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

Stefano Piasenti†, Müge Süer‡

March 30, 2026

Abstract

Are gender gaps in economic behavior and labor market outcomes associated with biological sex, or with gender-typical traits? Using an online U.S. sample and machine learning, we develop and validate a gender typicality measure capturing masculinity and femininity as separate dimensions. We show that confidence, competitiveness, and risk-taking are associated with masculinity, whereas altruism and equality and efficiency concerns are associated with femininity. Gender typicality also predicts labor market outcomes: masculinity is associated with higher income, managerial positions, performance-based pay, and willingness to negotiate. By leveraging latent heterogeneity, our framework offers a new empirical lens for analyzing gender differences.

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

*We thank Billur Aksoy, Dirk Engelmann, Tilman Fries, Lavinia Kinne, Dorothea Kübler, Nicola Lacetera, Huyen Nguyen, Regine Oexl, Davide Pace, 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 Behavioral 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 Milan (stefano.piasenti@unimi.it)
‡Halle Institute for Economic Research (IWH) (muege.sueer@iwh-halle.de)

1 Introduction

Gender gaps in economic behavior are widely documented, with significant implications for labor market inequality and social stereotypes (Markowsky and Beblo, 2022; Lozano et al., 2022). Yet empirical findings are often inconsistent and highly context dependent, limiting their usefulness for theory and policy (Croson and Gneezy, 2009; Sent and van Staveren, 2019). While differences between men and women are frequently interpreted as reflecting inherent distinctions, such interpretations typically rely on biological sex as the empirical proxy for gender.

A central but often implicit assumption in this literature is that biological sex sufficiently captures gender. In practice, sex is treated as synonymous with gender in empirical analyses. However, gender encompasses socially constructed norms and traits associated with masculinity and femininity, which may vary substantially within biological sex categories (West and Zimmerman, 1987). If biological sex does not adequately capture these socially defined dimensions, observed gender gaps may reflect variation in conformity to gendered norms rather than inherent differences between men and women.

In this paper, we examine whether observed differences in economic behavior and labor market outcomes are more closely associated with biological sex or with gender identity traits. Gender identity is a multidimensional construct that extends beyond categorical identification and includes dimensions such as typicality, contentedness, conformity pressures, and intergroup attitudes (Egan and Perry, 2001). We focus on the gender typicality dimension, defined as the extent to which individuals align with socially defined masculine and feminine attributes.

Recent work in economics has begun to incorporate continuous measures of gender identity (Brenøe et al., 2022, 2024). They propose a parsimonious single-item measure that captures self-presentation along a single continuum from masculine to feminine. While this approach captures self-presentation along a single continuum, our approach conceptualizes gender typicality as alignment with socially defined masculine and feminine traits along distinct dimensions. This allows us to measure the degree of alignment with each set of traits separately and to account for individuals who exhibit both masculine and feminine characteristics strongly (or weakly), allowing us to uncover within-gender heterogeneity that binary sex categories obscure.

We develop and validate a new two-dimensional measure of gender typicality, with separate masculinity and femininity components. Using two online experimental studies with a U.S. sample ($N = 2{,}017$), we first construct an updated inventory of 90 attributes commonly associated with gender roles, building on earlier work such as the Bem Sex Role Inventory (Bem, 1974). Participants of the first study evaluate attributes along two criteria, perceived desirability and perceived social norms, separately for society at large and for the workplace. This design allows us to identify which traits are socially coded as masculine or feminine across contexts, rather than imposing classifications a priori.

In a second study, conducted across two waves, we measure a comprehensive set of behavioral traits with well-documented gender gaps, including confidence, risk-taking, competitiveness, altruism, and concerns for equality and efficiency. Participants of this study also performed a self-assessment using the attribute inventory. By applying machine learning to the first wave

(the *training sample*), we identify the specific attributes that most strongly predict these economically relevant behaviors. We then validate the resulting gender typicality measure out of sample in a second wave as the *test sample*, mitigating concerns about overfitting.

Our findings reveal a consistent pattern. Masculinity, rather than biological sex per se, is strongly associated with confidence, competitiveness, and risk-taking. We extend the analysis to labor market outcomes. We show that masculinity is also positively associated with higher income, managerial positions, performance-based pay, and willingness to negotiate. Femininity, in contrast, is associated with altruism and stronger concerns for equality and efficiency. Importantly, once masculinity and femininity are included in the analysis, the explanatory role of biological sex declines substantially in several domains. These results suggest that some documented gender gaps may reflect variation in conformity to socially defined traits along with inherent differences between men and women.

Our framework builds on the identity utility model of Akerlof and Kranton (2000), which emphasizes that alignment with social identities affects utility and behavior. Whereas much empirical work proxies gender using binary sex categories, our approach measures the degree and dimensionality of conformity to socially defined norms, a construct we refer to as "gender typicality." Behavioral differences vary not only across sex categories but also with the intensity of alignment with masculine and feminine traits. This perspective provides a structured way to interpret heterogeneous findings in the gender literature.

This perspective is particularly relevant in light of recent cultural and generational shifts. Survey evidence from the United States indicates that a majority of Generation Z and Millennial respondents view traditional gender labels as outdated and expect weaker associations between gender and stereotypical traits in the future (Bigeye, 2021). Consistent with this trend, recent research shows that gender nonconformity and self-presentation outside traditional roles affect labor market behavior and perceptions (Coffman et al., 2017; Wilson and Meyer, 2021; Coffman et al., 2024). Together, these developments point to increasing heterogeneity in how individuals relate to gender norms, underscoring the importance of frameworks that move beyond binary sex categories.

