2-2023

The Measurement of Gender Expression in Survey Research

Dana Garbarski
Loyola University Chicago, dgarbarski@luc.edu

Recommended Citation

Garbarski, Dana. The Measurement of Gender Expression in Survey Research. Social Science Research, 110, : 1-16, 2023. Retrieved from Loyola eCommons, Sociology: Faculty Publications and Other Works, http://dx.doi.org/10.1016/j.ssresearch.2022.102845

The Measurement of Gender Expression in Survey Research

Dana Garbarski

Department of Sociology

Loyola University Chicago

1032 W. Sheridan Rd.

Chicago, IL 60660

dgarbarski@luc.edu

Abstract

Previous research on the survey measurement of sexual orientation, gender identity, and gender expression (SOGIE) often focuses on the measurement of identity, with comparably little research focused on *gender expression* as a key feature of how gender is lived and experienced. This study examines the reliability and validity of survey questions about gender expression in a 2-by-5-by-2 factorial experiment that varies the question order, type of response scale, and the order of gender presentation in the response scale.

The results indicate that the effect of which (side of the) scale is presented first on gender expression varies by gender for each of the unipolar items and one of the bipolar items (behavior). In addition, the unipolar items also show distinctions among the gender minority population in ratings of gender expression as well as more nuance with respect to concurrent validity in predicting health outcomes among cisgender respondents. The results of this study have implications for researchers who are interested in accounting for gender holistically in survey and health disparities research.

Keywords

Gender, gender expression, survey questions, health disparities, survey measurement

Introduction

The survey measurement of sexual orientation, gender identity, and gender expression (SOGIE) allows for the valid and reliable enumeration of sexual and gender minority populations, fuller representation of the scope of human identity and experience, and nuanced assessment of the predictors of health disparities. Several working groups and researchers have documented the best practices and emerging issues and research for the measurement of sexuality and gender in surveys (Federal Interagency Working Group 2016a, 2016b; GenIUSS 2014; NASEM 2022; SMART 2009). These summaries indicate that best practices for the survey measurement of sexuality and gender remain preliminary and incomplete, requiring rigorous empirical examination across a range of populations and survey conditions. Importantly, much of the previous research focuses on the measurement of gender and sexual identity in surveys, with comparably little focus on best practices for the measurement of other dimensions of gender in surveys.

Although theoretical and analytic approaches may vary, the consensus among social scientists that study gender is that sex and gender are distinct and nonbinary (Connell 2005; Courtenay 2000; Lorber 1994; Martin 2003; Risman 2018; West and Zimmerman 1987). Yet the survey measurement of gender often shows a strong adherence to the gender binary, conflates the concepts of sex and gender, and often assumes that gender can be easily determined by others, such as an interviewer (Westbrook and Saperstein 2015). As a result, the recommended way to determine gender identity in surveys is a set of questions that asks in two parts one's current gender identity, that is, their internal sense of gender (with response categories such as, e.g., woman, man, transgender woman, transgender man, gender nonbinary) and their assigned sex at

birth ("What was your sex assigned at birth, for example, on your original birth certificate?") (GenIUSS 2014; NASEM 2022; Saperstein and Westbrook 2020).

However, these discrete categorizations of gender identity are limited in their ability to explain the diversity of how gender manifests in the social world. Another key feature of the variation with which gender is lived, experienced, and has implications for life chances is in terms of one's gender expression. Gender expression is the presentational dimension of gender, that is, how gender is displayed through appearance and enacted through behavior (GenIUSS 2014; Lorber 1994; NASEM 2022; Spence 2011). While gender identity refers to categorical distinctions (e.g., woman), gender expression refers to external manifestations of masculinity and femininity in appearance (e.g., clothing) and behavior (e.g., mannerisms), and can refer to how one sees themselves or how others see them (further discussed in Background below). These more gradational measures of gender demonstrate diversity within categories of gender identity (e.g., femininity existing on a spectrum for cisgender women), as well as overlap between categories of gender identity (e.g., cisgender women and men having similar ratings of masculinity). One example of how gender expression is operationalized in surveys is described in the GenIUSS group's (2014) report: "A person's appearance, style, or dress may affect the way people think of them. On average, how do you think people would describe your appearance, style, or dress?" and "A person's mannerisms (such as the way they walk or talk) may affect the way people think of them. On average, how do you think people would describe your mannerisms?" followed with response scales "very feminine, mostly feminine, somewhat feminine, equally feminine and masculine, somewhat masculine, mostly masculine, very masculine."

