


# Scaling Up: Representing Gender Diversity in Survey Research

Socius: Sociological Research for  
a Dynamic World  
Volume 2: 1–11  
© The Author(s) 2016  
DOI: 10.1177/2378023116664352  
srd.sagepub.com  


Devon Magliozzi<sup>1</sup>, Aliya Saperstein<sup>1</sup>, and Laurel Westbrook<sup>2</sup>

## Abstract

Survey measures of gender have been critiqued for failing to reflect the diversity of the population. Conventionally, respondents to national surveys are categorized as female or male. Calls for improvement have centered on adding additional categories, such as transgender. We propose that in addition to revising categorical gender measures, national surveys should incorporate gradational measures of femininity and masculinity to better reflect gender diversity and sharpen models of gender inequality. Our results from two national pilot studies show that conventional measures mask significant variation within the categories of female and male. For example, less than a quarter of respondents reported that they are very feminine or masculine, respectively, and not at all the other. We also demonstrate that scale responses can be treated as independent variables in studies of inequality or as dependent variables that allow gender identification to be an outcome of social processes.

## Keywords

survey design, gender, sex

Ever since national surveys emerged in the mid-twentieth century as a means of tracking the attitudes and habits of the American population (Igo 2007), it has been the default to collect information on whether respondents are female or male.<sup>1</sup> The treatment of gender as a fundamental, universal demographic attribute in survey research is not surprising; in everyday interactions, individuals make nearly automatic assessments about others' gender in order to navigate the social world (Ridgeway 2011), and gender is widely recognized as an axis of inequality. Despite the consensus that surveys ought to measure respondents' gender, until recently there has been insufficient consideration of how surveys should measure this concept.

Calls to improve the measurement of gender have focused on representing categorical gender diversity by offering additional response options (e.g., transgender<sup>2</sup>), distinguishing between assigned sex at birth and current gender identity, and allowing respondents to self-identify (see e.g., GeNIUSS Group 2014; Harrison, Grant, and Herman 2011). These are important issues to address when updating the use of gender in surveys, yet the reach of these revisions is limited (see Westbrook and Saperstein 2015). For example, there are concerns that in many surveys, expanded categorical measures will yield some populations that would be too small for statistical analysis. Improved categorical measures also do not allow for variation within gender categories; such questions continue to

treat gender as a set of discrete attributes, each assumed to describe a relatively homogenous population.

To facilitate more nuanced analysis in national surveys of the general population, we propose that in addition to revising the standard categorical measure, survey researchers use more gradational measures of gender identification. In particular, we recommend measuring femininity and masculinity on separate scales to account for diversity within and overlap between gender categories. We tested femininity and masculinity scales on two national samples, alongside categorical measures of sex at birth and current gender. Less than one-quarter of respondents reported seeing themselves as very feminine or masculine and not at all the other, yet this strictly dichotomous gender identification is the only notion of gender allowed by current measures. In addition to showing that femininity and masculinity scales reflect the range of gender diversity better than standard categorical measures, we demonstrate that scale scores can serve as either an independent or a dependent variable, opening new avenues of research on gendered attitudes, behaviors, and inequality.

<sup>1</sup>Stanford University, Stanford, CA, USA

<sup>2</sup>Grand Valley State University, Allendale, MI, USA

## Corresponding Author:

Devon Magliozzi, Department of Sociology, Stanford University,  
450 Serra Mall–MC 2047, Stanford, CA 94305-2047, USA.  
Email: [dmaglioz@stanford.edu](mailto:dmaglioz@stanford.edu)



## Critique of Conventional Gender Measures

Survey research methods have long been a source of concern among gender scholars. The critique that survey researchers use gender as a variable rather than treating it as a socially constructed system (e.g., Stacey and Thorne 1985), along with broader criticisms of quantitative analysis, have led some scholars to avoid survey data entirely.<sup>3</sup> Others maintain that there is nothing inherently problematic about survey research (Oakley 1998; Sprague 2005) and that statistical analysis of large-scale surveys is necessary to provide entry into policy debates (Harnois 2013; Williams 2006). We take a similar approach that recognizes the role survey research can play in revealing broad patterns of inequality but also emphasizes that the method is only as good as its measures. Although previous work exposed patriarchal assumptions built into surveys, such as the assignment of men as “household heads” (Presser 1998), and has questioned the measurement of gender inequality (e.g., Permanyer 2010), until recently, the measurement of gender itself has remained largely taken for granted.

Survey measures of sex and gender began to attract attention in concert with calls for data collection on LGBT populations to help monitor health disparities and discrimination in employment (e.g., Balarajan, Gray, and Mitchell 2011; Human Rights Watch 2011; Institute of Medicine 2011). Subsequent proposals have focused primarily on the lack of transgender-inclusive response options in conventional measures. Surveys fielded among transgender respondents have demonstrated the feasibility of using a two-question method to measure sex at birth and current gender separately (Deutsch et al. 2013; GenIUSS Group 2014) and of including responses beyond female and male, such as intersex, transgender, genderqueer, and “a gender not listed here” (Harrison-Quintana, Grant, and Rivera 2015; Ingraham, Pratt, and Gorton 2015; Schilt and Bratter 2015).

Changing the categories used to measure sex and gender, “especially [moving] from binaries to multiplicities,” challenges the belief that sex and gender categories are natural and dichotomous (Lorber 2006:451) and grants recognition to otherwise uncategorized populations. However, adding more categories alone cannot solve all the dilemmas of representing population diversity. Any attempt to create a survey measure that is inclusive of all possible categorical responses will inevitably fall short of that goal. In the case of gender, individuals use myriad terms to self-identify—most commonly *woman* and *man* but also *transgender*, *genderqueer*, *androgynous*, *bigender*, *gender fluid*, and many others (see Singer 2015). Much like racial identification, some of these terms are in flux, so no closed set of answer options could anticipate all the terms respondents might use. Allowing open-ended responses addresses part of the problem, but analysts confronted with few respondents in any given gender category will likely exclude small populations from analysis or

aggregate all of them into a single umbrella category, such as transgender (Singer 2015). Furthermore, nontraditional gender practices are not confined to transgender people; all genders are part of a complex and unstable system of expectations and experiences (Butler 1993). This heterogeneity within and overlap between gender categories will not be captured by a categorical measure, regardless of how many answer options are offered. Thus, although expanded response options on surveys better reflect gender diversity than conventional measures do, survey measures of gender can be further improved by moving beyond categorical distinctions.