This paper makes three main contributions. First, we provide a validated, parsimonious, and scalable two-dimensional measure of gender typicality for use in economic research. Second, we demonstrate that this measure predicts a wide range of economic behaviors and labor market outcomes beyond biological sex, thereby offering a unified explanation for findings that have previously appeared fragmented or context dependent. Third, by leveraging machine learning to uncover latent structure in gender-related attributes, we show how modern empirical tools can reveal economically meaningful heterogeneity that coarse binary indicators overlook.

More broadly, our results suggest that interpreting gender gaps solely through biological sex may obscure important factors of economic behavior. Distinguishing between sex and socially constructed gender traits has implications for how economists conceptualize inequality, how organizations design interventions, and how policy targets disparities. Rather than asking only whether men and women differ, our framework encourages examining which identity-conforming traits shape observed gaps.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature. Sections 3 and 4 describe the two main studies. Sections 5 extend the analysis to labor market outcomes. Section 6 discusses limitations and potential extensions, and Section 7 concludes.

2 Literature Review

Conceptualizing and measuring gender identity has long posed challenges across the social sciences (Muehlenhard and Peterson, 2011; Hyde et al., 2019). Traditional economic research has largely operated within what Nicholson (1994) terms a biological foundationalist paradigm, treating gender as synonymous with biological sex and focusing primarily on binary comparisons between men and women. In contrast, research in psychology and sociology emphasizes that gender identity is multidimensional and socially embedded, encompassing constructs such as gender typicality, gender contentedness, conformity pressures, and intergroup attitudes (Egan and Perry, 2001).

A central strand of the literature focuses on the development of instruments to measure gender identity. A foundational contribution is the Bem Sex Role Inventory (BSRI; Bem, 1974), which conceptualizes gender identity as a continuous construct with distinct masculine and feminine dimensions. While highly influential, the BSRI relies on extensive item batteries, limiting its feasibility in many economic applications. Subsequent work has proposed alternative measures aimed at reducing respondent burden or refining dimensionality, including the Traditional Masculinity–Femininity Scale (Kachel et al., 2016), two-dimensional scaling approaches (Magliozzi et al., 2016), and survey-based instruments developed in psychology and sociology (Fleming et al., 2017). More recently, Gürel et al. (2025) introduced the Multidimensional Gender and Sexuality Inventory (MGSI), which integrates several established scales to capture a broad spectrum of gender identities, roles, and sexual orientation. While these instruments are designed to capture the complexity of gender identity and broaden the measurement toolkit available to researchers, none is specifically developed to identify the traits that predict economic decision-making.

A second strand links gender identity to economic preferences and outcomes. Early contributions demonstrate that gender-role orientation and related gendered traits, rather than biological sex per se, are predictive of financial decision-making (Meier-Pesti and Penz, 2008; Lemaster and Strough, 2014), risk-taking (Adamus, 2018), competitiveness (Kamas and Preston, 2012) and tax compliance (Kastlunger et al., 2010). Subsequent work examines the robustness and context dependence of these relationships in experimental settings (Fornwagner et al., 2022). Together, this literature establishes that gender identity captures economically meaningful variation beyond binary sex classifications, albeit primarily within narrowly defined behavioral domains. More recent work emphasizes parsimony in measurement. Brenøe et al. (2022) and Brenøe et al. (2024) introduce a single-item continuous measure of gender identity and document its predictive relevance across a wide range of economic preferences and outcomes.

A related literature constructs indices of gender typicality or conformity using detailed survey or behavioral data to study educational and labor market outcomes (Yavorsky and Buchmann, 2019; Burn and Martell, 2022; Sahi, 2023; Ayyar et al., 2024; Banan et al., 2025). While

valuable, these contributions remain fragmented. They typically focus on narrow outcomes or specific populations, such as children or students, and often rely on complex, context-specific data and methodologies that are difficult to generalize. As a result, none of these studies develops a unified and validated measure of gender typicality that can be systematically applied to predict a broad set of economic behaviors and labor market outcomes.

Recent advances in economics underscore the potential of machine learning methods to uncover latent structure and heterogeneity in complex data (Athey and Imbens, 2019). Rather than imposing coarse categories ex ante, these approaches allow the data to reveal predictive patterns that would otherwise remain obscured. Our approach builds on this insight by using machine learning to identify the subset of gender-related attributes with the greatest predictive power on economic behavior. In doing so, we uncover latent heterogeneity in gender typicality that is overlooked by binary sex indicators and unidimensional identity measures. To our knowledge, this is the first study to develop and validate a two-dimensional measure of gender typicality using machine learning and to demonstrate its predictive relevance in a broad set of behavioral traits and labor market outcomes.

