appendix 1, 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.

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. Participants 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.[9]

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.[10] 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.

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 Online Appendix 3.2). 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.

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

[9] 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.

[10] 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).

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 27 in Online Appendix 3.2). 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. 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. [11]

Competition

In Experiment 2, we captured competitiveness using a survey item recently developed by Fallucchi 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.[12]

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 3 and 4 in Online Appendix 2.1). 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.[13]

[11]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 $(30*1+10*3)/(0*1+40*3)=1/2$.

[12]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.

[13]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

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 our first goal was to find a set of attributes from the self-reported CSRI items listed in Table I 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 (See Online Appendix 2.2 for details of the LASSO procedure). 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. It is important to note however that the original BEMS gender measure included these two items. 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 for the first wave of Experiment 2.

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 associated with only one trait, while others predicted multiple traits. To focus on attributes with broader predictive relevance, we retained only those identified as important for at least two behavioral traits, refining our list to 41 attributes. To create our candidate identity measures, we classified these 41 attributes as feminine, masculine, or neutral based on the four categorizations from Study 1: 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—calculated as the arithmetic averages of their respective attributes, excluding neutral ones. Following the two-dimensional approach of Bem (1974), this method allows us to separately analyze the correlations between femininity and masculinity with behavioral traits, offering richer insights compared to a unidimensional scale. It also enables participants to rate themselves very high or very low on both femininity and masculinity, which would not be possible with a unidimensional scale. Notably, around 10% of participants scored very low (1–2) or very high (6–7) on both 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,

positive sign in front of the coefficient *Female* in the relative confidence measures in all regressions indicates that women are less confident than men.

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 5 in Online Appendix 2 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 6 in Online Appendix 2 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, in Study 1 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*.[14] 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", in only three out of ten (see Tables 7 to 20 in Online Appendix 2.4.1 for detailed regressions used in the model selection process).[15]

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.*

[14] 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 7 to 20 in Online Appendix 2.4.1

[15] 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.

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 II 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.

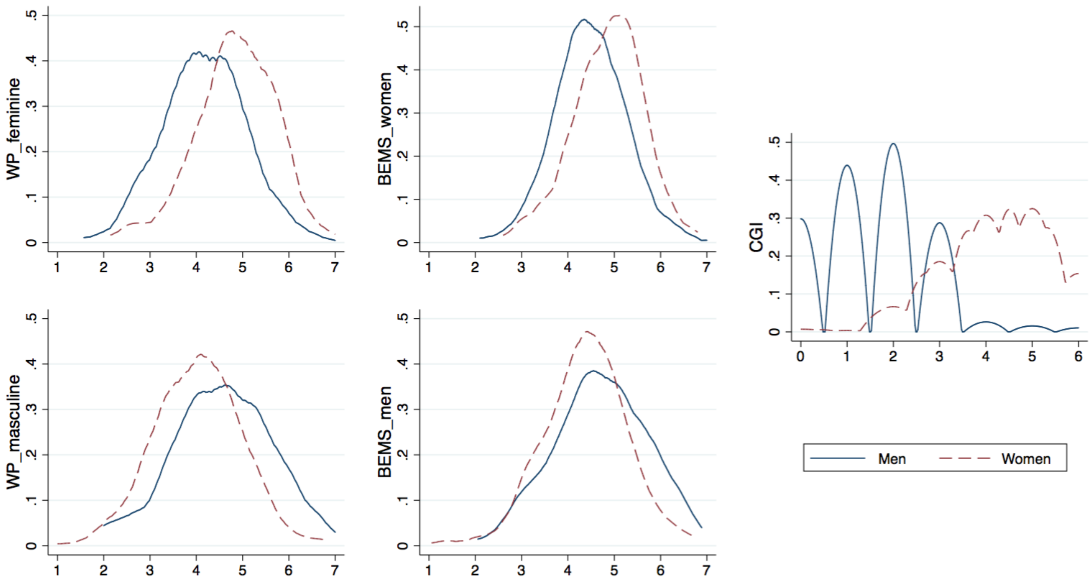
Moving forward to our model comparisons, we regress our new gender identity measure, the comparable BEMS classification (hereafter called "BEMS_men" and "BEMS_women")[16] and the continuous gender identity measure (CGI) on each of the ten behavioral traits: absolute confidence in math, relative confidence 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

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

Figure II: 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).

in explaining behavior. Table III presents all 40 regressions across the 10 behavioral traits, while Figure III visualizes the impact of each gender identity measure on the *Female* coefficient. Furthermore, results from Wald Tests, shown in Table 25 in Online Appendix 2.4.3, 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.[17]

Absolute and Relative Confidence. Considering absolute and relative confidence in a male task (Columns 1 and 2 of Table III), specifically solving math problems, the model

[17]Notice that if one is 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 fact, in this context, not including the gender identity measure is not a bias "per se", as one might be satisfied with a coarser measure that encompasses both biological and social traits.

