2020

Male/Female Is Not Enough: Adding Measures of Masculinity and Femininity to General Population Surveys

Jolene Smyth
University of Nebraska-Lincoln, jsmyth2@unl.edu

Kristen Olson
University of Nebraska-Lincoln, kolson5@unl.edu

Male/Female Is Not Enough: Adding Measures of Masculinity and Femininity to General Population Surveys

Jolene D. Smyth and Kristen Olson

University of Nebraska-Lincoln, NE, Lincoln, USA

Correspondence — J. D. Smyth
emails — jsmyth2@unl.edu; kolson5@unl.edu

Introduction

Survey research and sociological theory each provide insights into how and why people and groups act, think, and feel. Sociological theories identify what concepts are important for understanding and representing the social world. That is, sociological theories inform what to measure in surveys, and, to a certain extent, how to measure it. Survey research permits sociologists to carefully specify what is to be measured *vis a vis* sociological theory, setting surveys apart as a social research tool. It is this level of specification of concepts and measures that allow surveys to provide continued value at a time when "big data" proliferate. High quality survey measurement and estimation is necessary for sociologists to evaluate sociological theory among

Published as Chapter 11 in Philip S. Brenner (ed.), *Understanding Survey Methodology: Sociological Theory and Applications,* Frontiers in Sociology and Social Research 4, (2020), pp 247-275.
doi:10.1007/978-3-030-47256-6_11

generalizable samples with well-developed questions, leading to further refinement and improvement of the theory and improved understanding of the social world. High quality surveys also provide insights into where sociological theories fail and where they must be adjusted for different subgroups, as well as basic insights into the prevalence of outcomes of interest. Together, sociological theory and survey methods produce insights about society that can inform decision-making and social policy.

This mutually reinforcing relationship between sociological theory and survey methods requires sociological theory to evolve from insights obtained using survey methods and survey measurement to evolve with advances in in sociological theory. The measurement of sex and gender in surveys is one area where the development of survey measures has not kept pace with sociological theory and empirical, largely qualitative, findings. Contemporary gender theory sees sex and gender as separate concepts, both of which are important for understanding behaviors and outcomes. Yet, virtually all contemporary surveys measure sex as a binary "male" versus "female" categorization and fail to measure gender, ignoring important heterogeneity in gender identification that may exist within sex categories and any overlap that may occur across categories.

Both gender scholars and survey researchers are potentially affected by this shortcoming of modern survey measurement. Gender scholars lose an important tool for assessing gender theories, especially on generalizable samples, risking conclusions that are specific to a small group of individuals rather than the population at large. Survey researchers risk producing theoretically obsolete data, limiting the utility of the data or potentially generating misleading conclusions. Survey data that fail to capture and reflect modern and complex understandings of our social realities also face increased risk of being replaced by "big data" such as administrative and social media data. Survey data that do reflect modern and complex understandings can bring value not available in administrative or other data and are therefore unlikely to be replaced.

This paper is part of a growing chorus advocating for updates to how modern surveys measure sex and gender. We argue that the reliance on a single binary measure of sex (male or female) is out of step with current sociological understandings of sex and gender. In re-

sponse, we propose and test a new theoretically-informed gradational measure of gender identification in a nationally representative mail survey. We evaluate whether respondents answer the gender measure and examine the reliability and predictive validity of the measure. In particular, we examine whether measuring gender gradationally adds explanatory value beyond sex on important social outcomes such as sexuality, childcare, grocery shopping, housework, working for pay, and military service. We also examine whether sex moderates the effect of gender identification in the ways that sociological theory would suggest on these outcomes.

Background

Sociologists have pointed out that many people understand and thus organize their social worlds around the "sex and gender binary," or the belief that there are only two types of people, male-bodied masculine and female-bodied feminine (Lorber 1996; Wade and Ferree 2019). In this popular view, sex and gender are conflated. Men are masculine and women are feminine. In contrast, modern Sociologists understand sex and gender as separate social phenomena (Lorber 1996; Lucal 1999). Sex generally refers to biological sex, which is commonly determined during the social processes of sex categorization (i.e., usually determined at birth by a medical professional using socially-derived criteria) (Kessler 1990; Kessler and McKenna 1978).[1] Gender is separate from sex and exists at multiple societal levels, including as deep-seated ideologies that structure institutions and social lives at the macro level (Acker 1990, 1992; Britton 1997; Connell 1987; Hall 1993; Price 2008; Risman 2004) and as identities at the individual level and expressions of those identities at the interactional level (Ridgeway and Smith-Lovin 1999; West and Zimmerman 1987). Social behaviors are enabled and/or constrained by macro-level gender ideologies and structures and reflect micro-level identities (Burke 1991; Ridgeway and Smith-Lovin 1999). Because sex and gender are separate, sociologists understand that people of any sex can have masculine and/or feminine gender identification (Connell 1995; Lucal 1999), that gender identification can vary within sex categories (i.e., some men [women] may feel more masculine or feminine than other men [women]—Con-

nell 1995; Geist et al. 2017), and that there can be overlap in gender identification between the sexes (i.e., some women [men] may identify as just as masculine [feminine] as some men [women]—Magliozzi et al. 2016). Further, sociologists believe that both sex and gender relate in complex and often interacting ways to affect important social, economic, and health outcomes (Annandale and Hunt 1990; Geist et al. 2017; Hyde 2005; Saugeres 2002).

The gender binary is reflected strongly in the nearly ubiquitous survey practice of asking respondents or interviewers (e.g., the General Social Survey) to report respondent sex using only two categories, male or female, as a single demographic measure of sex/gender. The binary sex measure has allowed researchers to identify differences between men and women in types of paid and unpaid labor (Bianchi et al. 2000; Padavic and Reskin 2002); pay rates (Padavic and Reskin 2002); propensities and pathways to commit different types of crimes (Kruttschnitt 2013); health behaviors and outcomes (Verbrugge 1985); and in many other important domains, but fails to reflect current, more complex sociological understandings of sex and gender. Binary sex measures conflate sex and gender, as illustrated by the common interchange of the terms "sex" and "gender" in these questions, and obscure the variation in gender within and across sex categories that sociologists find of central importance. As a result, sociologists have increasingly critiqued survey research's heavy reliance on binary measures, calling instead for the addition of non-binary categories (e.g., "transgender"), measurement of both sex at birth and current sex to better reflect the sociological understanding that sex can change over the life course (Federal Committee on Statistical Methodology 2016a, b, c; Fraser 2018; The GenIUSS Group 2014), and measurement of individuals' gender identity and expression separate from sex (Magliozzi et al. 2016; Westbrook and Saperstein 2015; Geist and Ruppanner 2018; Geist et al. 2017).[2]

Psychologists have developed a handful of gender identification measures, most measuring femininity and masculinity, but these have

1 Work on the case management of intersex babies reveals the extent to which sex categorization is a social process (Epstein 1990, cited in Lorber 1996; Kessler 1990).

2 "Gender identity" is sometimes used to refer to cisgender versus transgender (i.e., one's sense of oneself of male or female regardless of sex), but in this paper we use it to refer to self-perceived masculinity/femininity.

not been widely adopted by population-based surveys because they are impractical (containing anywhere from 24 to 144 items [Bem 1974; Mahalik et al. 2005; Spence et al. 1974] or multiple vignettes [Kroska 2000]), rely heavily on stereotypically masculine or feminine traits that change over time (e.g., Bem 1974; Egen and Perry 2001; Mahalik et al. 2005; Spence et al. 1974), or have been developed using convenience samples of limited subpopulations (e.g., adolescent males, Oransky and Fischer 2009). These measures are costly for inclusion in a wide range of surveys, and may not be suitable for contemporary general adult populations. As a result, survey-based empirical gender literature, with its reliance on binary sex measures, has focused almost entirely on cisgender individuals, failing to capture gender diversity and its consequences (Geist and Ruppanner 2018) and undermining the use of surveys as a tool to study this fundamental organizing feature of society.

However, if measures that capture gender diversity can be developed and deployed alongside measures of sex, surveys should be reasonable tools for examining both sex and gender variation and their social correlates. This point has been made by the Federal Committee on Statistical Methodology, which has joined sociologists in pushing for more inclusive measures of sexual orientation and gender identification in surveys (FCSM 2016a, b, c). In a review of existing measures, only two studies (one an unpublished report) examined continuous measurements of masculinity and femininity (Correll et al. 2014 as cited in FCSM 2016b; Magliozzi et al. 2016), both of which use separate seven-point unipolar scales. The Magliozzi et al. (2016) measure asked respondents to rate how feminine and masculine they see themselves on unipolar endpoint-labeled scales ("not at all" to "very"). Consistent with sociological theory, they find considerable variability in gender identification within men and women, overlap in gender identification between men and women, and an association between gender identification and marital status such that higher gender polarization (i.e., high femininity and low masculinity or vice versa) is associated with being married. The Magliozzi et al. (2016) scales are practical from a survey standpoint because they take up far less space and respondent effort than prior multi-item or multi-vignette masculinity/femininity scales, making it more affordable to measure gender identification in surveys on a wide variety of topics and when respon-

dent burden is a concern. However, these measures were evaluated on an unrepresentative, convenience sample (Amazon Mechanical Turk) using only one predictive validity outcome (marital status). In addition, Magliozzi and colleagues' gender polarization operationalization failed to reveal how each gender identification is directly associated with marital status, whether this association is moderated by sex, how gender non-conforming polarizations are related to marital status, and they did not examine theoretically informed interactions between sex and gender.