3 Study 1: A New Sex Role Inventory

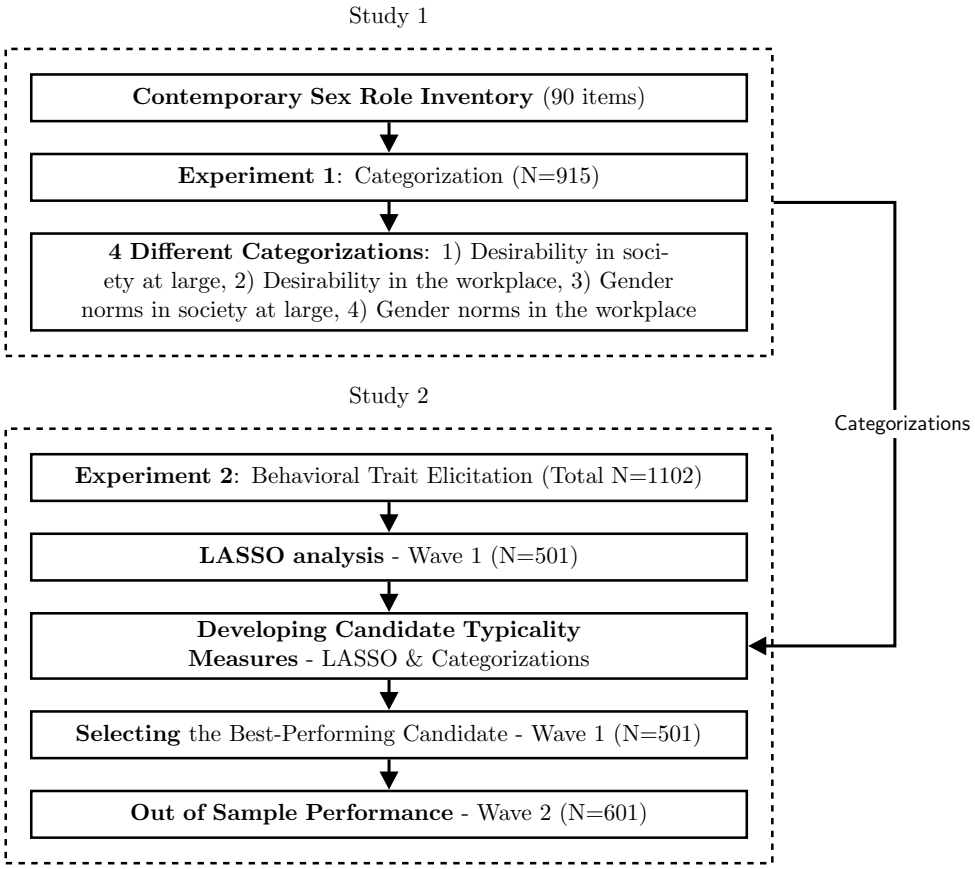
The goal of our first study is to construct a new sex role inventory and determine the masculinity and femininity of each attribute in four different categories: 1) desirability in society at large, 2) desirability in the workplace, 3) gender norm in society at large, and finally 4) gender norm in the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

3.1 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, Eberhardt 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

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



role inventory, we see considerable promise in combining the 60-item BSRI with the work-related attributes by Eberhardt et al. (2023).

3.2 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 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 a 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

Table I: Contemporary Sex-Role Inventory (CSRI)

Bem 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 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.

(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.[1]

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

3.2.1 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

[1]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.

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.[2] 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.[3] 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.[4] 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.

3.2.2 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 in the workplace (see Table II).

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 the entire population. *N* is the sample size and *Incentivization* is whether the question in the corresponding treatment is incentivized or not.

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.

[2]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.

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

[4]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.

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 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 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 believed was 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.[5]

[5]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.

3.3 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).[6] 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.[7]

4 Study 2: Generating the Gender Typicality Measure

We designed a second online experiment, Experiment 2, to elicit behavioral traits and collect participants' self-ratings on the CSRI attributes, which together form the basis for constructing the gender typicality measure. Experiment 2 was run in two waves. We followed four main steps, which were inspired by the Falk et al. (2023) 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 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 typicality 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

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.

[6] 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.

[7] 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.

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 typicality 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 eleven candidate gender typicality measures.

3. Selecting the best-performing gender identity measure out of all candidates: We compared the candidate measures within the *training sample* ($N = 501$) using a selection process. We first identified the top-performing measures that most effectively absorbed the biological sex dummy across the widest range of behavioral traits. From this group, we selected the one utilizing the fewest attributes to ensure model efficiency and simplicity.

4. Comparison to existing measures: Finally, using the fresh *test sample* ($N = 601$), we compared the predictive power of our new gender typicality measure on the behavioral traits against two existing measures from the literature, BSRI by (Bem, 1974) and CGI by (Brenøe et al., 2022).[8]

4.1 Experiment 2: Capturing Behavioral Traits

4.1.1 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]

[8] We benchmark our measure against BSRI and CGI for two reasons. First, BSRI is the direct methodological predecessor of our measure, making it a natural point of comparison. Second, CGI is the only other continuous gender identity measure that has been adopted in economics research. While other scales exist in the psychology and sociology literatures (e.g., Kachel et al., 2016; Magliozzi et al., 2016; Fleming et al., 2017), none have been used in economic applications, making them less relevant benchmarks for our purposes.

[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.

4.1.2 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 were further elicited as exit questionnaire survey items and they were not incentivized following Dohmen et al. (2011) and Fallucchi et al. (2020), respectively.