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 III), 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 III), 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 III). 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 25 in Online Appendix 2.4.3). 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 III). We further extend our analysis to different contexts of risk preference – life, financial, occupational, and health-related – in Online Appendix 2.4.2, and we observe the same pattern in terms of adjusted R^2 and the significance of the coefficients across all these domains.

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 III). 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.

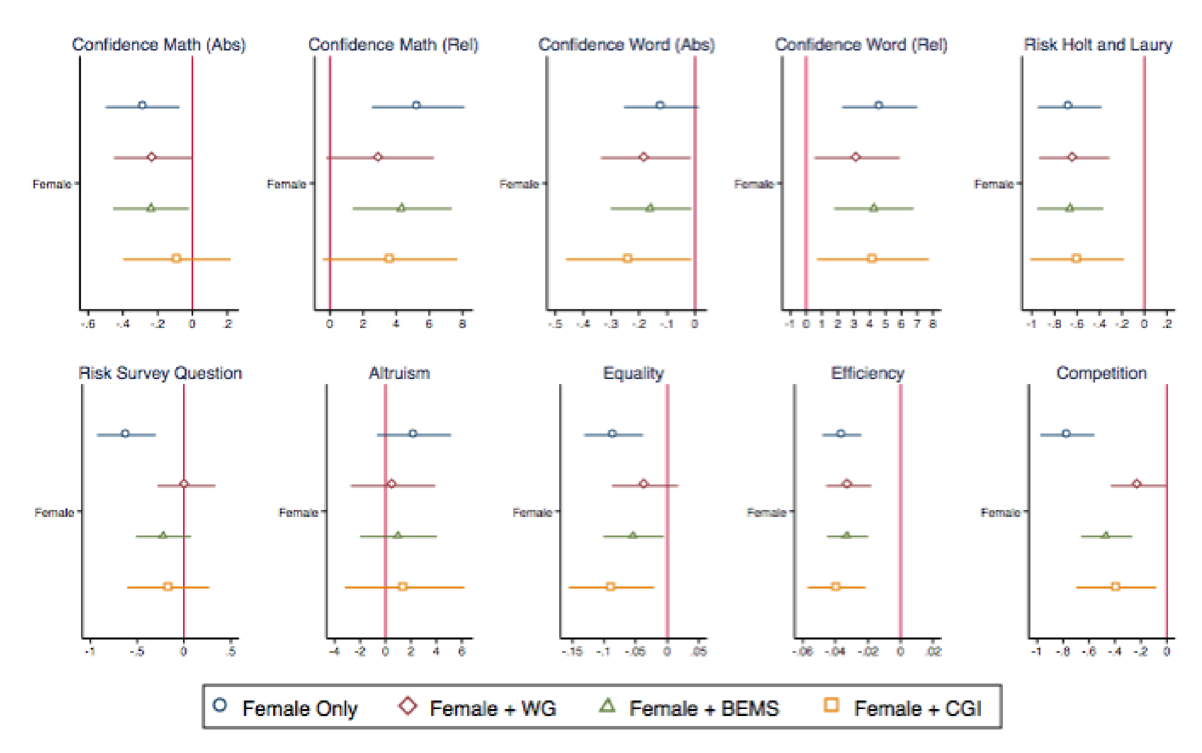
Altruism, Equality, Efficiency, and Competition. Across these four traits (Columns 7, 8, 9, and 10 of Table III), our model shows distinct strengths when compared to the other three models. In terms of absorbing the *Female* coefficient, our model performs better than the BEMS and CGI models, particularly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

Table III: Predictive Power of Gender Typicality 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure III: 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 III 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

significant ($p < 0.001$ in a Wald test) (see Table 25 in Online Appendix 2.4.3). 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 III).