In addition to these studies, Smyth (2007) and Smyth et al. (2018) introduced a gender self-perception measurement in which respondents placed marks representing themselves on a horizontal line labeled "completely feminine" at one end and "completely masculine" at the other. Using a ruler, they measured the number of millimeters from the completely feminine endpoint to the respondents' mark. Using this measure, they showed that gender self-perception is associated with women's involvement in farm and ranch work, with more involved women perceiving themselves as more masculine. However, while capturing the continuum of gender identification, this measure is unlikely to be widely adopted due to labor-intensive data entry. For a gradational gender identification scale to be useful and transportable enough to be widely adopted by population-based surveys, it needs to (1) be parsimonious to administer and process, (2) be a measure respondents are willing and able to answer, (3) exhibit high reliability, and (4) exhibit high validity.

In this paper, we test a new gradational measure of gender identification in which respondents are asked to report how masculine or feminine they are on a 21-point scale labeled "Completely Feminine" at one end and "Completely Masculine" at the other. By virtue of being only one item with explicit response options (i.e., no ruler needed), our measure of gender identification meets criteria #1 above. We assess the measure on the remaining three criteria.

We assess whether it meets criteria #2 by examining item nonresponse, which is a commonly used tool to evaluate survey item quality (e.g., Beatty and Herrmann 2002; Krosnick 2002). Item nonresponse rates that are higher than other commonly-asked items (in this case, sex) indicate respondent difficulty with the item while similar or lower item nonresponse rates indicate no such difficulty. An additional

desired outcome is that item missingness is not related to any of the variables that are of interest in the survey—that is, that nonresponse is missing completely at random (e.g., Little and Rubin 2002). Empirically, older respondents and respondents with lower levels of education often have higher item nonresponse rates than their younger and more educated counterparts (see review in de Leeuw et al. 2003). Other demographic characteristics of respondents (i.e., sex and race) are less consistently related to item missing data rates.

Criteria #3, reliability, will be assessed in two ways. First, we will examine the association between gender identification and other demographic variables, the most important of which is sex. Given that most of the U.S. population is cisgender, we expect considerable overlap between these two measures. However, we do not expect complete overlap because gender diversity within sex categories (Connell 1995; Geist et al. 2017) is what we are trying to capture with this measure. In addition to sex, we also examine other common demographics, following Magliozzi, et al. As a second assessment of reliability, we test whether responses to the gender identification item are influenced by questionnaire context. It is well established that the context of survey items can affect responses to these items by influencing how respondents understand questions, what information they use in responding, and how they incorporate that information into their responses (Tourangeau and Rasinski 1988; Sudman et al. 1996; Tourangeau et al. 2000). For example, context effects can occur when information from surrounding questions triggers social comparisons in the domain of interest (Schwarz and Strack 1999). To the extent that individuals have a well-formed gender identification, the context of surrounding questions will not change their gender identification ratings, indicating high reliability. However, asking about society's ideal man or woman before asking about one's self-placement may trigger important comparisons between themselves and this ideal, leading to different answers (e.g., "I don't meet this ideal, so I am going to answer differently from my answer there") and indicating lower reliability.

We use a series of predictive validity assessments based on sociological gender theory to examine criteria #4. The theory of "doing gender" says that individuals produce gender through everyday interactions with others (West and Zimmerman 1987). Individuals are expected to - and expect others to - act in accordance with macro-level

gender ideologies. When individuals act as expected by these societal norms, they are rewarded (i.e., socially accepted, complimented, etc.). When an individual's behavior challenges gender norms, they are held accountable (i.e., judged, devalued, and/or treated negatively) (Lucal 1999). Interactional behaviors that support gender norms or hold others accountable for doing so reproduce macro gender ideologies and sustain existing gender identities.

We may see this distinction between reward and accountability when examining sexuality. Individuals deploy gender displays strategically to try to manage their social experiences related to sexuality. Appearing straight (independent from one's sexuality) requires doing gender in a way that closely aligns with one's sex. Likewise, to be visible as a sexual minority (e.g., when dating or to challenge social norms) requires doing gender non-normatively (Wade and Ferree 2019; West and Zimmerman 1987; Frye 1983 as cited in West and Zimmerman 1987, p. 145). We do not expect the likelihood of being heterosexual or GLB to differ by sex alone or by gender identification alone, but because doing sexuality requires masculine or feminine gender presentation *relative to a sex category*, we do expect the relationship between gender identification and sexuality to differ for men and women. In particular, while there is undoubtedly variation in gender identities within sexuality categories, overall, we expect women who identify as more masculine to be more likely to self-identify as lesbian or bisexual and less likely to self-identify as straight and men who identify as more feminine to be more likely to self-identify as gay or bisexual.

Beyond this important social identity, a number of studies have uncovered interactional behaviors through which people commonly produce gender such as through the household division of labor (Berk 1985; South and Spitze 1994). One explanation for the fact that women do more housework on average than men (Bianchi et al. 2000; Bianchi et al. 2012) is that housework is a means for doing gender. Doing cooking, cleaning, and laundry is doing femininity for women (Berk 1985). Likewise, doing household repairs, mowing the lawn, and grilling meat are means for producing masculinity for men (Berk 1985; Sobal 2005). Even avoiding opposite-gendered tasks can be a way of doing gender, as is the case for U.S. men who contribute less to housework to bolster their masculinity when they lose their status as pri-

mary provider (Bittman et al. 2003; Brines 1994; Greenstein 2000). Consistent with much previous research, we expect sex to predict who sees themselves as the primary person in the household who does different types of tasks (housekeeping, household repairs, etc.) and how many hours they spend on tasks, but we also expect an association between these outcomes and gender identification. Moreover, we expect the association between gender identification and these outcomes to be moderated by sex (i.e., femininity will be associated with the likelihood claiming to be the primary housekeeper differently for women than men).

The division of labor in childcare is also a means through which men and women do gender (Dalton and Bielby 2000; Hays 1996; McMahon 1995; Walzer 1998). In trying to achieve ideological gendered standards of parenthood (or simply to function as a parent in a world that expects them), men and women parent in gendered ways, even when they desire otherwise (Walzer 1998). Women mother and men father (nobody is simply a parent), leading them to reproduce gendered parenting ideology and influence their own identities as parents and as women or men. Thus, we expect that women will be more likely to say that they are the primary person involved in childcare and spend more hours on care work than men. Gender identification may also be related to these tasks in that individuals who perceive themselves as more feminine may be more likely to engage in this kind of care work. Alternatively, engaging in care work may lead all respondents to perceive themselves as more feminine. While we cannot disentangle causal order on this issue, we do expect an association between gender identification and care work and we expect it to differ for men and women.

Paid employment also provides an arena for doing gender. To the extent that men are more likely to work for pay (Bureau of Labor Statistics 2018) and to hold fulltime work (Bureau of Labor Statistics 2014), they are doing masculine gender (Bittman et al. 2003). Thus, we expect men and those who report more masculine gender identification to report more hours working for pay than women and those reporting feminine gender identifications. Performing sex-typed work (i.e., driving a truck or teaching school) is another means for doing gender (England 1992; Padavic and Reskin 2002). Doing farm work, for example, leads women to be perceived by others and

to perceive themselves as more masculine (Brandth 2006; Smyth et al. 2018). Women in male-typed occupations commonly have to perform the "feminine apologetic" (i.e., put extra emphasis on feminine appearance and behaviors - Felshin 1974) to offset these perceptions. Women in the military have been shown to go to fairly extreme measures to do femininity to offset their masculine military jobs (Herbert 1998). Given that the military is largely a male/masculine domain and associated with masculine work (Enloe 2004), we expect men and those who are more masculine to be more likely to report having ever served.

In sum, based on theory and previous literature, we expect both sex and gender identification to be associated with each of these outcomes, and in some cases we expect sex to moderate the effect of gender identification. To the extent that gender identification explains variation in these outcomes above and beyond sex alone, or provides a more nuanced understanding of the joint roles of sex and gender, then we have evidence that our gender identification scale has predictive validity.