Scholars contend that measures of gender expression in survey research are essential to represent the multiplicity with which gender is lived and experienced and the scope of how gender inequality has implications for life chances (Garbarski and LaVergne 2020; Hart, Saperstein, Magliozzi, and Westbrook 2019; Magliozzi, Saperstein, and Westbrook 2016; Smyth and Olson 2020). With respect to health, some research links feminine behaviors with good health and masculine behaviors with poor health (Bird and Rieker 2008; Springer and Mouzon 2011). Yet other research shows that intersection of gender expression with identity is important for health: masculinity is associated with better self-rated health (SRH) for cisgender men, whereas femininity is associated with better SRH for cisgender women (Hart et al. 2019). In addition, those who are gender nonconforming are at an increased risk for poor health and discriminatory treatment by others (Austin et al. 2016; GenIUSS 2014; Gordon and Meyer 2007; Lowry et al. 2018; Miller and Grollman 2015). This research indicates that the boundaries and contours of gendered health disparities must be examined in a multifaceted way at individual, interactional, and institutional levels, starting with including a multidimensional accounting of gender in large-scale data collection efforts.

While a growing body of research demonstrates the utility of measuring gender expression in surveys, how gender expression is operationalized in survey research varies considerably across studies. As described below, researchers continue to grapple with the way to operationalize the survey measurement of gender expression, and this study seeks to further contribute to that effort in terms of three features of survey measurement: the type of response options offered, the order of the questions (offering self or reflected appraisals of gender expression first), and the order of presentation of masculine and feminine components in the questions.

Background

Gender expression as a multidimensional construct. In most operationalizations of gender expression, "feminine" and "masculine" are set up as bipolar opposites along the same unidimensional continuum (Bem 1993; Connell 2005; Constantinople 1973; Lorber 1994; Risman 2018) and the response options indicate the amount of masculinity or femininity, with the midpoint signaling equal amounts of each (Bittner and Goodyear-Grant 2017; Cassino 2020; Smyth and Olson 2020). However, the response scale need not be constructed to reify the duality in which masculine and feminine are defined in opposition as relational statuses. Indeed, psychology has long studied masculine and feminine personality traits as two distinct dimensions, albeit not necessarily decoupled from sex characteristics or gender identity (Constantinople 1973; Bem 1993; Risman 2018; Zucker et al. 2006).

It is an empirical question whether separating masculine and feminine into two response dimensions may allow for respondents to more fully locate their gender presentation in a survey (Garbarski and LaVergne 2020), but one that is worth exploring for both conceptual and empirical reasons. Research by Magliozzi and colleagues (2016) and Hart and colleagues (2019) separates masculinity and femininity into two unipolar response dimensions, one for masculine expression and one for feminine expression. Although femininity overlaps significantly with identifying as a woman and masculinity with men among cisgender respondents, variation in masculinity and femininity ratings exist within each gender identity (Magliozzi et al. 2016). Polarization—fulfilling conventional gender norms such that a cisgender man has a higher score for masculinity than femininity and vice versa for cisgender women—is associated with greater odds of being married (Magliozzi et al. 2016). Hart and colleagues (2019) use the same data and construct measures of nonconformity—cisgender women who are more masculine than feminine

and cisgender men who are more feminine than masculine—and examine how nonconformity is associated with SRH. These studies document how separate measures of masculinity and femininity allow for a more nuanced understanding of gender in studies of inequality.