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.[18]

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.*

To ensure the robustness of our results, we conducted multiple checks detailed in Section 2.4.5 of the online appendix. First, we excluded the biological sex variable *Female* to evaluate the predictive power of our gender identity measure relative to existing measures. Second, we ran regressions without demographic controls. Third, we performed iterative exclusion analyses to rule out the possibility of a single attribute driving the results. Fourth, we conducted a split-sample analysis, demonstrating consistent relationships between gender identity and behavioral outcomes across male and female subsamples. Finally, we compared our two-dimensional measure to a collapsed unidimensional version, showing that the latter has significantly lower predictive power. These analyses collectively confirm the robustness and value of our two-dimensional gender identity measure.

Our proposed measure of gender identity, concerning *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.This makes our approach not only more parsimonious but also more effective at capturing the relationship between gender identity and behavior, thereby contributing new insights to the ongoing discourse on gender differences in economic behavior.

4 Labor Market Outcomes

For approximately half of the sample (504 out of 1,082 participants), we collected additional information on labor market outcomes in a follow-up survey conducted after the main experiment. The follow-up survey was administered between October 31 and November 9, 2024, and

[18]The comparison between our model and CGI can also be found in Figure III and Table III.

all participants from the first and second waves of data collection were invited to participate. In the analyses of labor market outcomes and political preferences, we use the entire sample of returning respondents.

Labor market outcome measures capture key aspects of participants' employment situations, income, and professional characteristics. Below is a detailed description of the variables. *Employed* is a binary variable indicating whether the respondent was employed at the time of the follow-up survey. *EmploymentHours* records the number of hours usually worked per week during the past 12 months, coded into seven categories ranging from 0 hours to more than 50 hours per week. *Income* measures the respondent's yearly total gross personal income, including all sources of earned and unearned income before taxes (e.g., wages, bonuses, self-employment income, dividends, and interest). *Manager* is a binary indicator of whether the respondent holds managerial responsibilities, such as supervising staff or participating in hiring and firing decisions. *PerformanceBonus* captures whether the respondent's income includes performance-related pay or bonuses, with four response categories ranging from no bonus to more than two-thirds of total income, plus a "don't know" option. In addition, *Negotiation* measures the extent to which respondents report engaging in negotiation within their employment relationships, based on a five-point Likert scale ranging from "never or almost never true" to "always or almost always true".

We further include several constructed variables based on respondents' backgrounds and occupations. *Women%* represents the share of women employed in the respondent's industry, matched from external labor market data by sector. *NaturalScienceStudy* is a binary indicator equal to one for participants whose field of study falls within the natural sciences, economics, medicine, or STEM disciplines, and zero otherwise. *STEM* is a narrower indicator, taking the value one for respondents with a background in economics, STEM fields, or medicine. *HighEducation* equals one for respondents holding a graduate or doctoral degree and zero otherwise.

Table IV examines whether attrition between the baseline and the follow-up survey is systematically related to baseline individual characteristics. The table reports average marginal effects from a logit model in which the dependent variable is a binary indicator for responding to the second survey. We find no statistically significant differences in attrition by gender, education, ethnicity, or femininity scores. By contrast, age and masculinity are significantly associated with the probability of returning: older individuals and those with lower masculinity scores are more likely to participate in the follow-up.

Because participation in the follow-up survey is not random, we address selective attrition using an inverse-probability weighting approach. Specifically, we estimate a logit model of follow-up participation as a function of baseline covariates that predict attrition, including age and standardized measures of masculinity and femininity, as well as gender, education, and ethnicity. Based on this model, we construct stabilized inverse-probability weights equal to the overall response rate divided by each individual's predicted probability of returning. Applying these weights to the sample of follow-up respondents re-weights the observed data to recover the baseline distribution of observable characteristics. All subsequent analyses of labor market outcomes are therefore estimated on the subsample of follow-up respondents using these weights,

Table IV: Determinants of Returning in the Follow-up Survey

	(1)
	Return
Female	-0.034
	(0.033)
Age	0.011***
	(0.001)
Education	-0.001
	(0.013)
Ethnicity	-0.009
	(0.015)
Femininity	-0.001
	(0.017)
Masculinity	-0.036*
	(0.015)
Observations	1102

Standard errors in parentheses
# $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

so that results are representative of the original baseline population under the assumption that, conditional on observed characteristics, attrition is as good as random.

Comparing Panels A and B of Table V indicates that several labor market differences initially attributed to biological sex are attenuated once gender-typical traits are taken into account, particularly masculinity. In Panel A of Table V, being female is associated with a lower probability of employment, fewer hours worked, a lower likelihood of holding managerial positions, reduced reception of performance bonuses, and lower engagement in negotiations. Women are also significantly more likely to work in female-dominated occupations and substantially less likely to study natural sciences or STEM fields, while no statistically significant differences by sex are observed for income or overall educational attainment.