Absolute and Relative Confidence

To elicit absolute and relative confidence, we followed Dreber et al. (2014) and elicited them in stereotypically male and female domains. To do so, we used the math and verbal tasks mentioned above. The math task was a 6-item addition and multiplication of $1s$ and $0s$. Participants were asked to complete ten problems in 1 minute. The verbal task was a 7x8 word search matrix with ten hidden 4-letter words. Participants also had 1 minute to find the words. Both tasks were presented in a randomized order. When the math (word) task was randomly selected for payment, participants earned 10 points for each question (word) they answered (identified) correctly. Following these two real-effort tasks, each subject was asked to report their performance-related confidence on both tasks. Confidence was elicited in two ways: first, by asking how many problems they solved correctly (words they identified correctly), second, by asking where they thought their performance lay within a group of 100 randomly selected subjects. Both elicitations were incentivized. For the former, if the answer was correct, they earned an additional bonus payment of ten points. 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

[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).

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 the 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 efficiency variable was constructed, a participant who always maximized the total payoff of the pair would receive a value of 1, while one who never did would receive a value of 0. For the equality variable, a value of 0 corresponds to perfectly equalizing payoffs between oneself and the other person, while positive values indicate keeping more than equality would require and negative values indicate giving more.[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.

4.1.3 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

[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 $30 \times 1 - 10 \times 3 = 0$, indicating perfect equality. The efficiency for that decision would be $(30 \times 1 + 10 \times 3)/(0 \times 1 + 40 \times 3) = 1/2$, where the denominator represents the maximum total payoff achievable by the pair, obtained when the participant passes the entire endowment.

[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.

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]

4.2 Development of the Candidate Typicality Measures

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

4.2.1 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 BSRI 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.

4.2.2 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 typicality 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 typicality 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.

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

4.2.3 Candidate Typicality Measures

The first candidate measure, which we named "GSfeminine2" and "GSmasculine2" encompassed all the attributes selected by LASSO as important predictors of at least two behavioral traits, and classified as feminine and masculine based on Study 1. 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 typicality 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 typicality 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 typicality measures based on the frequency of appearance in LASSO analyses, each comprising 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 typicality measures, "KWWPfeminine2" and "KWWPmasculine2", "KWWPfeminine3" and "KWWPmasculine3", and "KWWPfeminine4" and "KWWPmasculine4".

These steps resulted in eleven candidate measures of gender typicality. Table 6 in Online Appendix 2 provides a summary of their names and the number of attributes that belong to each measure.

4.3 New Gender Typicality Measure

4.3.1 Selection of the Preferred Measure

Following the creation of our candidate typicality measures, the next step was to identify the preferred specification. We utilized the *training sample* to evaluate the candidates based on their predictive power and structural complexity.

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 typicality than those in society at large, and thus, the measure of gender typicality created by workplace categorizations would be a better tool for predicting gender differences in behavior.

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

To identify our preferred typicality measure among the eleven candidates, we employ a selection mechanism based on their performance within the *training sample.*[14] We evaluate the candidates across four criteria: (i) the extent to which the measure absorbs the significance and magnitude of the biological sex dummy, *Female*, (ii) the statistical significance of the feminine and masculine typicality components, (iii) the overall model fit (adjusted R^2), and (iv) the model simplicity (see Tables 7 to 16 in Online Appendix 2.3.1 for detailed regressions used in the model selection process).

Our analysis reveals a top-performing tier of measures, specifically GS2, KWGS2, KWWP2, and WP2, which demonstrate comparable and superior predictive power across the first three selection criteria. Within this cluster, no single measure strictly dominates across all behavioral domains. Following the model simplicity, we select from this top-tier the measure with the fewest underlying attributes. As detailed in Table 6, WP-2 (comprising 16 attributes) is substantially more parsimonious than its closest competitors, GS2 (32 attributes), KWGS2 (25 attributes) or KWWP2 (22 attributes), while delivering equivalent levels of significance and explanatory power.

Consequently, we select the workplace desirability measure with the LASSO 2+ threshold, identifying attributes that appeared in the prediction of at least two behavioral traits, as our primary typicality measure (previously "WPfeminine2" and "WPmasculine2" hereafter *WP_feminine* and *WP_masculine*). To ensure this selection is not an artifact of the training data, we then validate its performance out-of-sample on the fresh *test sample.* We find that WP2 remains the best-performing alternative in the test sample, confirming the robustness of our selection. *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.*

[14] For each behavioral trait, we have eleven models, each including one of the eleven measures and the biological sex dummy, *Female*, plus a model including only the biological sex dummy. Since we measured ten main traits, we have a total of 12*10 = 120 models.

Result 1. *Based on our selection process in the training sample, we find that the workplace desirability measure offers the best balance of predictive power and model simplicity.*

4.3.2 Internal Validity and Semantic Robustness

To ensure the internal validity and semantic robustness of the new 16-attribute gender typicality measure, we conducted two complementary analyses.

First, we assessed the internal consistency of each scale using Cronbach's α. Both the femininity scale, *WP_feminine* ($\alpha = 0.77$), and the masculinity scale, *WP_masculine* ($\alpha = 0.82$), exceed the standard psychometric threshold for research ($\alpha > 0.70$), indicating high internal consistency. The low correlation between the two aggregate scores ($r = 0.145$, $p < 0.01$) further confirms that the two scales capture genuinely distinct constructs rather than opposite poles of a single bipolar axis.