Methods

The data for this paper come from the National Health, Wellbeing, and Perspectives Survey (NHWPS; AAPOR RR1 = 16.7%, n = 1002; AAPOR 2016), a 12-page (77-item) mail survey conducted between April and August 2015. NHWPS was designed to examine mechanisms underlying sex differences in mental and physical wellbeing and to test methodological features of surveys. The design included a fully crossed 3×3×2 experiment with three within-household selection instruction treatments (instruction in the cover letter alone, in the cover letter and questionnaire, in the cover letter and questionnaire with a verification question; Olson and Smyth 2017), three incentive treatments (no incentive, $1 cash at first mailing, and $1 cash at third mailing; Smyth et al. 2019), and two versions of the questionnaire.[3] A simple random sample of 6000 addresses was selected from the USPS Delivery Sequence File by Survey Sampling International, and randomly as-

3 Each had the same questions, but design features within questions differed.

signed to one of the resulting 18 experimental treatments. The next birthday within-household selection procedure was used to sample an adult from each household (Gaziano 2005). Sampled households were contacted by postal mail up to four times (initial invitation, postcard reminder, and two full-packet reminders).

When examining item nonresponse, our analytic data set is n = 1002. Four questionnaires were returned with their identification numbers torn off, making it impossible to know their geographic region (a control variable, described below) or incentive treatment. Thus, when examining the other outcomes, our analytic data set is n = 922 cases with full data on sex and gender identification and intact questionnaire ID numbers. Although our imputation and missing data indicators (described below) ensure that no cases are dropped due to independent variables, we allow casewise deletion for missing data on dependent variables, resulting in some variation in sample size for the predictive validity analyses.

Measures

Independent Variables

Sex was measured with the categories "male" (coded 0) and "female" (coded 1) within a household roster that asked for information for up to six household members. Missing data (8.48%, unweighted estimate) on sex was logically imputed using information on sexuality, partnerships, sex of partners, the gender scale, and household tasks. 1.5% of missing cases could not be imputed. *Gender identification* was measured using a continuum with 21 unnumbered scale points ranging from "Completely Feminine" to "Completely Masculine" (See **Fig. 1)**. Respondents were asked to rate the femininity/masculinity of themselves, society's ideal man, and society's ideal woman and were randomly assigned to receive the items with the "yourself" scale first or last. Gender identification is a continuous measure ranging from 1 to 21 with higher numbers denoting more masculinity and less femininity. No imputation was used for this variable.

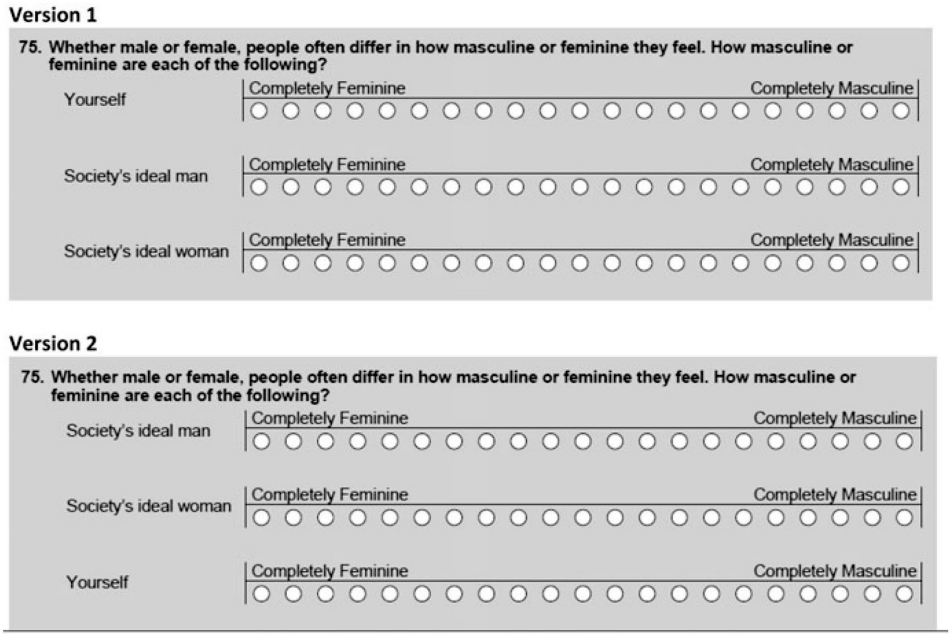

Fig. 1 NHWPS Survey Item Measuring Masculinity/Femininity

Dependent Variables for Predictive Validity Models

Sexuality was measured by an item asking, "Do you consider yourself to be: Heterosexual or straight, Gay or lesbian, Bisexual." An indicator variable coded 1 for those who reported being heterosexual or straight (92.85%) and 0 for those reporting being gay or lesbian or bisexual (GLB; 7.15%) was created (See **Table 1**). Logical imputation using information from the household roster was used to fill in missing data where possible. Remaining missing cases are casewise deleted when sexuality is used as a dependent variable and represented by a missing indicator to avoid the loss of cases when it is used as an independent variable, yielding 91.42% straight, 7.04% GLB, and 1.53% missing.

The next set of dependent variables capture whether respondents see themselves as the primary person in the household to do childcare, grocery shopping, housekeeping, and household repairs. After completing the household roster, respondents were asked, "Thinking about the people you listed in question #48, who is most likely to do each of the following tasks?" For each task, respondents who selected themselves are coded 1 and those who did not are coded 0.

Table 1 Descriptive Statistics for Dependent Variables (n = 922)

	Unweighted Frequency	*Weighted Mean or Percent*	*Standard Deviation*	*Minimum*	*Maximum*
Sexuality					
Straight	870	92.85			
GLB	38	7.15			
Missing	14				
Respondent is most likely person in the household to do. . .					
Childcare					
No	538	67.12			
Yes	274	32.88			
Missing	110				
Grocery shopping					
No	229	32.36			
Yes	666	67.64			
Missing	27				
Housekeeping					
No	232	35.78			
Yes	645	64.22			
Missing	45				
Household repairs					
No	367	44.52			
Yes	510	55.48			
Missing	45				
In a typical week, hours spent on. . .					
Working for pay					
Mean # hours		27.41	22.78	0	168
Missing	97				
Housework					
Mean # hours		9.28	10.56	0	112
Missing	81				
Caring for family					
Mean # hours		13.39	28.59	0	168
Missing	116				
Military service					
No	764	89.11			
Yes	114	10.89			
Missing	44				

When sexuality is used as a dependent variable, cases that remain missing after logical imputation are casewise deleted. When it is used as a control variable, a missing data indicator for these cases is included.

Three continuous variables capture the number of hours spent weekly on care work, housework, and working for pay, as measured by the question, "Thinking about how you spend your time in a typical week, how many hours do you spend on each of the following?"

One version of the questionnaire also included the instruction, "your best estimate is fine," which did not change responses (Timbrook et al. 2016). The items included, "Working for pay at all jobs, including overtime," "On household work, not including childcare and leisure time activities," and "Looking after family members (children, elderly, ill, or disabled family members)." Items on leisure time and sleep are not examined here. Responses to these items are top coded at 168 hours (24 hours × 7 days).

The final dependent variable is an indicator of having ever served in the military (1 = yes, 0 = no). This item was part of a separate experiment to examine full versus quasi-filters (Olson et al. 2018). Version 1 utilized a full filter asking, "Are you a veteran or currently serving in the military?" followed by items asking when the respondent served and if they served in a combat zone. Respondents were coded as having served if they answered affirmatively to the filter question or skipped the filter question but subsequently indicated a service time period or having served in a combat zone. In version 2, there was no filter question; rather the service dates and combat zone questions included the quasi-filter response option "Never served in the military." Respondents were coded as having served in the military if they reported any time period of service or having served in a combat zone, and coded as not having served if they selected "Never served in the military".

Control Variables

To account for other factors that may be associated with the outcomes and reduce the likelihood of spurious associations, we control for age, education, race, ethnicity, sexuality (when not the dependent variable), political affiliation, having dependents under age 18 living in the household, and region in all models. We also control for the experimental design factors. Missing data is accounted for with probabilistic single imputation and missing category indicators for the categorical variables and with group mean imputation for age. Descriptive statistics for these variables are shown in **Table 2.**

Table 2 Descriptive Statistics for Control Variables (n = 922)

	Unweighted Frequency	*Weighted Mean or Percent*	*Standard Deviation*	*Minimum*	*Maximum*
Age					
Mean		49.09	17.65	17.63	99.27
Education					
HS or less	172	34.04			
Some college	302	34.35			
BA+	441	30.91			
Missing	7	0.71			
Race and ethnicity					
Non-Hispanic white	713	64.73			
Hispanic white	31	7.41			
Non-Hispanic black	65	11.21			
Non-Hispanic other	63	8.70			
Hispanic other	16	4.92			
Missing	34	3.03			
Sexuality					
GLB	38	7.04			
Straight	870	91.42			
Missing	14	1.53			
Political party affiliation					
Democrat	318	35.72			
Independent	302	34.56			
Republican	259	24.94			
Missing	43	4.77			
Region					
South	316	40.02			
Northeast	175	17.31			
Midwest	246	22.62			
West	185	20.05			
Dependents under age 18					
No	650	62.10			
Yes	207	31.83			
Missing	65	6.06			
Experimental treatments					
Questionnaire version					
Version 1	485	52.16			
Version 2	437	47.84			
Within household selection					
Instruction in letter only	332	37.47			
Inst. In letter & questionnaire	301	30.52			
Inst. In Letter & Questionnaire w/ verification question	289	32.01			
Incentives					
No incentive	242	23.61			
$1 with first mailing	358	38.54			
$1 with third mailing	322	37.85			

N = 922 cases where both sex and gender had values and all experimental treatments were known

A continuous measure of age was calculated using reports of respondents' date of birth. Education was measured by a nominal item asking for highest degree. Indicator variables were created for high school or less, some college, a four-year degree or more (BA+), and missing data.