When masculinity and femininity are added to the specification, as shown in Panel B of Table V, the coefficient on being female becomes statistically insignificant across nearly all outcomes, indicating that biological sex per se explains relatively little residual variation once gender-typical traits are accounted for. By contrast, masculinity emerges as a strong and robust predictor of labor market outcomes: higher masculinity scores are associated with a greater likelihood of employment, higher income, increased probability of holding managerial positions, receiving bonuses, and engaging in negotiations, as well as sorting into less female-dominated occupations. Femininity exhibits no systematic association with labor market outcomes once masculinity is included. Allowing the returns to gender typicality to vary by biological sex reveals limited heterogeneity: interaction terms between femininity and sex are uniformly small and statistically insignificant, while the interaction between masculinity and being female is negative and statistically significant only for bonus receipt, suggesting that although masculinity is generally rewarded for both men and women, women experience weaker marginal returns to masculinity in contexts involving discretionary, performance-based compensation. Overall, these findings suggest that a substantial portion of the labor market disparities observed by biological sex operates through gender-typical traits, particularly masculinity, rather than sex itself.

Table V: Labor Market Outcomes: Biological Sex and Gender Typicality

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Employed	Hours	Income	Manager	Bonus	Negot	Women%	NatSci	STEM	HighEd
Panel A: Predictive Power of Biological Sex										
Female	-0.317#	-0.344***	-0.062	-0.141*	-0.318*	6.996***	-0.236***	-0.267***	-0.012	-0.002
	(0.163)	(0.104)	(0.043)	(0.068)	(0.133)	(1.582)	(0.052)	(0.051)	(0.035)	(0.027)
R-squared	0.145	0.525	0.155	0.065	0.081	0.167	0.124	0.140	0.089	0.635
Observations	504	504	504	481	489	489	384	384	504	504
Panel B: Gender Typicality and Labor Market Outcomes										
Female	0.216	-0.568	0.395	-0.024	0.719#	-0.185	-4.104	0.070	0.047	0.191
	(0.250)	(0.616)	(0.584)	(0.268)	(0.374)	(0.840)	(10.228)	(0.322)	(0.318)	(0.204)
Femininity	-0.018	-0.097	-0.055	0.013	-0.010	-0.034	0.386	0.035	0.033	0.012
	(0.029)	(0.072)	(0.090)	(0.037)	(0.047)	(0.102)	(1.526)	(0.045)	(0.045)	(0.028)
Female × Femininity	0.006	0.146	-0.003	-0.036	-0.006	-0.014	0.098	-0.092	-0.093	-0.033
	(0.045)	(0.110)	(0.120)	(0.055)	(0.069)	(0.158)	(1.997)	(0.062)	(0.061)	(0.040)
Masculinity	0.075**	0.182*	0.328***	0.066*	0.201***	0.322**	-3.045*	-0.000	0.002	0.029
	(0.029)	(0.072)	(0.079)	(0.033)	(0.059)	(0.105)	(1.192)	(0.043)	(0.043)	(0.023)
Female × Masculinity	-0.059	0.011	-0.118	0.042	-0.172*	0.045	2.074	0.030	0.030	-0.008
	(0.040)	(0.101)	(0.107)	(0.043)	(0.074)	(0.140)	(1.721)	(0.055)	(0.055)	(0.036)
R-squared	0.185	0.657	0.555	0.186	0.107	0.137	0.184	0.130	0.147	0.096
Observations	504	504	504	504	481	489	489	384	384	504

Notes: Robust standard errors are reported in parentheses. Regressions are estimated on respondents to the follow-up survey and weighted using inverse-probability weights to correct for selective attrition. All specifications control for age and ethnicity. Columns (2)–(10) additionally control for education level and employment category, except where the latter defines the outcome. In Panel B, columns (2)–(9) additionally control for education level and employment category; column (1) (Employed) excludes employment-category controls, and column (10) (HighEd) excludes education level as a control.
# $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5 Political Preferences

We next extend the same empirical framework to political preferences using the full sample of returning respondents, asking whether the role of gender typical traits observed in labor market outcomes also extends to ideological attitudes. Political affiliation is measured on a seven-point scale ranging from strong Republican (0) to strong Democrat (6), with higher values indicating more liberal preferences. As shown in Table VI Column 1, in a specification including only biological sex and standard controls, women display, on average, significantly more liberal political preferences than men.