Second, to verify that the predictive power of our measure does not simply reflect linguistic overlap between the attributes and the behavioral outcomes they predict, we performed a formal semantic robustness analysis. Following Budanitsky and Hirst (2006), we use the WordNet lexical database (Fellbaum, 1998) to evaluate the taxonomic distance between the 16 attributes and the 8 behavioral constructs. Specifically, for each attribute-behavior pair, we calculate the shortest path connecting them in the WordNet hierarchy, where words with closely related meanings are linked by shorter paths. The resulting mean path similarity score of 0.069 (where 1.0 indicates perfect synonymy) confirms that the attributes and behaviors occupy distinct branches of the English language hierarchy and that the predictive power of our typicality measure cannot be attributed to linguistic overlap with the outcomes it predicts.

4.4 Out-of-Sample Performance of the New Gender Typicality Measure

Having identified our preferred specification using the training data, we now evaluate its predictive power on the *test sample*. This out-of-sample validation is critical to ensure that our results are not driven by overfitting or the idiosyncratic features of the training sample. Our ultimate goal is to determine whether our selected gender typicality measure, *WP_feminine* and *WP_masculine*, can better predict behavioral traits and account for gender differences beyond biological sex, compared to the CGI measure by Brenøe et al. (2022) and the BSRI measure of masculinity and femininity by Bem (1974).

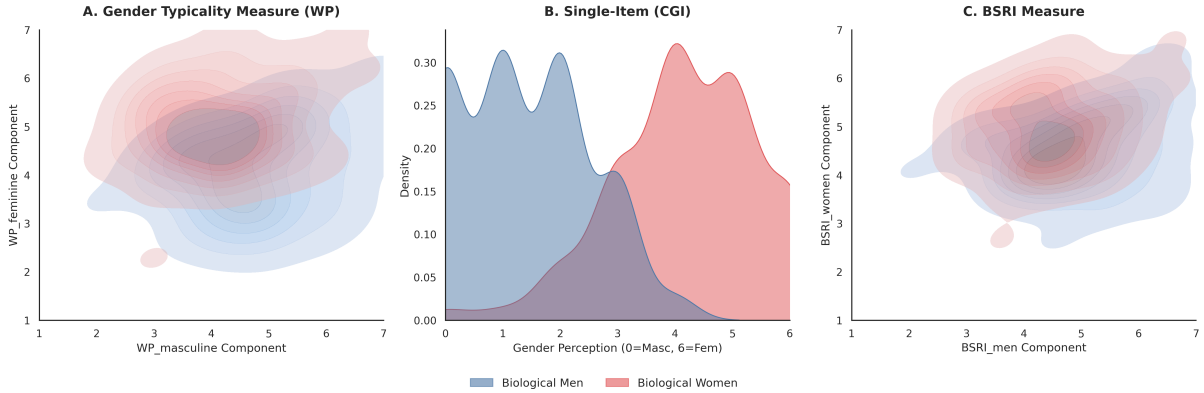
Hypothesis 2. *The new gender typicality measure outperforms binary sex, the single-item masculinity/femininity measure, CGI and BSRI measure in predicting confidence, altruism, risk, competition, efficiency, and equality preferences.*

We first illustrate the heterogeneity captured by the three measures. Using kernel density estimation, we visualize the distributions of masculinity and femininity separately for biological men and women. (Figure II). Our proposed measure (WP) and the BSRI measure, both of which treat masculine and feminine components as independent dimensions, reveal substantial within-sex variation. These two-dimensional clouds demonstrate that gender-typical traits are

not perfectly polarized by biological sex. Instead of two distinct, isolated clusters, the two-dimensional clouds show a substantial area of common support where the distributions for men and women overlap significantly.

This overlap demonstrates that a high score in one dimension does not necessitate a low score in the other, nor is any specific combination of traits strictly reserved for one biological sex. We observe a broad spectrum of trait combinations, including androgynous and undifferentiated profiles, within both groups. In sharp contrast, the CGI measure captures gender identity along a single dimension, resulting in a bimodal distribution that leaves substantial within-sex heterogeneity in masculine and feminine traits undetected. By allowing masculinity and femininity to vary independently, our framework brings this heterogeneity to light.

Figure II: Heterogeneity in Femininity and Masculinity Between Sexes



Notes: This figure displays the probability density functions (PDFs) of gender typicality for biological men and women in the *test sample*. Densities are estimated using Epanechnikov kernel density estimation (KDE) to reveal the latent heterogeneity within each biological sex.
Two-Dimensional Measures (Panels A and C): For the new gender typicality measure (*WP*) and the *BSRI*, distributions are mapped onto a two-dimensional coordinate system with axes ranging from 1 (low) to 7 (high). The overlapping "clouds" for men and women highlight significant within-sex variation that is obscured by binary indicators.
Single-Item Measure (Panel B): For the Continuous Gender Identity (*CGI*) measure, the distribution is shown on a unidimensional axis ranging from 0 (Very Masculine) to 6 (Very Feminine). In contrast to the multidimensional measures, the CGI distributions are sharply bimodal and clustered at the biological sex extremes, reflecting a relatively smaller within-sex heterogeneity in self-perceived identity.