We argue that a more thorough retooling of the use of gender in surveys should include using femininity and masculinity scales as measures of gender identification. Scale items will allow respondents to report a more nuanced sense of self regardless of how they might be classified in categorical gender terms. Our recommended measures avoid pitfalls of prior gender scales not only by allowing respondents to self-identify but also by measuring femininity and masculinity separately, ensuring the two concepts are neither treated as mutually exclusive nor operationalized as opposites. Incorporating such scales as a regular feature of social surveys will allow for assessments of variation in gender identification over time and across regions or contexts.

## Development of Scales

Although psychologists have long used and debated scale measures of femininity and masculinity (e.g., Spence 2011; Spence and Buckner 2000; Wylie et al. 2010), gradational measures of gender identification have not been widely employed in large-scale survey research (for exceptions, see Hunt et al. 2007; McLaughlin, Uggen, and Blackstone 2012). Early instruments in psychological studies scored femininity and masculinity along a single, bipolar scale (Gough 1952; Terman and Miles 1936). By the 1970s, the assumptions implied by the use of a bipolar scale—namely, that femininity and masculinity are mutually exclusive and opposite—were being called into question (Constantinople 1973). The Bem Sex Role Inventory (BSRI), a 60-item index of gendered trait ratings that are combined to assign feminine, masculine, and androgynous scores to respondents, subsequently became a standard instrument (Bem 1974). However, even the more abbreviated 30-item BSRI cannot feasibly be incorporated in the large-scale surveys that are key sources of data for social science research. Our aim is to build on the foundational work in psychology on scale measures of gender but adapt those insights for use in a wider range of studies.

Like the BSRI, the gender scales that we propose treat femininity and masculinity as distinct, orthogonal dimensions. However, we depart from the BSRI not only in the number of items we use but also by allowing respondents to determine which criteria contribute to their gender self-identification. Although the BSRI includes self-reported femininity and masculinity scales in its index, most of the instrument

asks how well stereotypically feminine and masculine attributes describe respondents. Thus, the BSRI and other similar instruments have been critiqued for imposing definitions of femininity and masculinity on respondents by relying on gender stereotypes to assign scale scores (see e.g., Connell [1995]2005; Gill et al. 1987; Hoffman and Borders 2001). For example, if a person reports on the BSRI that they are “often” or “always or almost always” gentle or compassionate, their femininity score increases, while reporting that one is “often” or “always or almost always” assertive or analytical increases a respondent’s masculinity score. The resulting score does not describe a person’s gendered sense of self but rather the extent to which they conform to a set of stereotypes. Moreover, because the BSRI’s index of stereotypes only includes personality traits, such as being gentle or assertive, the resulting score does not provide an overall measure of a person’s femininity or masculinity. Researchers have updated the traits included on the BSRI to reflect contemporary stereotypes about gendered personalities (Auster and Ohm 2000; Harris 1994), but even the updated index cannot account for how other factors, such as a person’s appearance or occupation, may bear on their gender identification (see Spence 2011).

In contrast to instruments like the BSRI, our scales measure gender identification directly and produce general measures of femininity and masculinity by allowing people to draw on whichever factors contribute to their gendered sense of self when responding. This approach is similar to other commonly used gradational measures, such as political ideology scales that ask respondents to rate themselves as liberal or conservative (see Jost, Federico, and Napier 2009). In addition, by granting respondents control over their gender identification rather than imposing a fixed set of criteria, we recognize that people construct complex gender identities and modes of expression as they navigate a system of gendered expectations and institutions (Risman 2004) and that responding to the scales is itself a means of “doing gender” (West and Zimmerman 1987).

Survey data are currently populated by females and males; to the extent that surveys are used to gauge social behaviors or institutional patterns, only females and males can act and institutions can only affect them on the basis of femaleness or maleness. Our brief measures of gender identification can be feasibly incorporated into general surveys to produce research that more closely approximates the complexity of the gendered social world. Repopulating survey data with a broader range of gendered individuals will enable researchers to sharpen explanations of gender inequality and study gender identification as an outcome of social processes.

## Implementing the Scales

We fielded two national surveys to assess the feasibility and potential applications of including femininity and

masculinity scales in social surveys aimed at the U.S. adult population. Each survey included a sex and gender module consisting of six questions measuring first-order femininity and masculinity (how do you see yourself), third-order femininity and masculinity (how do most people see you), as well as sex assigned at birth and current gender (Figure 1). Responses on first- and third-order scales were not statistically distinct on average, so for the sake of simplicity in demonstrating how gradational measures can augment categorical ones, we focus on findings from the first-order femininity and masculinity scales.<sup>4</sup>

The first survey was designed to gauge respondents’ ability to respond to unconventional sex and gender questions and therefore presented our sex and gender module as a standalone questionnaire with space for open-ended feedback.<sup>5</sup> Following the successful pilot, our second study aimed to approximate typical survey questionnaires used in the social sciences; thus, the sex and gender module was embedded within a series of more than 40 questions drawn from the General Social Survey (GSS). Unless otherwise specified, the findings presented refer to our second survey.<sup>6</sup>

We recruited respondents for both studies using Amazon Mechanical Turk (MTurk) in May and November 2014. MTurk is an online platform for recruiting workers to complete tasks, such as labeling images, testing hyperlinks, and responding to surveys. MTurk has been embraced by social scientists as a means to quickly and inexpensively recruit respondents for pilot surveys and experimental research (Paolacci and Chandler 2014). Although MTurk workers are not nationally representative, MTurk provides a more diverse pool of respondents compared to the undergraduate student samples often relied on in experimental research (including pretests for the BSRI). The quality of data provided by MTurk workers also compares favorably to online, population-based samples (Weinberg, Freese, and McElhattan 2014), and the use of performance ratings from past MTurk assignments can help to ensure the task of completing a survey will be taken seriously.<sup>7</sup>

The respondents we recruited to both surveys were required to be U.S. residents, 18 years or older, were only able to complete one of the two surveys, and were prevented from accessing either survey more than once. Respondents earned \$0.25 and \$1.50, respectively, which is consistent with standard pay rates on MTurk. After responses were collected, surveys submitted from IP addresses outside the United States or submitted from IP addresses with duplicate submissions were dropped from the samples. The surveys yielded 1,521 and 1,522 valid responses.