Related to the conceptual underpinnings of treating masculinity and femininity as distinct constructs is how this becomes operationalized in response scales. The studies by Magliozzi and colleagues (2016) and Hart and colleagues (2019) use two seven-point unipolar ratings scales, with "not at all" and "very" labeled at the endpoints and numbers 1 through 5 listed in the middle. However, such scales do not follow the best practices of survey methodology. Across several experimental and observational studies and a range of topics, research shows that labeling only the endpoints of response scales increases the likelihood of extreme responses (choosing the first and last response categories) and decreases reliability and criterion validity (see Schaeffer and Dykema [2020] for a summary of this research). This is likely because providing verbal labels for each scale point removes a step from the cognitive processing for respondents, who, when presented with numbers in a rating scale, must construct an implicit definition of what the numbers communicate in order to provide an answer (Krosnick and Presser 2010). Garbarski and LaVergne (2020) suggest the following response options for unipolar measures of masculinity and femininity based on research on the scaling of quantifiers (Beckstead 2014; Dobson and Mothersill 1979): "not at all," "a little," "somewhat," and "very." Schaeffer and Dykema (2020) suggest five categories for unipolar ratings scales. A fifth category of "extremely" could be considered, because it is not clear whether "very" is intense enough to be the highest category (Beckstead 2014). Overall, the limited research on different types of response scales for measures of gender expression leads to the following research question:

Research question 1: Do different types of response scales—bipolar or unipolar; unipolar with verbal or end-point only labels; "very" or "extremely" as the last category—impact the survey measurement of gender expression?

Gender expression as self or reflected appraisal. The survey items discussed in the GenIUSS (2014) report and included in the YRBS include reflected appraisals: asking respondents to report on how most people see them (Cooley 1902; Mead 1934; Ridgeway and Correll 2006), which reflects in part that gender is an interactional and performative accomplishment and determined both by oneself and others (West and Zimmerman 1987; Westbrook and Schilt 2014). However, the research done by sociologists focuses on both reflected appraisals as well as self-appraisals: one's report on their own gender expression (Garbarski and LaVergne 2020; Hart et al. 2019; Magliozzi et al. 2016; Smyth and Olson 2020). Although self and reflected appraisals tend to overlap substantially, a discrepancy between the two is associated with worse health outcomes among cisgender (Hart et al. 2019) and transgender (Miller and Grollman 2015) survey respondents.

When researchers are interested in both self and reflected appraisals, one feature of survey design that must be accounted for is question order, that is, which question should come first, one's own perception or the perceptions of others? As is well documented in survey methodological research, question order communicates meaning to respondents (Tourangeau, Rips, and Rasinski 2000). The former question can influence the interpretation of the definition or response scale of the latter question, as well as potentially activate a memory structure of beliefs, evaluations, and feelings about the broader topic which become salient when formulating an answer to the latter question (Tourangeau, Rips, and Rasinski 2000).

The evidence on the effects of self and reflected appraisal question order on the distribution of gender expression is mixed. In one sample from Amazon Mechanical Turk (MTurk), the researchers did not find differences in estimates of gender expression across ordering of self and reflected appraisals (Hart et al. 2019; Magliozzi et al. 2016). In a nationally representative survey, Smyth and Olson (2020) find evidence of question order effects with questions that ask for self-appraisal and society's ideals: cisgender women rate themselves as more masculine when asked about society's ideal man and woman before reporting a self-appraisal. Garbarski and LaVergne (2020) conducted cognitive interviews with queer women, in which they asked respondents the gender expression question from GenIUSS (2014)—that is, the reflected appraisal question—and asked respondents to report on what they were thinking about when they answered that question. Only about one-third of participants explicitly provide evidence of others' perceptions when describing how they arrived at their answer to this question in cognitive interviews; rather, they were more likely to convey what they are doing to present themselves (Garbarski and LaVergne 2020). Thus, Garbarski and LaVergne (2020) argue that reflected appraisals add an additional dimension to comprehension and processing and thus variability in terms of what respondents are considering when answering the question, and that self-appraisal should be administered first when both are included in the survey. Overall, it is an empirical question as to whether there are differences by question order, and qualitative research indicates that rather than varying the presentation, self-presentation should come first.