Race was measured with a check-all-that-apply question asking, "What is your race? White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Other." Ethnicity was measured by an item asking, "Are you Spanish, Hispanic, or Latino?". The race and ethnicity measures were combined to produce a set of five indicator variables for combinations of race and ethnicity (non-Hispanic white, Hispanic white, non-Hispanic black, non-Hispanic other, and Hispanic-other) plus a final missing data indicator for remaining missing cases.

Political party affiliation was measured by a question asking, "In politics today, do you consider yourself a ... Republican, Democrat, Independent." Indicator variables for Republican, Democrat, and Independent, and missing, were created.

Respondents were asked how many dependents from five age groups (under 1, 1–5, 6–11, 12–17, and 18 or older) were living with them. Responses were used to generate a dichotomous variable coded 0 for those with no dependents in the first four categories (i.e., under age 18) and 1 for those with dependents in these categories.

Geographic region, obtained from the sample file, is represented by a series of indicator variables for Northeast, Midwest, South, and West based on census regions. Indicator variables were also created to represent the experimental factors to account for any effects of the experimental design.

Analyses

Item nonresponse. We examine the item nonresponse rate for gender identification overall, for men versus women, and compared to that of the sex question using dependent t-tests. We then use logistic regression to evaluate whether certain subgroups (using the control variables described above) are more or less likely to fail to answer the gender identification question.

Reliability. Next, we examine response distributions for sex and gender identification and the response distribution for gender identification by sex category to determine how much variation in gender identification there is within and between sex categories. We determine how much variance in gender identification is shared with sex and how much is unique by regressing (OLS) the gender scale on sex. The resulting R^2 value reflects shared variance with sex, and 1 – R^2 reflects unique gender identification variance. Next, we examine whether the gender identification ratings change over different measurement contexts by comparing responses across the two question orders (yourself reported before or after society's ideals, **Fig. 1**). We look at the effect of this experiment on the mean gender identification ratings overall and separately for men and women using OLS regression. We then use OLS regression to examine whether the demographic variables described above predict gender identification ratings, and how much variation in the gender identification scale is explained by these demographic variables.

Predictive Validity. Finally, we examine the association between each of our dependent variables and sex and gender identification by estimating a series of logistic (for dichotomous outcomes) and negative binomial (for count outcomes with overdispersion) regression models using STATA 15.1. For each dependent variable, we estimate four models: sex alone, gender alone, sex with gender, and an interaction model. All models include the control variables.

All analyses account for the survey design using STATA's *svy* command and are weighted.

Findings

Item Nonresponse

Overall, 8.6% of respondents did not answer the gender identification question; in comparison, 7.6% did not answer the sex question.[4] These dependent proportions are not statistically different from each other ($F(1, 1001) = 0.37$, $p = 0.541$). Men and women were equally likely to

4 Weighted estimates.

answer the gender identification item (men 91%, women 93%, t = –0.88; p = 0.381), and those who skipped the sex question also were more likely to skip the gender identification question (42% of those who were missing on sex answered the gender identification question; t = 2.78, p = 0.006).

Next we examine whether item missing data on the gender identification measure was related to age, education, race/ethnicity, sexual orientation, political ideology, region of the country, having any dependents, and the experimental variation in question order, selection instructions, and incentive (results available from authors on request).[5] Across all of these characteristics, failing to answer the sex, education, political affiliation, and dependents questions were all the strongest predictors of failing to answer the gender identification question ($p < .03$ for all items). This makes sense as these items were all located in the same section of the questionnaire. Substantively, individuals with higher education levels were less likely to omit the item relative to those with a high school degree or less ($F(2,995) = 8.09$, $p < .0001$) and respondents whose race/ethnicity is Hispanic White were more likely to fail to answer this item compared to non-Hispanic White respondents (OR = 4.56, t = 2.72, p = 0.007). Additionally, those who received a questionnaire with the instructions on the front of it were more likely to omit answering this item compared to those who received the instructions in a cover letter (OR = 3.64, t = 2.60, p = 0.009). There was no association between item missing data rates on the gender identification question and age, sexual orientation, political affiliation, region of the country, having dependents, the question order experiment, or incentives.

Reliability

In this study, 46.7% of adults identified as male and 53.3% identified as female. **Figure 2** shows the percent of men and women selecting each point on the gender scale. As expected, the scale skews heavily masculine for men (42.6% chose "completely masculine") and heavily feminine for women (36.8% chose "completely feminine") (Design adjusted $F(16.36, 15099.08) = 20.65$, $p = 0.0000$). The average gen-

5 n = 998. Four cases were excluded because of missing questionnaire ID numbers, making it impossible to know experimental treatment and region (n = 998).

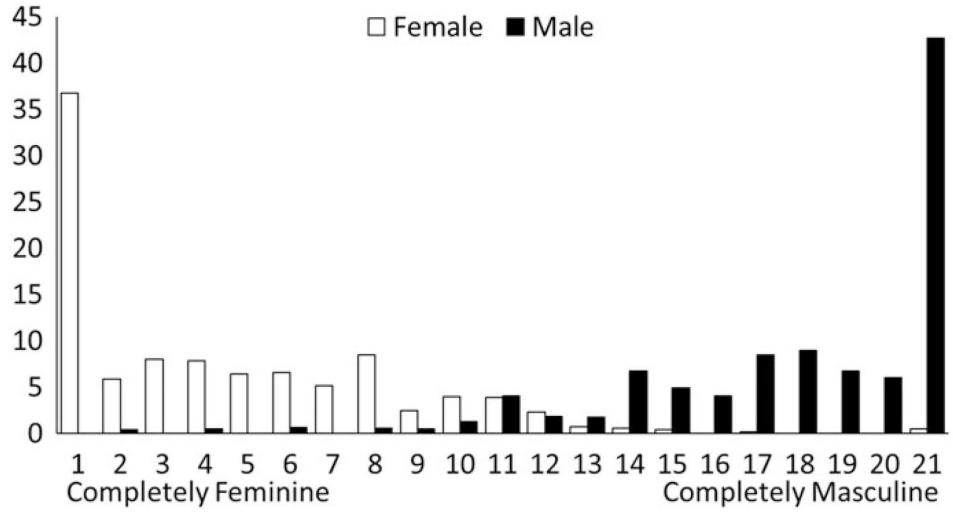

Fig. 2 Percent selecting each point on the gender identification scale by sex

der rating on the scale was 4.5 (SD = 4.05, IQR = 6) for women and 18 (SD = 3.38, IQR = 4) for men. Both sexes used almost the entire range of the gender scale (women 1 to 21; men 2 to 21). Thus, there is considerable range in gender identification within each sex category and considerable overlap between them.

Regressing the gender scale on sex reveals that sex is a significant predictor of gender (t = –36.21, p < 0.000) and explains 76.4% of the variance in the gender identification scale, leaving 23.6% of the variance not shared by sex. Some of this unexplained variation is explained by the experimental variation in question order. Although there is no difference in gender identification by question order overall (t = 1.26, p = 0.207), there are important sex differences (see **Table 3**). Men evaluate their gender identification similarly regardless of whether they evaluate themselves first or after society's ideals (t=–

Table 3 Mean reported gender self-identification by question order and sex of respondent

	Overall	*Men*	*Women*
Self-perception asked first	10.35	18.26	3.31
Society's ideal asked first	11.27	17.69	5.73
Difference in responses (society ideal-self-perception)	0.92	–0.57	2.42
P-value	0.207	0.33	< .0001
N	922	359	563

There are 5 people with a missing value for sex who answered the gender question. There is no difference in reports of gender identification across the questionnaire versions for these 5 people (p = 0.412)

1.00, p = 0.319). Women, on the other hand, evaluate themselves as 2.42 points *more masculine* (t = 5.39, p < .0001) when they are asked to evaluate society's ideal man and woman first versus when they evaluate themselves first. Thus, asking about society's ideals first creates a significant anchoring effect for women but not men. Adding the indicator for the question order experiment and the interaction between sex and the order experiment to the regression of the gender scale on sex explains an additional 1.4% of variance in gender identification, leaving 22.2% of the variance in gender unexplained.