Because our second survey was designed to resemble the GSS, we compared our respondents to the 2014 GSS respondents on a number of key demographic characteristics. Our survey sample aligns well with the 2014 GSS distribution for respondent’s sex, and the regional and political party

### First-order gender scale

In general, how do you see yourself? Please answer on both scales below.

	Not at all	1	2	3	4	5	Very
Feminine	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Masculine	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Third-order gender scale

In general, how do most people see you? Please answer on both scales below.

	Not at all	1	2	3	4	5	Very
Feminine	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Masculine	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Sex at birth

What sex were you assigned at birth?  
(For example, on your birth certificate.)

☐ Female

☐ Male

☐ Intersex

### Categorical gender identification

What is your current gender?

☐ Woman

☐ Man

☐ Transgender

☐ A gender not listed here (please specify)

**Figure 1.** Sex and gender survey module.

Note: Each question appeared on a separate page in a survey fielded by the authors on Amazon Mechanical Turk, November 2014.

affiliation distributions are also quite similar between the two surveys (Table A1). Relative to the GSS, which recruits respondents using a multistage area probability sample, our sample overrepresents whites and Asians; respondents also are younger, on average, and more highly educated. These patterns are consistent with previous studies conducted on MTurk (see e.g., Berinsky, Huber, and Lenz 2012). Given the nature of our sample, we expect that it will provide somewhat higher estimates of nontraditional gender identification than a nationally representative survey. However, the diversity of our sample also allows us to highlight key differences along these lines that will be fruitful avenues for future research.

## Weighing the Results

Conventional survey measures of gender are blunt tools. Our proposed femininity and masculinity scales enable respondents to better express nuanced gender identifications and researchers to better track patterns in gender inequality. The distribution of scale responses on both surveys confirms that conventional gender measures mask diversity among cisgender<sup>8</sup> and transgender respondents alike. Analyses using scale scores as an independent variable show that conventional gender measures can conceal gradational disparities in outcomes while treating scale scores as

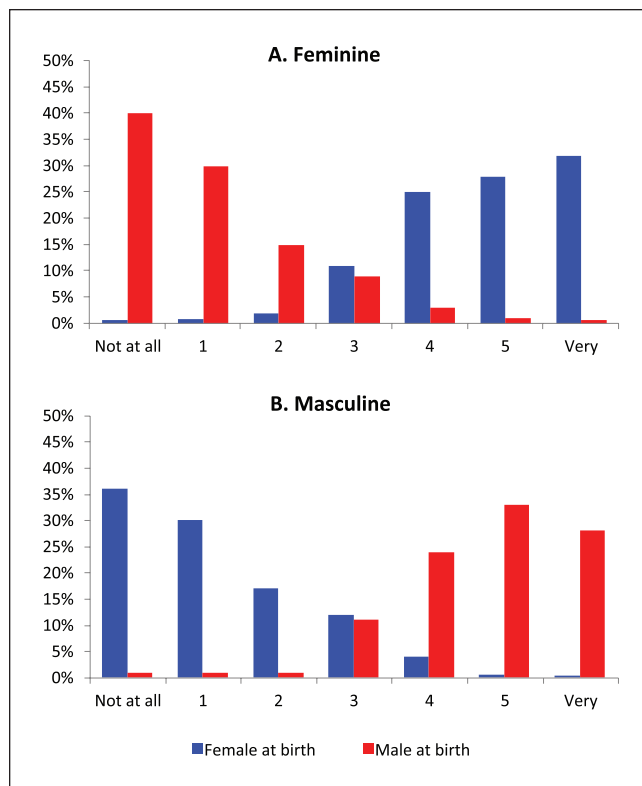
a dependent variable allows researchers to question the direction of causality underlying otherwise descriptive findings of gender difference.

## Diversifying Data

In both surveys, our scales uncovered significant variation in gender identification among both cisgender and transgender respondents. Aggregate results show that respondents made use of the full seven-point scales to describe their femininity and masculinity, with substantial overlap between the scale responses of respondents assigned female and male at birth (Figure 2). Further, although more than 99 percent of both samples would be classified as cisgender based on their categorical responses,<sup>9</sup> femininity and masculinity scale responses suggest that significantly fewer than 99 percent of respondents see their gender in traditionally dichotomous, categorical terms.

We also calculated *gender polarization* scores as the absolute value of the difference between each person's responses on the femininity and masculinity scales (Table 1). If the female/woman and male/man implied by standard measures are assumed to be only feminine or masculine, respectively, and not at all the other, polarization scores reveal how much respondents' gender scale responses deviate from these assumptions. Less than a quarter (24 percent) of respondents received a polarization score of 6, meaning they reported





**Figure 2.** Distribution of gender identification by sex at birth.  
Source: Authors' survey fielded on Amazon Mechanical Turk, November 2014.

seeing themselves as very feminine/masculine and not at all the other. Gender identification for the remaining 76 percent of respondents included either tempered or overlapping femininity and masculinity. Among them, 7 percent of the sample reported identical feminine and masculine responses, giving them polarization scores of zero, while nearly 4 percent of respondents reported a lower score on the gender scale that

“matches” their sex at birth than on the “cross-gender” scale—that is, 33 females saw themselves as more masculine than feminine, and 24 males saw themselves as more feminine than masculine. Thus, scales reveal greater gender diversity than is counted by conventional measures.