Research question 2: Does the question order of self and reflected appraisals of gender expression impact the survey measurement of gender expression?

Which comes first, masculine or feminine? One common practice in surveys is to "match" the presentation of a response scale with one's gender, e.g., present the feminine (side of the)

scale first if the respondent is presumed to be a woman and masculine first if the respondent is presumed to be a man. Another practice is to present the same side of the scale to everyone regardless of gender, e.g., in a self-administered paper survey such as the Youth Risk Behavioral Survey. When the order of the presentation of the response scale changes depending on one's gender identity, this decision is likely based on the assumption that the errors driven by response option order will be roughly the same in both directions (e.g., no difference in whether presenting feminine first to women and masculine first to men) (Garbarski and LaVergne 2020). One complicating factor is more fundamental to survey methodology in terms of starting with the most "desirable" set of response options, that is, the response option respondents are more likely to choose as a satisfactory answer for the purposes of the question. Indeed, starting with the least desirable response options first allows for respondents to be more likely to consider the range of responses before selecting an answer (Bradburn et al. 2004; Garbarski et al. 2015). However, this recommendation presumes that the underlying population considers the response options to be uniformly applicable (Garbarski and LaVergne 2020).

Research question 3: Does the order of gender presentation in the response scales (masculine or feminine first) impact the survey measurement of gender expression?

Current study

This study examines each of the three research questions about the survey measurement of gender expression in a 2-by-5-by-2 factorial experiment administered through MTurk that varies the type of response scale, question order, and the order of gender presentation in the response scale. Reliability and validity are used to assess the survey measurement of gender expression. Reliability refers to the extent to which responses to survey items are consistent,

stable, and dependable across context; validity refers to the extent to which the survey items measure what they are supposed to.

Measurement reliability is assessed in two ways: whether gender expression varies depending on question context (question order and the order of gender presentation in the response scales) and how much variance in gender expression is predicted by gender identity. First, to the extent that gender expression is an established, salient trait, question context can be used to assess measurement reliability (Smyth and Olson 2020). If assessments of gender expression vary with the context of the question—whether self-appraisal comes before or after reflected-appraisals or which (side of the) scale is presented first—this indicates that the item is less reliable than if the responses were the same regardless of question context. To further assess reliability of gender expression, consideration is given to how gender expression differs across cisgender, transgender, and nonbinary respondents.

This study examines the validity of gender expression in terms of concurrent validity; in particular, the extent to which gender expression is associated with mental and physical health in expected ways. Given the intersection of gender identity and expression in predicting health summarized in the Introduction, the interaction of gender identity and expression is expected to be associated with health. To the extent that the interaction of gender identity and expression predicts mental health and SRH beyond gender identity alone, this is evidence of the concurrent validity of gender expression.

Methods

Data

The survey was conducted online between May and August 2019 using workers from Amazon Mechanical Turk (MTurk; N = 5,872), a crowdsourcing method allowing researchers to find respondents to complete surveys and other "human intelligence tasks" (HITs). MTurk

workers register to complete tasks through the MTurk interface in exchange for small amounts of money. Although the rules of who can register change over time and are proprietary, all registrants must be 18 years old, and many tend to be US citizens or residents with verifiable identities. This study uses data from the 4,569 respondents who reside in the US and did not have the same IP address as another respondent (which may omit a small number of legitimate responses, for example, from members of the same household) (Aguinis et al. 2020).

The HIT announcement noted that the purpose of this study was to examine different measures of gender expression that are used in surveys and how these are associated with health, and indicated that respondents would be asked about gender expression, health, and demographic information in a survey that would take no more than 10 minutes to complete. After reading the announcement, workers were given the option to choose to participate in the survey. Those who chose to participate were redirected to a Turkitron page (www.turkitron.com) that asked them to input their MTurk ID to prevent repeat respondents. Those who had not previously completed the questionnaire were redirected to the Qualtrics interface that hosted the survey. At the end of the survey, respondents were given a number to enter in the HIT page for remuneration. Respondents received 75 cents for completing the task. This amount was determined by estimating that the task would take four minutes to complete (23 questions*10 seconds per question) and wanting to compensate respondents for their time at a rate above the federal minimum wage in the United States ($7.25 per hour).