We next compare the predictive power of the three measures by estimating regressions of ten behavioral traits (absolute and relative confidence in math, absolute and relative confidence in words, risk (Holt and Laury), risk survey question, altruism, equality, efficiency, and competition) on each gender typicality measure in turn, always including the biological sex dummy. For each trait, we estimate four models: (1) the biological sex dummy only; (2) the biological sex dummy with *WP_feminine* and *WP_masculine*; (3) the biological sex dummy with the two BSRI sub-scales (BSRI women and BSRI men); and (4) the biological sex dummy with CGI.

Model 2 performs well on two fronts. First, including *WP_feminine* and *WP_masculine* substantially reduces the Female coefficient for traits such as risk, competition, and efficiency, indicating that these components capture variation previously attributed to biological sex. Second, Model 2 achieves higher adjusted R^2 than both the BSRI and CGI models for traits such as absolute confidence and risk, suggesting stronger overall explanatory power. Table III presents

all forty regressions, and Figure III visualizes the effect of each measure on the Female coefficient. Wald test results, shown in Table 21 in Online Appendix 2.4.2, provide formal significance comparisons across models.

A key interpretive note before examining each trait in detail: in Models 2, 3, and 4, a reduction in the magnitude and significance of the Female coefficient relative to Model 1 indicates that the gender identity measure is absorbing variation previously attributed to biological sex, that is, that gender-typical traits rather than sex per se drive the observed behavioral differences.[15]

Absolute and Relative Confidence. Considering absolute and relative confidence in the math task (Columns 1 and 2 of Table III), the model including WP_feminine and WP_masculine 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 the math task is positively correlated with masculinity, while the feminine component shows no significant correlation.

In terms of absolute confidence in the verbal task (Column 3 of Table III), none of the continuous gender measures provides extra fit in terms of statistical significance or absorbs the effect of the *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, absolute confidence in the math task has a greater variance (Variance Ratio Test, p-value < 0.0001), and our study may therefore be underpowered to detect gender differences in this specific task.

Regarding relative confidence (Column 4 of Table III), our measure reduces the magnitude of the *Female* coefficient the most. In terms of model fit, our model performs the best, though the difference is not substantial. It is also worth noting that women exhibit lower confidence levels than men in the verbal task, both in absolute and relative terms, which 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. For the incentivized risk task based on Holt and Laury (2002) (Column 5 of Table III), none of the three models provides a better fit or additional explanatory power. The low-stakes version of this task, as implemented here, has been shown not to correlate with self-reported risk preferences (Andreoni and Kuhn, 2019; Galizzi et al., 2016), which may explain why the gender difference in this type of risk preference remains unexplained by the additional gender identity measures.

For the survey-based risk measure (Column 6 of Table III), our gender typicality measure performs best in terms of absorbing the *Female* coefficient, absorbing more than both the BSRI and CGI models (see Figure III). Additionally, the difference in the *Female* coefficient between the model using our measure and the BSRI model is statistically significant ($p < 0.001$ in a Wald test) (see Table 21 in Online Appendix 2.4.2). In terms of model fit, our measure achieves

[15]Notice that if one is interested in the average gap between men and women, the biological sex dummy Female remains a valid and useful summary measure. In this context, not including the gender typicality measure is not a bias per se, as the two approaches simply answer different empirical questions. Controlling for biological sex alone estimates the average behavioral difference between men and women, whereas our framework additionally reveals how much of that difference is associated with variation in gender-typical traits within each biological sex group.

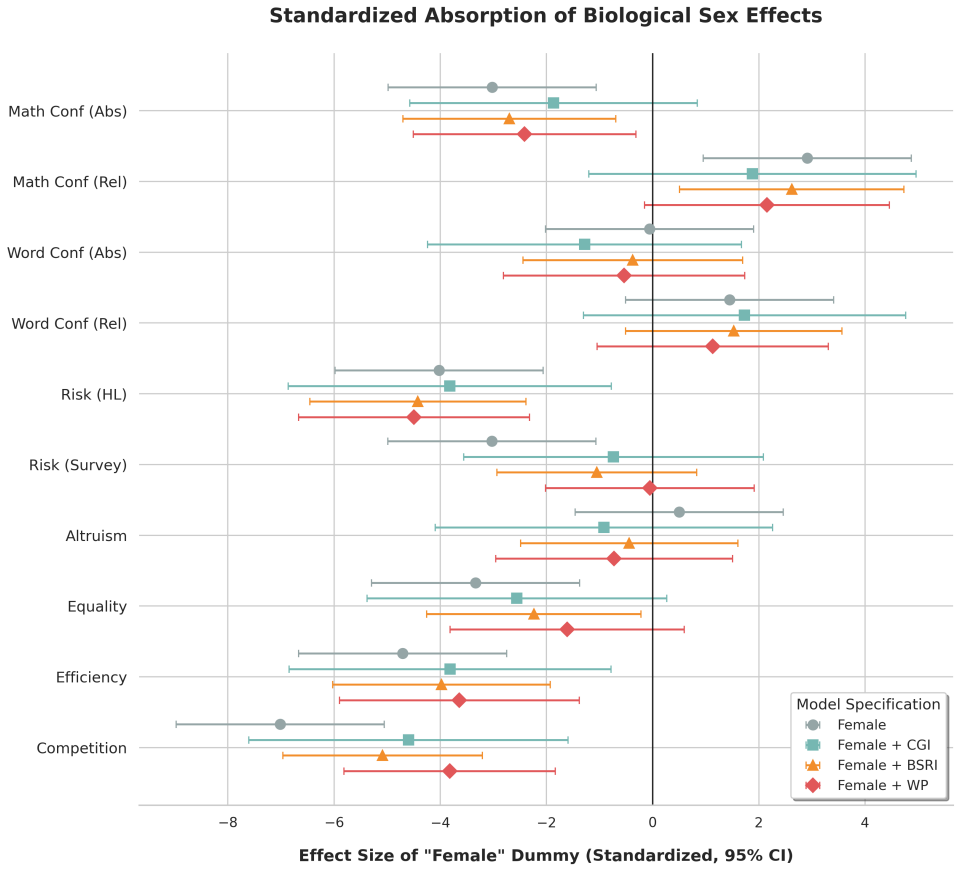
Table III: Predictive Power of Gender Typicality Measures on Behavioral Traits