The scales additionally expose the social contingency of gender identification. We find statistically significant patterns in scale responses by region, age, sexual orientation, and self-identified race, among other factors (Table 2).<sup>10</sup> For example, respondents in the South were significantly more likely to give very polarized responses than people elsewhere in the country, affirming that gender is not a natural attribute but rather is culturally inflected. Respondents over the age of 30 (the median age in our sample) also reported significantly more polarized gender identifications than their younger counterparts. Repeated inclusion of femininity and masculinity scales will allow researchers to determine if this age gap is related to cohort differences or changes in gender identification over the life course (or both). Heterosexual respondents reported more polarized gender identifications than their gay, lesbian, or bisexual counterparts as well, suggesting that scale responses will enable researchers to disentangle whether inequality associated with sexual orientation is mediated or moderated by conformity to gender expectations. With more racially diverse samples, femininity and masculinity scales will also allow researchers to advance studies of intersectionality by exploring how racial and gender identification co-vary (e.g., Galinsky, Hall, and Cuddy 2013).

### Situating Scale Responses

Open-ended feedback from the first survey sheds light on how respondents understood the scales and selected responses. Though it was optional, over a third of respondents provided feedback, with 14 percent ( $N = 209$ ) opting to explain their responses to the sex and gender questions. These results

**Table 1.** Gender Polarization and Scale Response Distributions.

	Scale Response Distributions (Percentage)							Mean	Standard Deviation
	Not at All	1	2	3	4	5	Very		
Female at birth									
Polarization	8	8	11	18	18	12	24	3.7	1.9
Feminine	0.6	0.8	2	11	25	28	32	4.7	1.2
Masculine	36	30	17	12	4	0.6	0.3	1.2	1.2
Male at birth									
Polarization	7	6	13	15	21	14	23	3.7	1.8
Feminine	40	30	15	9	3	1	0.7	1.1	1.2
Masculine	1	1	1	11	24	33	28	4.7	1.2

Source: Survey fielded on Amazon Mechanical Turk, November 2014.

Note:  $N = 1,522$ ; 805 females at birth and 717 males at birth. Scales were coded from 0 to 6. Polarization is the absolute value of the difference between the two scales.

**Table 2.** Gender Polarization by Demographic Characteristics.

	Percentage Very Polarized	N
<i>Gender</i>		
Cisgender	24	1,514
Transgender	13	8
<i>Sex at birth</i>		
Female	24	805
Male	23	717
<i>Region</i>		
South	27*	571
West	23	374
Midwest	19*	312
Northeast	24	262
<i>Education</i>		
College degree	22	883
No college degree	26	639
<i>Age</i>		
Over 30	28***	813
30 and younger	19	709
<i>Sexual orientation</i>		
Heterosexual or straight	26***	1,375
Gay, lesbian, homosexual, or bisexual	4	147
<i>Hispanic origin</i>		
Hispanic	31	110
Not Hispanic	23	1,412
<i>Self-identified race</i>		
White	22**	1,237
Black or African American	49***	101
All other responses	21	184
<i>Party affiliation</i>		
Democrat	26	619
Republican	31**	263
Independent	19***	580
All other responses	25	60
<b>Total</b>	<b>24</b>	<b>1,522</b>

Source: Authors' survey fielded on Amazon Mechanical Turk, November 2014.

Note: For polytomous variables (region, race, and party affiliation), each comparison is tested separately, as if it were dichotomous (e.g., South versus all else, white versus all else, black versus all else).

\* $p < 0.05$ . \*\* $p < 0.01$ . \*\*\* $p < 0.001$  (two-tailed tests).

indicated that people took into account a range of factors when reporting their femininity and masculinity, including their appearance, hobbies, and occupation, in addition to citing personality traits, such as those measured in the BSRI. For example, a cisgender woman who sees herself as a 4 on the femininity scale and a 2 on the masculinity scale explained, "I grew up as a tomboy, and I'm an electrical engineer, a field that is almost all male. I also can have quite an assertive personality at times, so I feel somewhat masculine, but more feminine still."<sup>11</sup> This respondent identifies with the gender category "woman" but also sees herself as masculine with regards to her assertiveness and occupation. A cisgender man who sees himself as a 2 on the femininity scale and a 5 on the masculinity scale commented, "I

consider myself in the metrosexual sort of group. I'm a male who likes females, who is concerned about his skin, clothes, and looks a bit more than most of my friends." This respondent reflected on his categorical sex, his sexuality, and his self-presentation and ultimately decided that he has both feminine and masculine characteristics. By asking for an overall sense of their femininity and masculinity, our scales allowed respondents to weigh multiple dimensions of their gender identification.<sup>12</sup>

The scale responses and feedback provided by transgender respondents further highlight the heterogeneity hidden by categorical measures. For example, a respondent who was assigned male at birth and currently identifies as transgender reported a femininity score of 5 and a masculinity score of 0. In explanation, this person said, "I am small and have a female's body, it has always been a curse when trying to make it with the ladies. But now, I have accepted it and embrace it." On a conventional gender measure, this respondent could be recorded as either female or male, but neither response would fully reflect their gendered sense of self. The same is true for a respondent who was assigned female at birth, wrote in "genderqueer or gender neutral," and selected a 3 on the femininity scale and 4 on the masculinity scale. They commented, "This is the best survey ever—I get so sick of surveys asking which 'gender' I am and then providing only two options, neither of which are my gender identity." On a survey with expanded categorical gender measures, these respondents would be better able to report their current gender and sex at birth. However, categorical measures alone would render these two respondents difficult to distinguish—both identify outside the conventional gender binary, and thus they likely would be collapsed into a single group for statistical analysis. Including gradational measures of femininity and masculinity alongside categorical items allows these respondents to report their gender in a more nuanced way and reveals that in terms of femininity and masculinity, they are far from similar. The first respondent's gender identification is highly polarized and feminine, while the second reports femininity and masculinity scores in the middle of each scale. Averaging their experiences under an umbrella categorical variable would hide the extent to which their differing gender identifications bear on their life experiences.