Measures

The main variables of interest, which are both independent and dependent variables depending on the analysis, are the four questions on gender expression: self-appraisal of appearance, reflected appraisal of appearance, self-appraisal of behavior, and reflected appraisal

of behavior. The questions about gender expression are slightly revised from the more common version in the YRBS and GenIUSS report (listed in the Background section) (see Appendix A). A 2-by-5-by-2 factorial experiment was used to examine different versions of gender expression questions, listed in Appendix A. The first experimental factor varies whether the self or reflected appraisal questions appear first. The second experimental factor varies the type of response scale presented: bipolar (the version used by YRBS and described in the GenIUSS (2014) report); unipolar (two scales, one for masculine and one for feminine) with endpoints labeled "not at all" and "very" and numbers for the middle categories (Magliozzi et al. 2016; Hart et al. 2019); unipolar with endpoints labeled "not at all" and "extremely" and numbers for the middle categories; unipolar with verbal labels "not at all," "a little," "somewhat," and "very," (Garbarski and LaVergne 2020); and unipolar with verbal labels adding "extremely." The third experimental factor varies whether the feminine or masculine (side of the) scale is presented first.

Given the focus on the order of self- and reflected appraisals, type of response option, and order of gender presentation, several features of survey design were not experimentally manipulated in this study. The order of which feature of gender expression is presented first is held constant—appearance always precedes behavior. Similarly, the order of the blocks of questions are held constant—gender expression (because it is the primary focus of the study), SRH, depressive symptoms, additional health questions, then sociodemographics (starting with sex assigned at birth and current gender identity).

Gender identity is ascertained using the "two-step" approach that first asks about sex assigned at birth and current gender identity (GenIUSS 2014). Respondents are coded as cisgender men and women if their current gender identity aligns with their sex assigned at birth. Respondents are coded as transgender women and men if their current gender identity is "man"

or “woman” and their sex assigned at birth is “female” or “male,” respectively. Respondents who indicate a transgender or nonbinary identity are coded as such.

To examine concurrent validity, two measures of health are included in the survey. The first is a measure of mental health: the Patient Health Questionnaire depression module (PHQ-9), a set of nine items reflecting the nine criteria for major depression in the Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (DSM-5) and used in both clinical settings and general population survey research. These items score the severity of depressive symptoms experienced in the last two weeks from not at all (1) to nearly every day (4) and are summed to create a score of depressive symptoms (Kroenke and Spitzer 2002). Because the data are positively skewed, a natural log transformation is applied. The second variable is self-rated health (SRH), one of the most common survey measures used to examine health (Garbarski 2016).

Descriptive statistics for these health measures as well as a series of control variables are located in Table 1. As expected based on previous research and discussed in the Discussion section under limitations, the MTurk sample is more educated than the U.S. population and underrepresents the Black and Hispanic/Latine population and overrepresents the white population compared to the U.S. population (https://www.census.gov/quickfacts/fact/table/US/PST045219).

Table 1. Descriptive statistics for demographic variables, MTurk 2019

	N	Mean or Percent	Standard Deviation	Minimum	Maximum
Self-rated health	4,569				
Poor		2.06 %			
Fair		14.73 %			
Good		33.31 %			
Very good		37.97 %			
Excellent		11.93 %			
Mental health: PHQ-9 score	4,565	15.87	6.66	9	36