We now examine demographic predictors of gender identification. Notably, none of the demographic predictors (age, education, race/ethnicity, political affiliation, region of the country, presence of dependents, the selection experiment, or the incentive experiment) other than sex are associated with gender identification at the p < .05 level when men and women are included in the same model, likely because masculinity and femininity operate in opposite directions on the scale.

However, there are significant associations in self-perceived gender identification with many of these demographic variables when examining men and women separately (**Table 4**). For instance, older men rate themselves as more masculine and older women rate themselves as more feminine than their younger counterparts. Overall, education is not associated with gender identification for men (F = 0.44, p = 0.722) or women (F = 1.58, p = 0.193). Race/ethnicity is associated with evaluations of gender identification for both men (F = 8.24, p < .0001) and women (F = 4.81, p = 0.0002)—Hispanic white men and women and non-Hispanic black men and women evaluate themselves as more gender normatively polarized than their non-Hispanic white counterparts. Republican men evaluate themselves as more masculine and Republican women evaluate themselves more feminine than their Democrat counterparts (men: F = 9.96, p < .0001; women: F = 2.36, p = 0.070). Gender identification varies by region for men (F = 4.74, p = 0.0027), but not for women (F = 0.14, p = 0.936). Having dependents and the other experimental conditions are not associated with gender identification for either men or women. These results mirror many of the associations between gender polarization and a similar set of demographic variables examined by Magliozzi et al. (2016), indicating that gender identification is socially contingent (p. 5). This collection of demographic variables explains about 29% of the variation in gender identification for both men and women.

Table 4 Linear Regression Coefficients Predicting Gender Identification for Men and Women

	Men		*Women*	
	Coef.	*SE*	*Coef.*	*SE*
Question order experiment				
Self-perception asked first	–		–	
Society's ideal asked first	−0.81+	0.449	2.29***	0.399
Age	0.08***	0.012	−0.05***	0.012
Education				
HS or less	–		–	
Some college	−0.55	0.593	0.61	0.563
BA+	−0.12	0.549	1.06*	0.526
Missing	−0.50	1.859	−0.40	1.819
Race/ethnicity				
Non-Hispanic white	–		–	
Hispanic white	2.76***	0.612	−1.40+	0.753
Non-Hispanic black	1.78*	0.843	−2.82***	0.627
Non-Hispanic other	−1.28	1.160	0.76	0.934
Hispanic other	−2.00	2.864	−1.63*	0.768
Missing race & ethnicity	5.68*	2.376	0.86	1.307
Political affiliation				
Democrat	–		–	
Independent	0.90	0.613	0.06	0.510
Republican	1.15+	0.626	−0.98+	0.550
Missing	−5.98***	1.664	−1.73	1.089
Region				
South	–		–	
Northeast	−2.75***	0.773	−0.23	0.612
Midwest	−0.33	0.622	−0.03	0.510
West	−0.92+	0.536	0.16	0.615
Any dependents				
No	–		–	
Yes	0.49	0.540	−0.39	0.507
Missing	0.85	1.307	−0.55	0.652
Cover experiment				
Letter only	–		–	
Instructions	0.15	0.448	0.49	0.461
Verification question	0.49	0.560	0.32	0.501
Incentive experiment				
No incentive	–		–	
Pre-paid incentive	−0.44	0.631	−0.04	0.510
Incentive with reminder	−0.89	0.667	0.16	0.520
Intercept	14.70***	1.410	5.70***	1.149
N	359		563	
Model F	6.55***		8.07***	
R^2	29.14%		29.04%	

+ p < .10 ; * p < .05 ; ** p < .01 ; *** p < .001

Table 5 Odds Ratios Predicting Heterosexual Sexuality (n = 902)

	Model 1	*Model 2*	*Model 3*	*Model 4*
Female	1.33		1.59	22.70*
Gender identification		0.99	1.02	1.15*
Female * gender identification				0.79*

All models controlled for age, education, race, ethnicity, political affiliation, region, dependents under 18 in the household, and experimental treatments.
+ p ≤ 0.100 ; * p ≤ 0.050 ; ** p ≤ 0.010 ; *** p ≤ 0.001

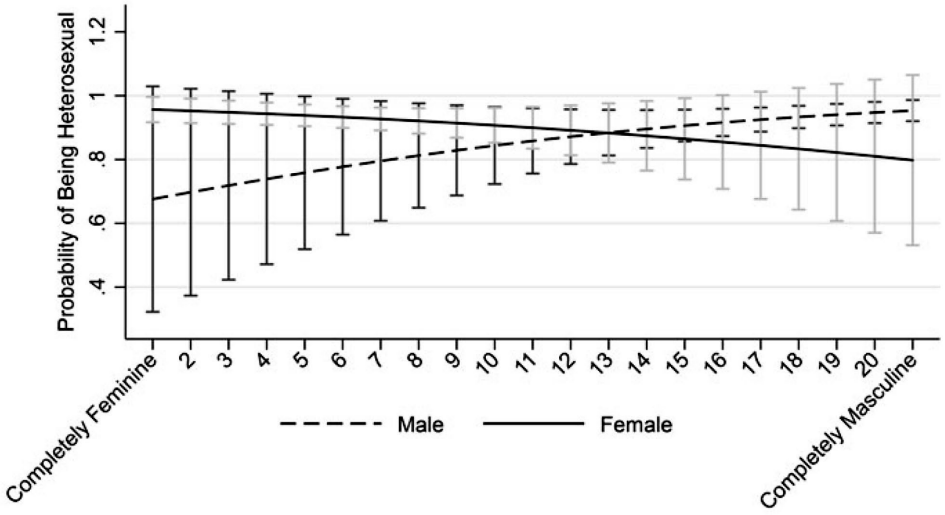

Fig. 3 Predicted Probability of being Heterosexual by Sex and Gender Identification

Predictive Validity

The first set of associations we examine are between sex, gender identification, and sexuality (**Table 5**—full models in online supplement). As expected, neither sex nor gender identification on their own (Models 1 and 2) nor the two of them together (Model 3) are significantly associated with the likelihood of reporting being heterosexual. However, there is a significant interaction effect between sex and gender identification (t = −2.25, p = 0.025), graphed in **Fig. 3**. As men report higher masculinity, the likelihood of them reporting being heterosexual increases, but as women report higher masculinity, the likelihood

Table 6 Odds Ratios Predicting Reporting Self as the Person in the Household Most Likely to Do Tasks

	Model 1	*Model 2*	*Model 3*	*Model 4*
Childcare (n = 812)				
Female	4.78***		3.34*	1.51
Gender identification		0.91***	0.97	0.94
Female * gender identification				1.07
Grocery shopping (n = 895)				
Female	9.04***		5.20***	9.75*
Gender identification		0.87***	0.96	0.98
Female * gender identification				0.94
Housekeeping (n = 865)				
Female	13.80***		3.87**	14.95**
Gender identification		0.83***	0.90**	0.95
Female * gender identification				0.89+
Household repairs (n = 877)				
Female	0.09***		0.33*	1.89
Gender identification		1.19***	1.12***	1.22**
Female * gender identification				0.86+

All models controlled for age, education, race, ethnicity, sexuality, political affiliation, region, dependents under 18 in the household, and experimental treatments.
+ p ≤ 0.100; * p ≤ _ 0.050 ; ** p ≤ 0.010 ; *** p ≤ 0.001

of them reporting being heterosexual decreases.[6] Thus, the effect of gender identification depends on sex.

Table 6 shows the association of sex and gender identification with the likelihood of reporting oneself as the household member most likely to do childcare, grocery shopping, housekeeping, and household repairs. Women are significantly (p = 0.023 and p = 0.001) and substantively (OR from 3.34 to 5.20) more likely to report being the primary person to do childcare and the primary grocery shopper. Gender identification is not associated with either of these reports above and beyond sex (Model 3, childcare t=−0.74, p = 0.458, groceries t=−1.33, p = 0.184), nor is there a significant interaction between gender identification and sex for either outcome (Model 4, childcare t = 0.90, p = 0.369, groceries t = −0.92, p = 0.357).

6 This is not to imply that all sexual minority males are feminine or that all sexual minority females are masculine. Gender identity ratings varied within these groups, from 10 to 21 for sexual minority men and from 1 to 13 for sexual minority women. In this survey, the most feminine men and masculine women were heterosexual, counter stereotypes.