### Complicating Gender Inequalities

Because traditional survey measures of gender allow just two mutually exclusive responses—female and male—quantitative studies of inequality can only depict and model gender gaps in binary, categorical terms. Incorporating femininity and masculinity scales into surveys can expose hidden dimensions of gender inequality. Health researchers, for example, have found that relying on binary sex categories in analysis masks an association among men between self-identified femininity and decreased risk of death from heart disease (Hunt et al. 2007; see also Hammarström and Annandale 2012). Similarly, studies that incorporate femininity and masculinity scales have revealed associations between gender identification and abuse in dating

(Burke, Stets, and Pirog-Good 1988) and workplace harassment (McLaughlin et al. 2012) as well as how gender identification may be altered by marriage (Burke and Cast 1997).

We further illustrate how scales can sharpen models of inequality by examining the association between femininity and masculinity scale responses and marital status. Marital status has been mechanically related to conventional categorical measures of gender in the United States—because the vast majority of marriages have been heterosexual, women and men have been represented equally among married people. However, our results indicate that marital status is related to gender polarization scores (Table 3). People with very polarized responses have 50 percent greater odds of being married, net of sex at birth, compared to people with less polarized gender identifications. The direction of causality in this association is unclear—traditional, dichotomous gender identification could increase the odds of marriage, or marriage could increase conformity to traditional gender norms (Burke and Cast 1997). Repeat measures of femininity and masculinity in panel studies will allow researchers to further investigate the association, extending conversations about the role of marital status in maintaining inequality (Waite and Lehrer 2003).

More broadly, if gender scales become a regular feature of large-scale surveys, researchers will be able to study gender identification as both a determinant and a consequence of one's life experiences, allowing for the possibility that gender inequality is produced and reproduced in a positive feedback loop. For example, feminine behaviors and tasks are often devalued in the workplace, but sanctions for violating expectations compel women to engage in them nonetheless (Eagly and Carli 2007; England 2010). Similarly, men often engage in behavior known to be risky or detrimental to health in the interest of conforming to masculine expectations (Courtenay 2000). We need better measures of gender to understand, and potentially to interrupt, such cycles of inequality. We know that some women break into professional leadership and some men avoid health risks; would they describe their gender differently, in terms of femininity or masculinity, than their peers? If so, would differences in gender identification precede or follow differences in gendered behaviors or experiences? Femininity and masculinity scales will allow researchers to account for variation within gender categories and offer the potential to disentangle causal mechanisms that are missed in standard studies of gender inequality.

## Conclusion

We argue that including femininity and masculinity scales in survey research better reflects the complexity of the social world than current categorical measures alone and thus can shed new light on axes of inequality. Unlike standard gender scales that rely on a series of trait ratings, reducing femininity and masculinity to stereotype conformity, our proposed items give primacy to a more comprehensive measure of self-identification. Recording such gradational measures of gender, in

**Table 3.** Revealing a Relationship Between Marital Status and Gender Polarization.

	GSS	MTurk	
	(1)	(2)	(3)
Gender polarization			
Very polarized (6)	—	—	1.57** (3.33)
Sex (ref: Female)			
Male	1.11 (1.18)	0.53*** (-5.42)	0.52*** (-5.46)
Region (ref: South)			
West	0.79 (-1.89)	0.86 (-1.05)	0.87 (-0.96)
Midwest	1.09 (0.70)	0.89 (-0.78)	0.90 (-0.64)
Northeast	0.75* (-2.11)	0.70* (-2.08)	0.70* (-2.09)
Education (years)	1.13*** (7.60)	1.12*** (5.21)	1.12*** (5.37)
Age	1.01*** (4.17)	1.03*** (6.66)	1.03*** (6.32)
Sexual orientation (ref: Heterosexual or straight)			
Gay, lesbian, homosexual or bisexual	0.25*** (-5.24)	0.44*** (-3.70)	0.48** (-3.29)
Hispanic origin	1.04 (0.25)	0.86 (-0.66)	0.82 (-0.83)
Race (ref: White)			
Black	0.34*** (-6.92)	0.73 (-1.34)	0.64 (-1.85)
All other responses	0.98 (-0.11)	0.75 (-1.55)	0.75 (-1.52)
Political Party (ref: Democrat)			
Republican	1.84*** (4.66)	2.50*** (5.65)	2.46*** (5.53)
Independent	1.06 (0.58)	1.34* (2.24)	1.38* (2.43)
All other responses	0.92 (-0.27)	0.90 (-0.31)	0.89 (-0.35)
Constant	0.09*** (-7.77)	0.05*** (-7.74)	0.04*** (-7.99)
N	2269	1518	1518

Sources: General Social Survey (GSS) 2014 cross-sectional sample and authors' survey fielded on Amazon Mechanical Turk, November 2014.

Note: Logistic regressions predicting being currently married. Coefficients presented as odds ratios. Z-scores in parentheses. Estimates using data from the GSS presented for comparison, to demonstrate similar effects for our covariates; marital status coded using POSSLQ variables (MARITAL provides similar results). Models for the Mturk sample include controls for survey condition (not shown); there were no significant differences between the different question orders or sex/gender module placement within the questionnaire.

\* $p < 0.05$ . \*\* $p < 0.01$ . \*\*\* $p < 0.001$  (two-tailed tests).

addition to improving categorical measurements, demonstrates that gender does not affect all women, all men, or all members of any other category in a uniform way. Rather, gender may insinuate itself into social processes along various dimensions. Future studies can use scales to investigate outcomes of conventional and unconventional gender identification while also treating gender identification as the outcome of social processes. To understand how gender inequality is created and maintained, researchers need measures of gender that reflect its complexity and contingency.

## Appendix

**Table A1.** Demographic Comparison between General Social Survey and MTurk samples.

	GSS 2014	MTurk Sample
<i>Gender</i>	—	
Woman	—	52.6
Man	—	46.9
Transgender (direct)	—	0.1
Transgender (indirect)	—	0.2
All other responses		0.2
<i>Sex</i>		
Female	55	53
Male	45	47
Intersex	—	0
<i>Region</i>		
South	36	38
West	24	25
Midwest	23	21
Northeast	17	17
Education (mean years)	13.7	14.6
Age (mean)	49.0	34.4
Married	44	38
Heterosexual or straight	95	90
Hispanic origin	16	7
<i>Self-identified race</i>		
White	70	81
Black or African American	14	7
Specific Asian response	3	5
Selected two or more	7	5
All other responses	6	2
<i>Party affiliation</i>		
Democrat	33	41
Republican	21	17
Independent	43	38
All other responses	2	4
N	2,538	1,522

Sources: General Social Survey (GSS) 2014 cross-sectional sample and authors' survey fielded on Amazon Mechanical Turk, November 2014. Note: Frequencies are reported as percentages unless otherwise noted, and may not sum to 100 due to rounding. Indirect transgender categorization is based on reporting a sex at birth that does not "match" one's current gender. Region is based on self-reported state of residence. Specific Asian responses included categories such as Asian Indian, Chinese, Filipino, and so on.