Gender	4,557					
Cisgender women		49.62	%			
Cisgender men		48.19	%			
Transgender women		0.44	%			
Transgender men		0.50	%			
Genderqueer or nonbinary		1.25	%			
Race	4,548					
White		70.54	%			
Latine		4.82	%			
Black		9.70	%			
Asian		8.18	%			
2 or more races		5.83	%			
Other		0.95	%			
Age	4,546	36.73		11.86	18	84
Sexuality	4,550					
Gay		4.02	%			
Straight		81.45	%			
Bisexual		11.87	%			
Multiple or other sexuality listed		2.20	%			
Not sure		0.46	%			
Education Level	4,569					
Less than high school		0.42	%			
High school		10.37	%			
Some college		21.62	%			
Associate degree		10.66	%			
Bachelor's degree		43.09	%			
Graduate or professional degree		13.83	%			
Marital Status	4,569					
Never married		45.04	%			
Married		43.69	%			
Separated		1.29	%			
Divorced		7.44	%			
Widowed		1.07	%			
Something not listed		1.47	%			
Employment Status	4,563					
Full-time job		64.04	%			
Part-time job		18.26	%			
Not employed for pay other than MTurk		17.71	%			

Household size	4,517	2.89	1.50	1	15
Language spoken at home	4,567				
English		97.13 %			
Another language		2.87 %			
Device used to take survey	4,569				
Desktop or laptop		94.18 %			
Smartphone		4.16 %			
Tablet		1.60 %			
Something not listed		0.07 %			

Notes. Analytic sample is N=4,569. SRH=self-rated health. PHQ-9=Patient Health Questionnaire-9. Percentages may not sum to 100% due to rounding.

Analytic strategy

Given the conceptual distinctions between self and reflected appraisals (Garbarski and LaVergne 2020; Hart et al. 2019), this study focuses on analyzing one of these dimensions: self-appraised appearance and behavior. In order to examine differences across the four types of unipolar scales (endpoint labeled with "very" as the top category, endpoint labeled "extremely," verbal labels with "extremely," and verbal labels with "very"), the measures are standardized (mean=0, standard deviation=1) and combined into a single scale for each of the following: appearance-unipolar-masculine, appearance-unipolar-feminine, behavior-unipolar-masculine, and behavior-unipolar-feminine. Because the bipolar scale cannot be meaningfully combined with the unipolar scales, separate analyses are conducted for unipolar and bipolar results. Thus, there are six measures of gender expression examined: appearance-bipolar, behavior-bipolar, appearance-unipolar-masculine, appearance-unipolar-feminine, behavior-unipolar-masculine, and behavior-unipolar-feminine. There were no significant two-way (for unipolar and bipolar) or three-way (for unipolar) interactions among the experimental factors in their effect on self-appraised appearance and behavior (Supplementary Appendix Table A.1), thus the standardized unipolar measures are used in the subsequent analysis.

The reliability of measures of gender expression is assessed in in two ways (Smyth and Olson 2020). The first set of results (Table 2) presents whether self-appraised gender expression varies depending on question context (question order and order of gender presentation for the response scales) by regressing (OLS) each of the 6 measures of gender expression on the experimental factors. These models also include an interaction term by gender identity, as we might expect the question context effects to vary by gender for the features of context that are tied to gender (this model is restricted to cisgender women and men as these two groups have sufficient sample size to examine the effects of question context by gender). The next set of results examines how much of the variation in self-appraised gender expression is predicted by gender identity by regressing each of the six measures of gender expression on gender identity and examining pairwise comparisons of differences in means (Table 3).

The concurrent validity of the measures of gender expression is examined by regressing (OLS) two measures of health (PHQ-9 and SRH) on gender identity, gender expression, their interaction, and relevant sociodemographic covariates (Tables 4 and 5 and Supplementary Appendix Tables A.2-A.5). The validity analyses are restricted to cisgender women and men because of the statistical power needed to estimate interaction effects. Evidence of concurrent validity of gender expression is demonstrated by the extent that the interaction of gender identity and expression predicts mental health and SRH beyond gender identity alone.

Finally, the results also present the correlations among the gender expression items (Table 6), as the data offer an opportunity to help researchers prioritize among the many possible measures of gender expression.