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)
BSRI_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)
BSRI_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 typicality measure (*WP_feminine* and *WP_masculine*), 3. *Female* + BSRI without attributes *willingness to take risk* and *competitiveness* (*BSRI_women* and *BSRI_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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

the best performance, as indicated by the highest adjusted R^2. We further extend our analysis to different contexts of risk preference, namely life, financial, occupational, and health-related, in Online Appendix 2.4.1, and we observe the same pattern in terms of adjusted R^2 and the significance of the coefficients across all these domains.

Figure III: Comparison of the *Female* Coefficient Among Four Models for Each Behavioral Trait



Notes: This figure plots the standardized coefficients for the biological sex dummy (*Female*) from the 40 regression models presented in Table III, grouped by behavioral trait. To facilitate comparison across varying scales of the dependent variables, all coefficients are standardized by the baseline (*Female* alone model) standard error.
For each trait, the markers represent the point estimate of *Female* and its 95% confidence interval across four specifications: (1) *Female* alone (gray circles), (2) *Female* + *CGI* (teal squares), (3) *Female* + *BSRI* (orange triangles), and (4) *Female* + *WP* typicality components (red diamonds).
The movement of markers toward the vertical dashed line at zero illustrates the extent to which each typicality measure absorbs the biological sex effect. Intervals crossing the zero-line indicate that the sex dummy is no longer statistically significant at the 5% level.

Altruism, Equality, Efficiency, and Competition. Across these four traits (Columns 7, 8, 9, and 10 of Table III), our measure outperforms BSRI and CGI primarily in its ability to absorb the *Female* coefficient, particularly for equality, efficiency, and competition (see Figure III). For equality and competition, the reduction in the *Female* coefficient between our model and the BSRI model is statistically significant ($p < 0.001$ in a Wald test) (see Table 21 in Online Appendix 2.4.2). Regarding model fit, our model closely matches the BSRI model in terms of adjusted R^2 for altruism, equality, and efficiency. For competition, our model outperforms the others, achieving the highest adjusted R^2 with only 16 attributes compared to BSRI's 38, suggesting that our measure offers a more parsimonious solution without sacrificing fit. In

terms of the significance of the gender identity components, femininity is the primary driver of altruism, equality, and efficiency, while competition is more strongly associated with masculinity.

In summary, our measure reduces the significance of the *Female* coefficient relative to the model including *Female* alone in risk (survey question), competition, altruism, equality, and efficiency (Table III). In terms of model fit, it outperforms BSRI and CGI in adjusted R^2 for the risk survey question and competition, while performing comparably for the remaining traits. For all traits where the *Female* coefficient is significant in the baseline model, our measure reduces its significance more than the BSRI measure (Table III). Crucially, it achieves this with only 16 attributes compared to BSRI's 38, making it a more parsimonious tool. Moreover, its two-dimensional nature allows us to identify whether observed differences are driven by masculinity or femininity, providing richer insights than either a binary sex dummy or a unidimensional continuous measure.

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 those traits for which the Female coefficient is initially significant.*

To ensure the robustness of our results, we conducted multiple checks detailed in Section 2.4.4 of the online appendix. First, we excluded the biological sex variable *Female* to evaluate the predictive power of our gender typicality 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 typicality 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 typicality measure.

5 Labor Market Outcomes

In the third phase of data collection, we obtained additional information on labor market outcomes through a follow-up survey conducted after the main experiment. This survey was administered between October 31 and November 9, 2024, and all participants from the first and second waves were invited to participate. Approximately 46% of the original sample (504 out of 1,102 participants) completed the follow-up, which took place two years after the initial wave. For the analysis of labor market outcomes, we utilize the full 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.

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

Notes: Robust standard errors are reported in parentheses.
$^{\#}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

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, 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 fewer hours worked, significantly lower income, 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 overall educational attainment (HighEd).

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 in the saturated specification the partial effect of *Female* is imprecisely estimated. By contrast, masculinity emerges as a strong and robust predictor of labor market outcomes: higher masculinity scores are associated with increased hours, higher income, a greater probability of holding managerial positions, receiving bonuses, and engaging in negotiations, as well as sorting into less female-dominated occupations and a higher likelihood of high educational attainment.

Femininity exhibits a more limited association with labor market outcomes, though it is significantly and negatively correlated with hours worked. Allowing the returns to gender typicality to vary by biological sex reveals limited heterogeneity: interaction terms between femininity and sex are mostly small and statistically insignificant, with only marginal negative interactions appearing for natural sciences and STEM studies. 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 is associated with variation in gender-typical traits, particularly masculinity, rather than biological sex per se.