Once masculinity and femininity, along with their interactions with biological sex, are introduced, as reported in Table VI Column 2, this sex difference is substantially attenuated and no longer statistically significant, indicating that biological sex per se explains little residual variation in political affiliation. Instead, political preferences align closely with gender conformity. Among men, greater masculinity is associated with more conservative political views, whereas greater femininity is associated with more liberal preferences. Among women, the pattern reverses, with higher femininity associated with more conservative political affiliation and gender atypical women who score higher on masculinity more likely to report liberal preferences. Consistent with this interpretation, the interaction between femininity and being female is large and negative, while the interaction between masculinity and being female substantially offsets the baseline association between masculinity and conservatism. Taken together, these results suggest that ideological differences commonly attributed to biological sex are to a substantial extent associated with conformity to gender norms. Individuals whose traits align with traditional gender expectations, namely masculine men and feminine women, tend to hold more conservative political views, whereas gender nonconforming individuals, namely feminine men and masculine women, are systematically more liberal.

Table VI: Predictive Power of Biological Sex and Gender Typicality on Political Affiliation

	(1) Political Affiliation	(2) Political Affiliation
Female	0.322**	0.176
	(0.104)	(0.117)
Femininity		0.277***
		(0.079)
Female × Femininity		-0.397***
		(0.116)
Masculinity		-0.270***
		(0.078)
Female × Masculinity		0.198#
		(0.115)
Observations	1,082	1,082
R-squared	0.034	0.053

Notes: Robust standard errors in parentheses. All specifications control for age, education, employment status, and ethnicity.
# $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6 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 relatively 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.

7 Conclusion

This paper develops a gender identity measure with separate feminine and masculine components, thereby avoiding multicollinearity issues inherent in unidimensional measures. Each component is constructed from distinct attributes, resulting in a gender identity measure that minimizes the influence of demand effects. The proposed measure improves model fit in predicting behavioral traits and reduces the risk of misinterpreting the biological sex dummy when used in isolation. Moreover, compared to the Bem Sex Role Inventory, our measure substantially reduces the number of attributes, resulting in a more streamlined and practical tool for survey-based research.

Our findings provide new insights into previously documented gender disparities in economic decision-making. In particular, our findings demonstrate a correlation between confidence, risk-taking, competitiveness and labor market outcomes are strongly associated with masculine traits, whereas altruism, efficiency, and equality concerns are more closely related to feminine traits.

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.

Conversely, 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 can help individuals with low masculine traits to build competitiveness, confidence, and risk-taking abilities, while those with low feminine traits altruism and concern for equality, contributing to a more balanced and diverse economic decision-making process.

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.

Ayyar, S., U. Bolt, E. French, and C. O'Dea (2024). Imagine your life at 25: Gender conformity and later-life outcomes. Technical report, National Bureau of Economic Research.

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.

Banan, A. R., T. Santavirta, and M. Sarzosa (2023). Childhood gender nonconformity and gender gaps in life outcomes. *Unpublished Manuscript*.

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.

Bigeye (2021). Gender: Beyond the binary – a national study on gender identity and expression. https://www.them.us/story/gen-z-study-traditional-gender-norms-outdated. Accessed: 2025-05-30.

Brenøe, A. A., Z. Eyibak, L. Heursen, E. Ranehill, and R. A. Weber (2024). Gender identity and economic decision making. Technical report, Working Paper.

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.

Burn, I. and M. E. Martell (2022). Gender typicality and sexual minority labour market differentials. *British Journal of Industrial Relations 60*(4), 784–814.

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.

Caliendo, M., F. M. Fossen, A. Kritikos, and M. Wetter (2015). The gender gap in entrepreneurship: Not just a matter of personality. *CESifo Economic Studies 61*(1), 202–238.

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.

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.

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

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.

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.

Egan, S. K. and D. G. Perry (2001). Gender identity: a multidimensional analysis with implications for psychosocial adjustment. *Developmental psychology 37*(4), 451.

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.

Fleming, P. J., K. M. Harris, and C. T. Halpern (2017). Description and evaluation of a measurement technique for assessment of performing gender. *Sex Roles 76*, 731–746.

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*.

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.

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.

Krupka, E. L., R. Weber, R. T. Croson, 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.

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.

Powell, M. and D. Ansic (1997). Gender differences in risk behaviour in financial decision-making: An experimental analysis. *Journal of economic psychology 18*(6), 605–628.

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.

Yavorsky, J. E. and C. Buchmann (2019). Gender typicality and academic achievement among american high school students. *Sociological Science 6*, 661–683.