Results

Reliability

The first analysis considers how question context—response scale, question order, and the order of gender presentation in the response scale—affects measures of gender expression (self-appraised appearance and behavior), and how the effects of question context vary by gender identity (Table 2). We might expect the question context effects to vary by gender for the features of context that are tied to gender—in this case, the presentation order of the (side of the) scale. This part of the analysis focuses on cisgender women and men as these two groups have sufficient sample size to examine the effects of question context by gender.

The results indicate that the effect of which (side of the) scale is presented first on gender expression varies by gender identity for each of the unipolar items and one of the bipolar items (behavior). Figure 1 shows the model-predicted means for some of these results. With the bipolar item on behavior, cisgender men rate themselves as more masculine when the scale starts with "very feminine" compared to when the scale starts with "very masculine," and this is significantly different from the scale presentation effect for cisgender women (Figure 1a). With each of the unipolar items, cisgender women rate themselves as less masculine (Figure 1b) and more feminine (Figure 1c) when the masculine scale is presented first compared to the feminine scale, while cisgender men rate themselves as more masculine (1b) and less feminine (1c) when the masculine scale is presented first compared to the feminine scale (the pattern is the same for unipolar behavior items that are not included in the figure).

With respect to question order, respondents rate their appearance and behavior as more masculine when self-appraisal comes before reflected appraisal rather than after for the unipolar measures of gender expression, and this does not vary by gender identity (no significant interaction between question order and gender identity) (Table 2). There is no significant difference in how respondents rate their appearance and behavior across types of (unipolar)

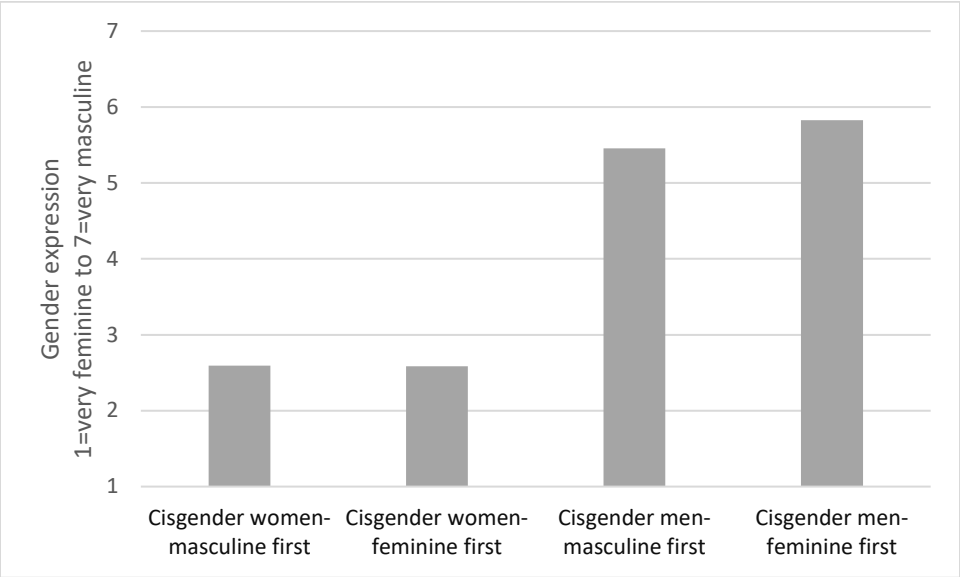
response options (using other unipolar response scales as reference groups show no significant differences), and no interaction with gender identity.

Table 2. Regression of self-appraisals of gender expression on scale order, question order, type of response scale, and gender, MTurk 2019