## Acknowledgments

We are grateful to Shelley Correll, Jocelyn Hollander, Cecilia Ridgeway, Tom Smith, Robb Willer, and Christine Williams for their helpful comments and discussions and to Chrystal Redekopp for her research assistance.

## Funding

This research was supported by the American Sociological Association Fund for the Advancement of the Discipline and the Clayman Institute for Gender Research.

## Notes

1. Survey documentation routinely treats "female" and "male" as gender categories rather than sex categories and conflates the terms *sex* and *gender*. It is therefore frequently unclear whether conventional survey questions are intended to measure sex, gender, or both (Westbrook and Saperstein 2015). We conceptualize sex as a socially constructed system of categorization that divides bodies based on biological criteria, such as genitals and chromosomes, while gender is the set of behaviors generally associated with membership in a sex category (West and Zimmerman 1987). These behaviors are treated as central to one's sense of self, and so most people develop a gender identity. Although normatively encouraged to do so, people's gender identities do not always "match" the sex they were assigned at birth, so that, for example, people labeled female at birth can identify as women, men, genderqueer, gender fluid, non-binary, and so on. We focus our recommendations on improving the measurement of gender in particular. The scales we propose measure a person's public gender identification at a specific point in time, as communicated through the survey (cf. Harris and Sim 2002).
2. Our use of *transgender* includes any person who identifies as a sex and/or gender other than the ones they were labeled at birth.
3. Feminist scholars have long critiqued survey research as emblematic of masculinist, dualistic, and positivist science (Du Bois 1983; for a review, see Stacey 1988). More recently, queer methodologists have contended that surveys not only reify sex, gender, and sexuality categories but also expose people outside of dominant categories to sanctions for their perceived transgressive behavior (Currah and Stryker 2015; Labuski and Keo-Meier 2015; Spade 2015).
4. We included both first- and third-order scales to investigate variation between gendered self-conceptions and reflected appraisals, expecting that large differences might be associated with negative social outcomes. In the second survey, 19 percent of respondents reported distinct first- and third-order responses on both the femininity and masculinity scales, while 29 percent reported distinct responses with regards to either femininity or masculinity. These differences warrant further research and highlight the utility of using scales to represent gender's multidimensionality.
5. The first survey randomly assigned question order, answer order, and answer options. Half of respondents saw conventional sex and gender categories (male/female, man/woman) while the other half also saw intersex, transgender, and "a gender not listed here (please specify)." The second survey



- randomly assigned respondents to four conditions varying question order (scales first or last in the module) and the placement of the sex and gender questions within the questionnaire (in the middle or at the end). The first survey confirmed that respondents understood and responded to questions with unconventional sex and gender answer options, so all respondents to the second survey saw nonstandard answer options. All findings presented are consistent across survey conditions.
6. Between the two studies, we changed the labels used to mark the poles of the 7-point femininity and masculinity scales. On the first survey, the poles were labeled *not at all* and *extremely*, with points in between numbered 1 to 5. Due to concerns that identifying as “extremely” feminine or masculine could carry negative connotations, the second survey labeled the high end *very*. Arguably, the term *completely* would have been a more symmetrical label for the high end than either *extremely* or *very*, but we sought to avoid suggesting that anyone who does not report a 6 on the femininity or masculinity scale is somehow “incomplete.” In the second survey, with the high end labeled *very*, more respondents selected the highest response category on their sex-typical scale (i.e., feminine for females, masculine for males), likely due to the change in wording. All other response patterns were consistent between the surveys (results available on request).
  7. MTurk workers select “Human Intelligence Tasks” (HITs) to complete and then submit completed HITs to the “requesters” who posted them. Requesters are able to approve or reject a worker’s submission depending on whether it meets the assignment’s criteria. Performance ratings report the percentage of HITs a worker has submitted that were approved by requesters. HIT approval ratings have been shown to be effective predictors of data quality (Peer, Vosgerau, and Acquisti 2014), with higher performance ratings related to increased attention and response validity. Workers recruited for our first survey were required to have a minimum HIT approval rating of 80 percent. We compared segments of the sample with minimum HIT approval ratings of 80 percent, 90 percent, and 95 percent and found no differences in response patterns. Workers recruited to the second survey had a minimum HIT approval rating of 90 percent. In addition to using HIT approval ratings to ensure data quality, we dropped observations from respondents who completed the surveys in under 25 seconds and 4.5 minutes, respectively (representing the fastest 1 percent of responses). However, our findings do not change when these responses are included.
  8. Cisgender people identify with the same sex and gender categories to which they were assigned at birth (Schilt and Westbrook 2009).
  9. For example, 99.5 percent of the sample in the second survey reported a sex at birth and current gender that “match”—that is, female/woman and male/man—and are therefore classified as cisgender. It was not possible for respondents to directly identify as “cisgender” unless they were to write this term into the open response field (which none did). Using categorical responses, the gender composition of the second survey sample is: 52.6 percent cisgender women (N = 800), 46.9 percent cisgender men (N = 714), and 0.5 percent transgender (N = 3 male/woman, N = 2 female/transgender, N = 1 female/“Agender,” N = 1 female/“Gender-fluid,” N = 1 female/“I’m not comfortable disclosing assigned gender . . .”). The proportion of transgender respondents in our sample is in line with previous estimates (Flores et al. 2016; Gates 2011; Ponce et al. 2016).
  10. Respondents to the second survey were asked to self-report their home state and zip code, which we could check against the region implied by their IP address. Differences between the two methods of assigning region were minor and did not affect our substantive results.
  11. Typographical errors in open-ended responses were corrected to improve readability. Otherwise, responses appear as they were entered.
  12. We allow respondents to determine which criteria contribute to their femininity and masculinity because our scales are intended to provide a general measure of gender self-identification. This measure is not ideal for all research purposes; researchers who seek to measure a single dimension of gender identification, such as appearance, could adapt the scale introduction to suit their needs.