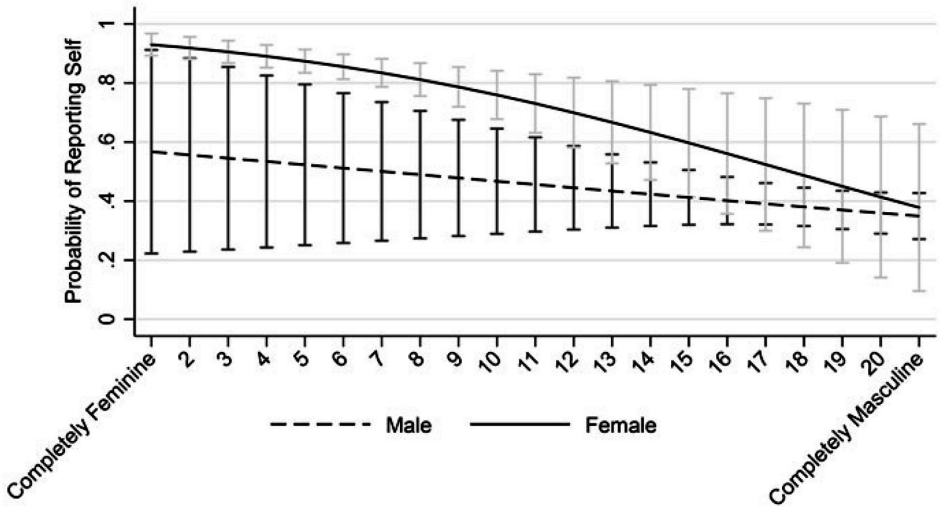

Fig. 4 Predicted Probability of Reporting Oneself as the Person in the Household Most Likely to do Housekeeping

For housekeeping, Model 3 shows that women are more likely than men to report being the primary housekeeper (t = 2.66, p = 0.008) and those who rate themselves as more masculine, net of sex, are less likely to do so (t = −3.13, p = 0.002). Additionally, there is a marginally statistically significant interaction between sex and gender identification (Model 4, t= −1.80, p = 0.073, see **Fig. 4**). Both men and women are less likely to report being the primary person to do housekeeping as they rate themselves more masculine, but the decline is steeper for women and gap between men and women closes as women approach the masculine side of the gender identification scale.

The results for household repairs also indicate that both sex and gender identification matter. Net of gender identification, women are less likely than men to report this status (Model 3, t = −2.26, p = 0.024), and net of sex, those who rate themselves as more masculine are more likely to report this status (Model 3, t = 3.29, p = 0.001). However, as Model 4 shows, there is a moderately statistically significant interaction between sex and gender (t = −1.94, p = 0.052, see **Fig. 5**). On the feminine end of the scale, the probability of claiming to be the primary person to do household repairs is similarly low for males and females. As both sexes rate themselves more masculine, the likelihood of reporting this status increases, but at a faster rate for men than women, resulting in the largest differences between the sexes at the completely masculine endpoint of the scale.

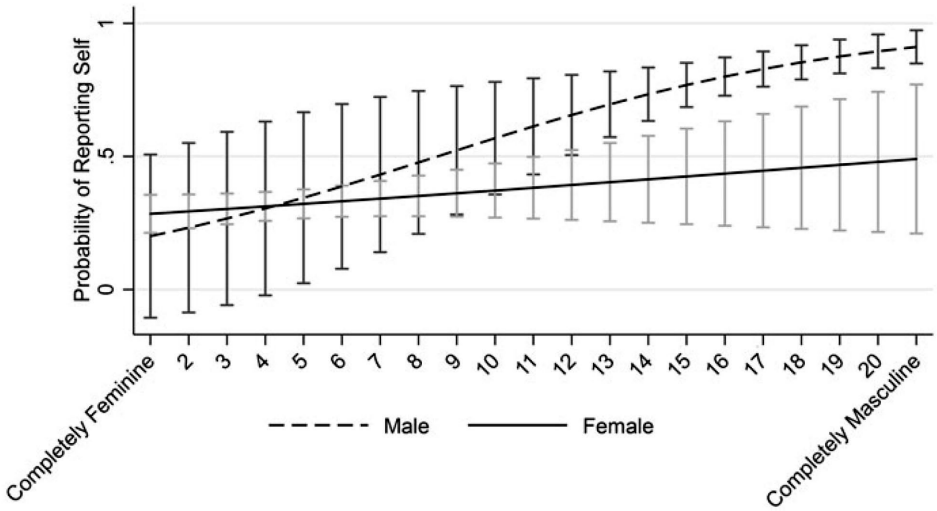

Fig. 5 Predicted Probability of Reporting Oneself as the Person in the Household Most Likely to do Household Repairs

Table 7 shows how sex and gender identification predict hours spent working for pay, on housework, and on family care work. For all three outcomes, sex and gender identification individually are statistically significant (Models 1 and 2). Women report fewer hours working for pay (t = −3.32, p = 0.001) and more hours on housework (t = 4.38,

Table 7 Coefficients from Negative Binomial Regression of Number of Hours Spent on Tasks on Sex and Gender Identification

	Model 1	*Model 2*	*Model 3*	*Model 4*
Hours working for pay (n = 825)				
Female	−0.31***		−0.29	0.18
Gender identification		0.02**	0.00	0.02
Female * gender identification				−0.04
Hours on housework (n = 841)				
Female	0.43***		0.36+	0.68+
Gender identification		−0.03***	−0.01	0.01
Female * gender identification				−0.03
Hours caring for family (n = 806)				
Female	0.91***		0.53	−0.10
Gender identification		−0.06***	−0.03	−0.06
Female * gender identification				0.05

All models controlled for age, education, race, ethnicity, sexuality, political affiliation, region, dependents under 18 in the household, and experimental treatments.
+ p ≤ 0.100 ; * p ≤ 0.050 ; ** p ≤ 0.010 ; *** p ≤ 0.001

Table 8 Odds Ratios Predicting Having Ever Served in the Military (n = 849)

	Model 1	*Model 2*	*Model 3*	*Model 4*
Female	0.14***		1.26	0.40
Gender identification		1.16***	1.17**	1.12
Female * gender identification				1.09

All models controlled for age, education, race, ethnicity, sexuality, political affiliation, region, dependents under 18 in the household, and experimental treatments.
+ p ≤ 0.100 ; * p ≤ 0.050 ; ** p ≤ 0.010 ; *** p ≤ 0.001

p = 0.000) and care work (t = 4.62, p = 0.000) than men. More masculine ratings on the gender identification scale are associated with more hours working for pay (t = 2.95, p = 0.003) and fewer hours on housework (t = −3.73, p = 0.000) and care work (t = −4.40, p = 0.000). However, these effects appear to be due to the overlap (i.e., shared variance) between these two measures. When they are entered into the model simultaneously (Model 3) the effects of sex and gender are virtually eliminated, and there are no statistically significant interactions for any outcome.

Table 8 shows results for military service. Both sex and gender identification, entered individually (Models 1 and 2), are significantly associated with having ever served in the military. Women are less likely to have served than men and increases in masculinity ratings are associated with increased likelihood of having served. However, Model 3 shows that when sex and gender identification are both accounted for, only gender identification remains significant. Net of sex, each additional point toward the masculine end of the gender identification scale increases the odds of having served by 17 percent. There is no significant interaction between sex and gender identification. Thus, apparent sex differences in military service may in fact be gender differences. Including a gender measure in longitudinal studies would help reveal whether more masculine identifying people are more likely to enlist, the experience of military service leads people to identify as more masculine, or perhaps both.

Discussion and Conclusions

While binary sex measures have helped reveal inequalities between men and women in many domains, the heavy reliance on binary measures is out of step with contemporary sociological understanding and theorizing of sex and gender. The lack of good gender measurement, separate from sex, inhibits the ability to understand how this concept shapes peoples' lives, above and beyond sex. To overcome this glaring limitation, measures of gender that can practically be included in population based surveys need to be developed. In this paper, we examine one such measure that asks respondents to rate their femininity/masculinity on a 21-point scale. Our criteria for assessing the measure is that (1) it must be parsimonious, (2) respondents must be able and willing to answer it, (3) it must exhibit measurement reliability, and (4) it must exhibit validity and have explanatory value above and beyond traditional binary measures of sex.

Our short one-item measure meets all of these criteria. Respondents are just as likely to answer it as they are to answer the binary sex measure, and the predictors of item nonresponse are what is seen in other studies examining item nonresponse more generally in mail surveys (e.g., low education). The strongest predictors of item nonresponse on this measure were item nonresponse on other demographic measures, suggesting a general tendency for people to skip demographic items and no specific problems with the gender identification measure. This is noteworthy, given that the gender identification question was the third to last question in a 12-page questionnaire.

The gender identification item also exhibits reasonable reliability. As expected, we found considerable (although not complete) overlap between this measure and the binary sex measure. In addition, many demographic characteristics were significantly associated in the same direction with gender identification here as in Magliozzi et al. (2016). Thus, while the two scales are operationalized differently, they seem to be reliably tapping into the same underlying construct. Our gender identification measure was subject to measurement context effects, particularly for women, who rated themselves more masculine when asked about society's ideal man and woman before themselves. This context effect only accounts for 1.4% of the

variance in the gender identification scale. Nevertheless, future research should examine why the comparison to the ideals affects women's responses but not men's.