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 BSRI 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

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

	(1) Hours	(2) Income	(3) Manager	(4) Bonus	(5) Negot	(6) Women%	(7) NatSci	(8) STEM	(9) HighEd
Panel A: Predictive Power of Biological Sex									
Female	-0.304$^{\#}$	-0.526***	-0.079$^{\#}$	-0.145*	-0.301*	7.373***	-0.248***	-0.277***	-0.055
	(0.161)	(0.128)	(0.044)	(0.066)	(0.131)	(1.500)	(0.051)	(0.050)	(0.046)
R-squared	0.145	0.256	0.093	0.038	0.067	0.153	0.103	0.122	0.047
Observations	504	504	504	481	489	489	384	384	504
Panel B: Gender Typicality and Labor Market Outcomes									
Female	-0.555	0.260	-0.066	0.756*	-0.048	-6.209	0.148	0.106	0.555$^{\#}$
	(1.061)	(0.756)	(0.275)	(0.381)	(0.861)	(9.858)	(0.321)	(0.317)	(0.287)
Femininity	-0.207*	-0.092	0.009	-0.013	-0.045	0.253	0.034	0.032	0.021
	(0.105)	(0.107)	(0.038)	(0.047)	(0.103)	(1.486)	(0.045)	(0.044)	(0.037)
Female × Femininity	0.197	0.008	-0.031	-0.002	-0.009	0.365	-0.105$^{\#}$	-0.104$^{\#}$	-0.078
	(0.181)	(0.148)	(0.056)	(0.070)	(0.161)	(1.936)	(0.062)	(0.061)	(0.055)
Masculinity	0.285*	0.390***	0.071*	0.206***	0.328**	-3.159**	-0.000	0.000	0.093**
	(0.111)	(0.099)	(0.032)	(0.057)	(0.104)	(1.203)	(0.042)	(0.042)	(0.032)
Female × Masculinity	-0.092	-0.125	0.045	-0.186*	0.013	2.369	0.024	0.026	-0.046
	(0.173)	(0.135)	(0.043)	(0.075)	(0.138)	(1.677)	(0.056)	(0.055)	(0.048)
R-squared	0.164	0.299	0.129	0.082	0.120	0.171	0.111	0.130	0.077
Observations	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 account for selective attrition. All specifications include session fixed effects and control for age and ethnicity. Columns (1)–(9) additionally control for education level. In column (10) (High Education), the education control is excluded as it defines the outcome.
$^{\#}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

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 BSRI 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 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

ine and masculine components, thereby accounting for the latent heterogeneity within biological sex categories, providing additional explanatory power for previously detected gender differences. Each component is constructed from distinct attributes, resulting in a gender typicality measure that minimizes the influence of experimenter 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. The two-component nature of the measure also mitigates potential multicollinearity issues that typically arise from the inherent bipolarity of single-item scales, which treat masculinity and femininity as mutually exclusive. Moreover, compared to the BSRI measure, our measure substantially reduces the number of attributes from 38 to 16, 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. Specifically, we document that confidence, risk-taking, and competitiveness are strongly correlated with masculine traits, whereas altruism, efficiency concerns, and equality concerns are more closely correlated with feminine traits. Importantly, once gender-typical traits are accounted for, the explanatory role of biological sex declines substantially across several domains, suggesting that some documented gender gaps may reflect variation in conformity to socially defined norms rather than inherent differences between men and women. We extend these findings to labor market outcomes, showing that masculinity is associated with higher income, managerial positions, performance-based pay, and willingness to negotiate.

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 provide a validated and parsimonious two-dimensional measure of gender typicality that outperforms existing measures in predicting a broad set of economic behaviors and labor market outcomes. Third, by leveraging machine learning to uncover latent structure in gender-related attributes, we show how modern empirical tools can reveal economically meaningful heterogeneity that coarse binary indicators overlook.

Our results have potential implications for how economists and policymakers think about gender gaps. By showing that observed disparities in economic behavior and labor market outcomes are associated with gender-typical traits rather than solely with biological sex, our framework suggests that future research should move beyond binary sex categories when study-

ing the origins and persistence of these gaps. Identifying which traits are associated with specific disparities, namely, masculinity for confidence, risk-taking, competition, and labor market success, and femininity for altruism and equality concerns, can help direct future causal research toward the mechanisms underlying these patterns, and ultimately inform the design of targeted interventions. We view establishing causality as an important and promising avenue for future work.

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), 16–30.

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.

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.

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 (2025). *Childhood gender nonconformity and gender gaps in life outcomes*. Helsinki Graduate School of Economics.

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.

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.

Budanitsky, A. and G. Hirst (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics 32*(1), 13–47.

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.

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.

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 (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

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.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT press.

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.

Gürel, B., H. Fornwagner, and S. Palan (2025). Rethinking gender identities and sexual orientation: The multidimensional gender and sexuality inventory (mgsi). *Available at SSRN 5501859*.

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.

Kamas, L. and A. Preston (2012). The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior & Organization 83*(1), 82–97.

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.

Lemaster, P. and J. Strough (2014). Beyond mars and venus: Understanding gender differences in financial risk tolerance. *Journal of Economic Psychology 42*, 148–160.

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.

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

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.

West, C. and D. H. Zimmerman (1987). Doing gender. *Gender & Society 1*(2), 125–151.

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.