## References

- Auster, Carol, and Susan Ohm. 2000. “Masculinity and Femininity in Contemporary American Society.” *Sex Roles* 43(7/8): 499–528.
- Balarajan, Meera, Michelle Gray, and Martin Mitchell. 2011. *Monitoring Equality: Developing a Gender Identity Question*. London: Equality and Human Rights Commission.
- Bem, Sandra L. 1974. “The Measurement of Psychological Androgyny.” *Journal of Consulting and Clinical Psychology* 42(2):155–62.
- Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. “Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk.” *Political Analysis* 20:351–68.
- Burke, Peter J., and Alicia D. Cast. 1997. “Stability and Change in the Gender Identities of Newly Married Couples.” *Social Psychology Quarterly* 60:277–90.
- Burke, Peter J., Jan E. Stets, and Maureen A. Pirog-Good. 1988. “Gender Identity, Self-esteem, and Physical and Sexual Abuse in Dating Relationships.” *Social Psychology Quarterly* 51:272–85.
- Butler, Judith. 1993. *Bodies That Matter: On the Discursive Limits of Sex*. New York: Routledge.
- Connell, R. W. [1995]2005. *Masculinities*. 2nd ed. Berkeley, CA: University of California Press.
- Constantinople, Anne. 1973. “Masculinity-femininity: An Exception to the Famous Dictum?” *Psychological Bulletin* 80(5):389–407.
- Courtenay, Will H. 2000. “Constructions of Masculinity and Their Influence on Men’s Well-being: A Theory of Gender and Health.” *Social Science & Medicine* 50(10):1385–401.
- Currah, Paisley, and Susan Stryker. 2015. “Introduction.” *Transgender Studies Quarterly* 2(1):1–12.
- Deutsch, Madeline B., Jamison Green, JoAnne Keatley, Gal Mayer, Jennifer Hastings, Alexandra M. Hall, and Rebecca Allison. 2013. “Electronic Medical Records and the Transgender Patient: Recommendations from the World Professional Association for Transgender Health EMR Working Group.” *Journal of the American Medical Informatics Association* 20(4):700–03.
- Du Bois, Barbara. 1983. “Passionate Scholarship: Notes on Values, Knowing and Method in Feminist Social Science.” Pp. 105–17 in *Theories of Women’s Studies*, edited by G. Bowles, R. Duelli Klein and P. Kegan. London: Routledge.