The predictive validity analyses illustrate the contributions of both sex and gender to many important outcomes. For example, the results showed that gender identification has different effects for men and women on the probability of reporting being heterosexual versus a sexual minority. Consistent with expectations, higher ratings of masculinity were associated with increased likelihood of men but decreased likelihood of women identifying as heterosexual. This is consistent with the theory that people "do heterosexuality" by "doing gender" in a way that is consistent with their biological sex. We also found significant interactions between sex and gender on the likelihood of identifying as the person in the household most likely to do housekeeping and household repairs. These findings add considerable nuance to previous research focusing on sex differences in these types of labor and are consistent with the notion that the labor one does is intricately linked to gender.

In contrast to housekeeping and household repairs, the outcomes of being the person most likely to do childcare or grocery shopping and hours working for pay, doing housework, and doing care work did not see added explanatory value from gender identification. It seems the shared variance between the sex and gender identification measures is what is associated with these outcomes. Finally, our data suggest that gender identification is more important than sex in explaining military service. Future research could examine whether this gender effect is caused by serving in the military or by more masculine individuals selecting into the military (and other similar occupations).

Taken together, our results confirm the findings of Magliozzi et al. (2016) that measures of gender identification add considerable explanatory value beyond measures of binary sex. In addition, our one-item measure is parsimonious, answerable, and reliable and valid. Our results also extend the work of Magliozzi et al. (2016) by evaluating the gender identification scale in a national probability sample survey and with considerably more outcomes motivated by contemporary gender theory. In addition, we were able to examine the moderating effect of gender identification on existing sex differences findings.

There are several notable differences between our measure and that of Magliozzi et al. (2016). Whereas Magliozzi and colleagues utilized separate scales for masculinity and femininity, we utilized only one scale, placing masculinity and femininity in contrast to one another. The use of a single bipolar scale for gender identification has been critiqued for treating masculinity and femininity as mutually exclusive and opposites (Constantinople 1973), but these critiques were developed in the context of multi-item, domain- or trait-specific batteries that were largely designed to discriminate between males and females. In essence, these early bipolar masculinity/ femininity scales were designed to be social/psychological proxies for biological sex, not to capture gender as something separate from (or in addition to) sex. The new gradational measures of gender identification proposed here do not rely on specific domains or traits. Instead, they are "more comprehensive" (Magliozzi et al. 2016:7) measures. Thus, it is unclear the extent to which old critiques apply to these new measures.

Certainly separate masculinity and femininity scales can capture more nuance than a single scale and more closely match sociological theory about masculinity and femininity. However, it is unclear whether these sociological ideas make sense to general population members who will be asked to answer surveys. General understanding of a construct has direct bearing on how people answer the questions, what the questions are actually measuring, and how much measurement error they produce. Several researchers have demonstrated through qualitative work that gender nonconforming people understand and find utility in separate feminine and masculine scales (Kasabian 2015; Magliozzi et al. 2016; Garbarski and LaVergne, Chap. 9 this volume), but little research has examined how cisgender individuals understand and answer them. These are unsettled questions that will need to be addressed in order to move gradational gender measures into wide-scale general population survey use.

In addition to the theoretical debate about one versus two scales, this choice has practical implications for data collection, processing, and analyses that should be considered. Two scales require more space in the questionnaire and data entry time, which have direct cost implications. Respondents' ability and willingness to answer may also differ. A direct comparison of item nonresponse rates across one and

two-item measures should be made, and item-nonresponse rates should be compared across the items in the two-item format (higher item nonresponse rates to the second item would indicate respondent difficulty understanding masculinity and femininity as separate concepts).[7,8] Since the goal is to be able to use these measures in general population surveys, these tests should be conducted in general population surveys.

As gender identification measures continue to be developed and refined, thought should also be given to whether and how they can be administered in different survey modes, especially in telephone surveys with no visual cues. An open question is how respondents understand these scales when they cannot see them. Whether placing the two concepts on a single continuum or separating them into two scales helps or hinders this process is also an open question.

Measuring gender identification with one versus two items also has implications for the operationalization of variables for analyses. Given that most people in the general population are cisgender, we would expect a high correlation between the separate masculinity and femininity scales, making it difficult to use them in their original form in analyses because of multicollinearity. Magliozzi, et al. do not report correlations between their two scales or discuss multicollinearity, but they also do not include the two separate items in their predictive validity regression model, opting instead to combine them into a single measure of gender polarization. Using this measure, they find that more polarized people (i.e., high masculinity and low femininity or high femininity and low masculinity) are more likely to be married, but we don't know from their analysis the effect of gender identification in and of itself on marital status (i.e., are femininity and masculinity associated with marital status net of sex?) or if it varies by sex. We also do not know whether the reported association is the same for polarized masculine versus polarized feminine people or whether this depends on sex (i.e., are women who are polarized masculine more

7 Magliozzi, et al. did not report item nonresponse rates. Even so, the rates are not comparable across the two studies because of other design differences such as sample type and survey mode (web surveys typically have lower item nonresponse than mail – see *Survey Practice* 2012, volume 5, issue 2).

8 That Magliozzi, et al. included an instruction to "Please answer on both scales below" to prompt responses to both the feminine and masculine scales suggests respondents may not understand these concepts as separate in the way gender scholars do.

likely to be married, or just those who are polarized feminine?). Essentially, this choice to combined the two measures, which may have been driven by multicollinearity challenges, has the effect of eliminating the explanatory power that motivated asking about masculinity and femininity separately in the first place. This is a direct result of the heavily skewed (by sex) distribution of gender identification in the population at large. In less gender-conforming subpopulations, the masculinity and femininity scales may be less correlated, eliminating this challenge (see Garbarski and LaVergne, Chap. 9, this volume, for example), but it is a problem that will likely persist in general population usage. Using a single, bipolar scale eliminates these challenges in general population usage, allowing for a direct assessment of the association between gender identification and outcomes of interest. It also eliminates the potential for people to report being high on both femininity and masculinity, and thus may not fully capture existing gender variation. A direct experimental comparison between the two scales in a general population survey would help illuminate how many and what types of people might be affected by this omission.

A second difference between the two scales that raises important questions for future research is the number of scale points used. Whereas Magliozzi and colleagues used seven-point scales for each measure of masculinity and femininity, we used a 21-point scale to capture both, allowing for finer gradation in reports. How much gradation is needed to accurately capture gender variation is another open question.

While many empirical questions remain about how best to measure gender identification in general population surveys, this paper has demonstrated that it can be done in practical and affordable ways with reasonable reliability and validity, and that doing so adds considerable explanatory value. It is no longer sufficient to rely solely on binary measures of sex.

References

AAPOR. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys. American Association for Public Opinion Research*. Retrieved November 29, 2018, from www.aapor.org

Acker, J. (1990). Hierarchies, jobs, bodies: A theory of gendered organizations. *Gender & Society, 4*(2), 139–158.

Acker, J. (1992). From sex roles to gendered institutions. *Contemporary Sociology, 21*, 565–569.

Annandale, E., & Hunt, K. (1990). Masculinity, femininity, and sex: An exploration of their relative contribution to explaining gender differences in health. *Sociology of Health and Illness, 12*(1), 24–46.

Beatty, P., & Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item nonresponse. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse* (pp. 71–69). New York, NY: Wiley.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*(2), 155–162.

Berk, S. F. (1985). *Gender factory: The apportionment of work in American households*. New York: Plenum Press.

Bianchi, S. M., Milkie, M. A., Sayer, L. C., & Robinson, J. P. (2000). Is anyone doing the housework? Trends in the gender division of household labor. *Social Forces, 79*(1), 191–228.

Bianchi, S. M., Sayer, L. C., Milkie, M. A., & Robinson, J. P. (2012). Housework: Who did, does or will do it, and how much does it matter? *Social Forces, 91*(1), 55–63.

Bittman, M., England, P., Sayer, L., Folbre, N., & Matheson, G. (2003). When does gender trump money? Bargaining and time in household work. *American Journal of Sociology, 109*(1), 186–214.

Brandth, B. (2006). Agricultural body-building: Incorporations of gender, body and work. *Journal of Rural Studies, 22*(1), 17–27.

Brines, J. (1994). Economic dependency, gender, and the division of labor at home. *American Journal of Sociology, 100*(3), 652–688.

Britton, D. M. (1997). Gendered organizational logic: Policy and practice in Men's and Women's prisons. *Gender & Society, 11*, 796–818.

Bureau of Labor Statistics. (2014). *Women in in the labor force: A Databook*. U.S. Bureau of Labor Statistics Report 1052, Washington DC. Retrieved November 9, 2018, from https://www.bls.gov/opub/reports/womens-databook/archive/women-in-the-labor-force-a-databook-2014.pdf

Bureau of Labor Statistics. (2018). *Employment status of the civilian noninstitutional population 16 years and over by sex, 1970s to date*. U.S. Bureau of Labor Statistics, Washington DC. Retrieved November 9, 2018, from https://www.bls.gov/cps/tables.htm#empstat

Burke, P. J. (1991). Identity processes and social stress. *American Sociological Review, 56*(6), 836–849.

Connell, R. W. (1987). *Gender and power: Society, the person, and sexual politics*. Stanford, CA: Stanford University Press.

Connell, R. W. (1995). *Masculinities: Knowledge, power and social change*. Berkeley, CA: University of California Press.

Correll, S., Ridgeway, C., Saperstein, A., & Westbrook, L. (2014). *Gender Identity and Diversity: Revision and Updates*. (unpublished report)

Dalton, S. E., & Bielby, D. D. (2000). 'That's our kind of constellation': Lesbian mothers negotiate institutionalized understandings of gender within the family. *Gender & Society, 14*(1), 36–61.

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics, 19*(2), 153–176.

Egen, S. K., & Perry, D. G. (2001). Gender identity: A multidimensional analysis with implications for psychosocial adjustment. *Developmental Psychology, 37*(4), 451–463.

England, P. (1992). *Comparable worth: Theories and evidence*. New York: Aldine de Gruyter.

Enloe, C. (2004). Wielding masculinity inside Abu Ghraib: Making feminist sense of an American military scandal. *Asian Journal of Women's Studies, 10*(3), 89–102.

Epstein, J. (1990). Either/or-neither/both: Sexual ambiguity and the ideology of gender. *Genders, 7*, 100–142.

Federal Committee on Statistical Methodology. (2016a). *Current measures of sexual orientation and gender identity in federal surveys*. August 2016. Retrieved December 10, 2018 from https://nces.ed.gov/FCSM/interagency_reports.asp

Federal Committee on Statistical Methodology. (2016b). *Evaluations of sexual orientation and gender identity survey measures: What have we learned?* September 2016. Retrieved December 10, 2018, from https://nces.ed.gov/FCSM/interagency_reports.asp

Federal Committee on Statistical Methodology. (2016c). *Toward a research agenda for measuring sexual orientation and gender identity in federal surveys: Findings, recommendations, and next steps*. October 2016. Retrieved December 10, 2018, from https://nces.ed.gov/FCSM/interagency_reports.asp

Felshin, J. (1974). The triple option for women in sport. *Quest, 21*, 36–40.

Fraser, G. (2018). Evaluating inclusive gender identification measures for use in quantitative psychological research. *Psychology & Sexuality, 9*(4), 343–357.

Frye, M. (1983). *The Politics of Reality: Essays in Feminist Theory*. Trumansburg, NY: The Crossing Press.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly, 69*(1), 124–157.

Geist, C., Reynolds, M. M., & Gaytán, M. S. (2017). Unfinished business: Disentangling sex, gender, and sexuality in sociological research on gender stratification. *Sociology Compass, 11* (4), e12470.

Geist, C., & Ruppanner, L. (2018). Mission impossible? New housework theories for changing families. *Journal of Family Theory and Review, 10*, 242–262.

The GenIUSS Group. (2014). *Best practices for asking questions to identify transgender and other gender minority respondents on population-based surveys*. Los Angeles, CA: The Williams Institute. Retrieved December 10, 2018, from https://williamsinstitute.law.ucla.edu/wp-content/uploads/geniuss-report-sep-2014.pdf

Greenstein, T. N. (2000). Economic dependence, gender, and the division of labor in the home: A replication and extension. *Journal of Marriage and the Family, 62*, 322–335.

Hall, E. J. (1993). Waitering/waitressing: Engendering the work of table servers. *Gender & Society, 7*, 329–346.

Hays, S. (1996). *The cultural contradictions of motherhood*. New Haven, CT: Yale University Press.

Herbert, M. S. (1998). *Camouflage Isn't only for combat: Gender, sexuality, and women in the military*. New York, NY: New York University Press.

Hyde, J. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592.

Kasabian, A. (2015). *Capturing the Gendiverse: A test of the gender self-perception scale, with implications for survey data and labor market measures. Unpublished doctoral dissertation*. Lincoln, NE: University of Nebraska-Lincoln.

Kessler, S. J. (1990). The medical construction of gender: Case Management of Intersexed Infants. *Signs, 16*(1), 3–26.

Kessler, S. J., & McKenna, W. (1978). *Gender: An Ethnomethodological approach*. New York: Wiley.

Kroska, A. (2000). Conceptualizing and measuring gender ideology as an identity. *Gender & Society, 14*, 368–394.

Krosnick, J. A. (2002). The causes of no-opinion responses to attitude measures in surveys: They rarely are what they appear to be. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 88–100). New York: Wiley.

Kruttschnitt, C. (2013). Gender and crime. *Annual Review of Sociology, 39*, 291–308.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Lorber, J. (1996). Beyond the binaries: Depolarizing the categories of sex, sexuality, and gender. *Sociological Inquiry, 66*, 143–159.

Lucal, B. (1999). What it means to be gendered me: Life on the boundaries of a dichotomous gender system. *Gender & Society, 13*, 781–797.

Magliozzi, D., Saperstein, A., & Westbrook, L. (2016). Scaling up: Representing gender diversity in survey research. *Socius, 2*, 1–11.

Mahalik, J. R., Morray, E. B., Coonerty-Femiano, A., Ludlow, L. H., Slattery, S. M., & Smiler, A. (2005). Development of the conformity to feminine norms inventory. *Sex Roles, 52*, 417–435.

McMahon, M. (1995). *Engendering motherhood: Identity and self-transformation in Women's lives*. New York: Guilford Press.

Olson, K., & Smyth, J. D. (2017). Within-household selection in mail surveys explicit questions are better than cover letter instructions. *Public Opinion Quarterly, 81*(3), 688–713.

Olson, K., Watanabe, M., & Smyth, J. D. (2018). A comparison of full and quasi-filters for autobiographical questions. *Field Methods, 30*(4), 371–385.

Oransky, M., & Fischer, C. (2009). The development and validation of the meanings of adolescent masculinity scale. *Psychology of Men and Masculinity, 10*(1), 57–72.

Padavic, I., & Reskin, B. (2002). *Women and men at work*. Thousand Oaks, CA: Sage.

Price, K. (2008). Keeping the dancers in check: The gendered Organization of Stripping Work in the Lion's Den. *Gender & Society, 22*, 367–389.

Ridgeway, C. L., & Smith-Lovin, L. (1999). The gender system and interaction. *Annual Review of Sociology, 25*, 191–216.

Risman, B. J. (2004). Gender as social structure: Theory wrestling with activism. *Gender & Society, 18*, 429–450.

Saugeres, L. (2002). 'She's not really a woman, She's half a man': Gendered discourses of embodiment in a French farming community. *Women's Studies International Forum, 25*(6), 641–650.

Schwarz, N., & Strack, F. (1999). Reports of subjective Well-being: Judgmental processes and their methodological implications. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 61–84). New York, NY: Russell Sage Foundation.

Smyth, J. D. (2007). *Doing gender when home and work are blurred: Women and sex-atypical tasks in family farming. Unpublished dissertation*. Pullman, Washington: Washington State University.

Smyth, J., Olson, K., & Stange, M. (2019). Within-household selection methods: A critical review and experimental examination. Chapter 2. In P. J. Lavrakas, M. W. Traugott, C. Kennedy, A. L. Holbrook, E. D. de Leeuw, & B. T. West (Eds.), *Experimental methods in survey research: Techniques that combine random sampling with random assignment* (pp. 23–46). Hoboken, NJ: Wiley.

Smyth, J. D., Swendener, A., & Kazyak, E. (2018). Women's work? The relationship between Farmwork and gender self-perception. *Rural Sociology, 83*(3), 654–676.

Sobal, J. (2005). Men, meat, and marriage: Models of masculinity. *Food and Foodways, 13*, 135–158.

South, S. J., & Spitze, G. (1994). Housework in marital and NonMarital households. *American Sociological Review, 59*, 327–347.

Spence, J. T., Helmreich, R., & Stapp, J. (1974). The personal attributes questionnaire: A measure of sex-role stereotypes and masculinity and femininity. *JSAS: Catalog of Selected Documents in Psychology, 4*, 43–44.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Timbrook, Jerry, Jolene D. Smyth, and Kristen Olson. (2016). *Does Adding 'Your Best Estimate is Fine' Affect Data Quality?* Paper presented at the International Conference on Questionnaire Design, Development, Evaluation, and Testing, Miami, FL, November 9–13, 2016.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299–314.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, MA: Cambridge University Press.

Verbrugge, L. M. (1985). Gender and health: An update on hypotheses and evidence. *Journal of Health and Social Behavior, 26*(3), 156–182.

Wade, L., & Ferree, M. M. (2019). *Gender: Ideas, interactions, institutions* (2nd ed.). New York: W. W. Norton & Company.

Walzer, S. (1998). *Thinking about the baby: Gender and transitions into parenthood*. Philadelphia, PA: Temple University Press.

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

Westbrook, L., & Saperstein, A. (2015). New categories are not enough: Rethinking the measurement of sex and gender in social surveys. *Gender & Society, 29*(4), 534–560.