- Eagly, Alice H., and Linda L. Carli. 2007. *Through the Labyrinth: The Truth about How Women Become Leaders*. Cambridge, MA: Harvard Business School Press.
- England, Paula. 2010. "The Gender Revolution: Uneven and Stalled." *Gender & Society* 24(2):148–66.
- Flores, Andrew R., Jody L. Herman, Gary J. Gates, and Taylor N. T. Brown. 2016. *How Many Adults Identify as Transgender in the United States?* Los Angeles, CA: The Williams Institute.
- Galinsky, Adam D., Erika V. Hall, and Amy J. C. Cuddy. 2013. "Gendered Races: Implications for Interracial Marriage, Leadership Selection, and Athletic Participation." *Psychological Science* 24(4):498–506.
- Gates, Gary J. 2011. *How Many People Are Lesbian, Gay, Bisexual and Transgender?* Los Angeles, CA: The Williams Institute.
- GenIUSS Group. 2014. *Best Practices for Asking Questions to Identify Transgender and Other Gender Minority Respondents on Population-based Surveys*. Los Angeles, CA: Williams Institute.
- Gill, Sandra, Jean Stockard, Miriam Johnson, and Suzanne Williams. 1987. "Measuring Gender Differences: The Expressive Dimension and Critique of Androgyny Scales." *Sex Roles* 17(7):375–400.
- Gough, Harrison G. 1952. "Identifying Psychological Femininity." *Educational and Psychological Measurement* 12:427–39.
- Hammarström, Anne, and Ellen Annandale. 2012. "A Conceptual Muddle: An Empirical Analysis of the Use of 'Sex' and 'Gender' in 'Gender-specific Medicine' Journals." *PLoS One* 7(4):e34193.
- Harnois, Catherine. 2013. *Feminist Measures in Survey Research*. Thousand Oaks, CA: Sage.
- Harris, Allen C. 1994. "Ethnicity as a Determinant of Sex Role Identity: A Replication Study of Item Selection for the Bem Sex Role Inventory." *Sex Roles* 31(3/4):241–73.
- Harris, David R., and Jeremiah Joseph Sim. 2002. "Who Is Multiracial? Assessing the Complexity of Lived Race." *American Sociological Review* 67(4):614–27.
- Harrison, Jack, Jaime Grant, and Jody L. Herman. 2011. "A Gender Not Listed Here: Genderqueers, Gender Rebels and OtherWise in the National Transgender Discrimination Survey." *LGBTQ Policy Journal* 2:13–24.
- Harrison-Quintana, Jack, Jaime M. Grant, and Ignacio G. Rivera. 2015. "Boxes of Our Own Creation: A Trans Data Wo/Manifesto." *Transgender Studies Quarterly* 2(1):166–74.
- Hoffman, Rose Marie, and L. DiAnne Borders. 2001. "Twenty-five Years after the Bem Sex-role Inventory: A Reassessment and New Issues Regarding Classification Variability." *Measuring and Evaluation in Counseling and Development* 34(1):39–55.
- Human Rights Watch. 2011. *Controlling Bodies, Denying Identities: Human Rights Violations against Trans People in the Netherlands*. New York: Human Rights Watch.
- Hunt, Kate, Heather Lewars, Carol Emslie, and G. David Batty. 2007. "Decreased Risk of Death from Coronary Heart Disease amongst Men with Higher 'Femininity' Scores: A General Population Cohort Study." *International Journal of Epidemiology* 36(3):612–20.
- Igo, Sarah. 2007. *The Averaged American*. Cambridge, MA: Harvard University Press.
- Ingraham, Natalie, Vanessa Pratt, and Nick Gorton. 2015. "Counting Trans\* Patients: A Community Health Center Case Study." *Transgender Studies Quarterly* 2(1):136–47.
- Institute of Medicine. 2011. *The Health of Lesbian, Gay, Bisexual, and Transgender People*. Washington, DC: National Academies Press.
- Jost, John T., Christopher M. Federico, and Jaime L. Napier. 2009. "Political Ideology: Its Structure, Functions, and Elective Affinities." *Annual Review of Psychology* 60:307–37.
- Labuski, Christine, and Colton Keo-Meier. 2015. "The (Mis) Measure of Trans." *Transgender Studies Quarterly* 2(1):13–33.
- Lorber, Judith. 2006. "Shifting Paradigms and Challenging Categories." *Social Problems* 53(4):448–53.
- McLaughlin, Heather, Christopher Uggen, and Amy Blackstone. 2012. "Sexual Harassment, Workplace Authority, and the Paradox of Power." *American Sociological Review* 77(4):625–47.
- Oakley, Ann. 1998. "Gender, Methodology and People's Ways of Knowing: Some Problems with Feminism and the Paradigm Debate in Social Science." *Sociology* 32(4):707–31.
- Paolacci, Gabriele, and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23(3):184–88.
- Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti. 2014. "Reputation as a Sufficient Condition for Data Quality of Amazon Mechanical Turk." *Behavioral Research Methods* 46(4):1023–31.
- Permanyer, Iñaki. 2010. "The Measurement of Multidimensional Gender Inequality: Continuing the Debate." *Social Indicators Research* 95(2):181–98.
- Ponce, Ninez, Matt Jans, Gary J. Gates, Bianca Wilson, Jody Herman, and David Grant. 2016. "Putting the 'T' in LGBT: Testing and Fielding Questions to Identify Transgender People in the California Health Interview Survey." Paper presented at the Population Association of America Annual Meeting, Washington, DC, April 1.
- Presser, Harriet. 1998. "Decapitating the U.S. Census Bureau's 'Head of Household.'" *Feminist Economics* 4:145–58.
- Ridgeway, Cecilia. 2011. *Framed by Gender: How Gender Inequality Persists in the Modern World*. New York: Oxford University Press.
- Risman, Barbara. 2004. "Gender as Social Structure: Theory Wrestling with Activism." *Gender & Society* 18(4):429–50.
- Schilt, Kristen, and Jennifer Bratter. 2015. "From Multiracial to Transgender? Assessing Attitudes toward Expanding Gender Options on the US Census." *Transgender Studies Quarterly* 2(1):77–100.
- Schilt, Kristen, and Laurel Westbrook. 2009. "Doing Gender, Doing Heteronormativity: 'Gender Normals,' Transgender People, and the Social Maintenance of Heterosexuality." *Gender & Society* 23(4):440–64.
- Singer, T. Benjamin. 2015. "The Profusion of Things: The 'Transgender Matrix' and Demographic Imaginaries in US Public Health." *Transgender Studies Quarterly* 2(1):58–76.
- Spade, Dean. 2015. *Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law*. Durham, NC: Duke University Press.
- Spence, Janet. 2011. "Off with the Old, On with the New." *Psychology of Women Quarterly* 35(3):504–09.
- Spence, Janet T., and Camille E. Buckner. 2000. "Instrumental and Expressive Traits, Trait Stereotypes, and Sexist Attitudes: What Do They Signify?" *Psychology of Women Quarterly* 24(1):44–53.

- Sprague, Joey. 2005. *Feminist Methodologies for Critical Researchers*. Lanham, MD: AltaMira Press.
- Stacey, Judith. 1988. "Can There be a Feminist Ethnography?" *Women's Studies International Forum* 11(1):21–27.
- Stacey, Judith, and Barrie Thorne. 1985. "The Missing Feminist Revolution in Sociology." *Social Problems* 32(4):301–16.
- Terman, Lewis Madison, and Catharine Cox Miles. 1936. *Sex and Personality*. New York: McGraw-Hill.
- Waite, Linda J., and Evelyn L. Lehrer. 2003. "The Benefits from Marriage and Religion in the United States: A Comparative Analysis." *Population and Development Review* 29(2): 255–75.
- Weinberg, Jill D., Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-based and Crowdsourced-recruited Sample." *Sociological Science* 1:292–310.
- West, Candace, and Don Zimmerman. 1987. "Doing Gender." *Gender & Society* 1(2):125–51.
- Westbrook, Laurel, and Aliya Saperstein. 2015. "New Categories Are Not Enough: Rethinking the Measurement of Sex and Gender in Social Surveys." *Gender & Society* 29(4):534–60.
- Williams, Christine. 2006. "Still Missing? Comments on the Twentieth Anniversary of 'The Missing Feminist Revolution in Sociology.'" *Social Problems* 53(4):454–58.
- Wylie, Sarah A., Heather L. Corliss, Vanessa Boulanger, Lisa A. Prokop, and S. Bryn Austin. 2010. "Socially Assigned Gender Nonconformity: A Brief Measure for Use in Surveillance and Investigation of Health Disparities." *Sex Roles* 63(3-4): 264–76.

### Author Biographies

**Devon Magliozi** is a PhD candidate in sociology at Stanford University. Her research considers how gender, race, and socioeconomic hierarchies are constructed and contested. Her dissertation investigates how social control strategies contribute to neighborhood inequality.

**Aliya Saperstein** is an assistant professor of sociology at Stanford University. Her research examines how categories of difference, such as race/ethnicity and sex/gender, are operationalized in survey research and the consequences of these methodological decisions for studies of stratification.

**Laurel Westbrook** is an associate professor of sociology at Grand Valley State University. Her research focuses on the inner workings of the sex/gender/sexuality system, including how knowledge production, gendered violence, and social movements both maintain and alter the system.