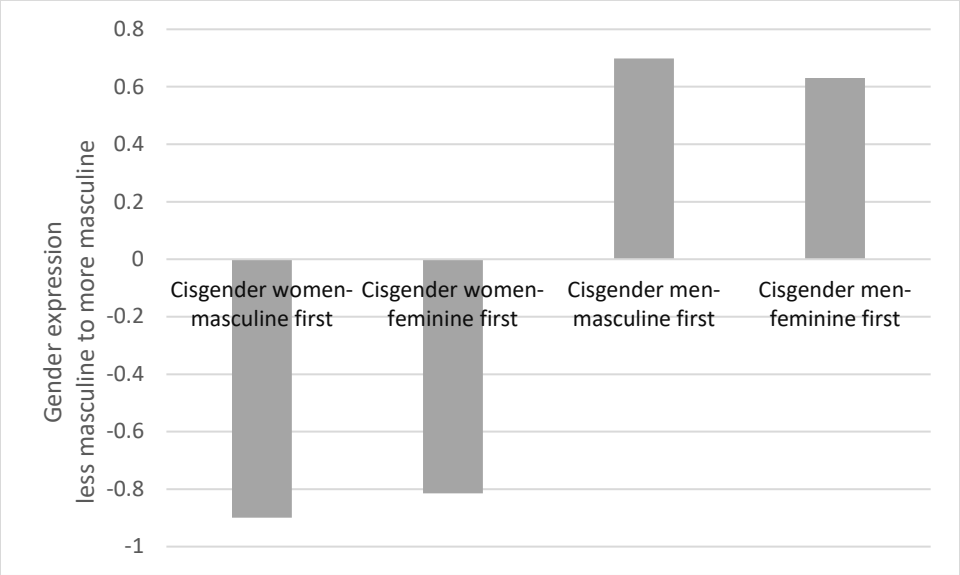
	Appearance Bipolar			Behavior Bipolar			Appearance Unipolar Masculine			Appearance Unipolar Feminine			Behavior Unipolar Masculine			Behavior Unipolar Feminine		
	Coef.	SE		Coef.	SE		Coef.	SE		Coef.	SE		Coef.	SE		Coef.	SE	
Cisgender men (vs. women)	3.121	0.136	***	2.934	0.141	***	1.628	0.053	***	-1.561	0.056	***	1.547	0.055	***	-1.481	0.059	***
Feminine (side of) scale first	0.096	0.110		-0.010	0.114		0.084	0.031	**	-0.069	0.033	*	0.095	0.033	**	-0.061	0.035	
Self-appraisal first	-0.061	0.110		0.121	0.114		0.101	0.031	***	-0.020	0.033		0.087	0.032	**	-0.038	0.035	
Response options																		
Endpoint labels very (reference)							--	--		--	--		--	--		--	--	
Endpoint labels extremely							0.023	0.043		0.000	0.046		0.004	0.045		0.009	0.048	
Verbal labels extremely							-0.043	0.045		-0.059	0.048		-0.043	0.048		-0.057	0.051	
Verbal labels very							0.026	0.043		-0.039	0.046		0.034	0.045		-0.042	0.048	
Men*feminine (side of) scale first	0.244	0.157		0.384	0.163	*	-0.151	0.044	***	0.113	0.047	*	-0.150	0.046	***	0.101	0.049	*
Men*self-appraisal first	-0.088	0.157		-0.144	0.163		-0.040	0.044		0.081	0.047		-0.031	0.046		0.094	0.049	
Response options																		
Men*endpoint labels very (reference)							--	--		--	--		--	--		--	--	
Men*endpoint labels extremely							-0.052	0.061		0.025	0.065		0.018	0.065		-0.020	0.069	
Men*verbal labels extremely							-0.016	0.063		0.126	0.067		-0.003	0.066		0.139	0.071	
Men*verbal labels very							0.023	0.061		0.029	0.066		0.030	0.065		0.039	0.069	
Intercept	2.476	0.097	***	2.531	0.100	***	-0.951	0.037	***	0.746	0.040	***	-0.912	0.040	***	0.716	0.042	***
N	911			912			3544			3543			3544			3545		
Adjusted R-squared	0.645			0.606			0.580			0.514			0.535			0.459		

Notes. Coef.=coefficient, SE=standard error. ***p<.001, **p<.01, *p<.05. N=sample size. Sample is restricted to cisgender women and men due to small samples of gender minority respondents. Regression is an ordinary least squares regression. No other differences among response options.

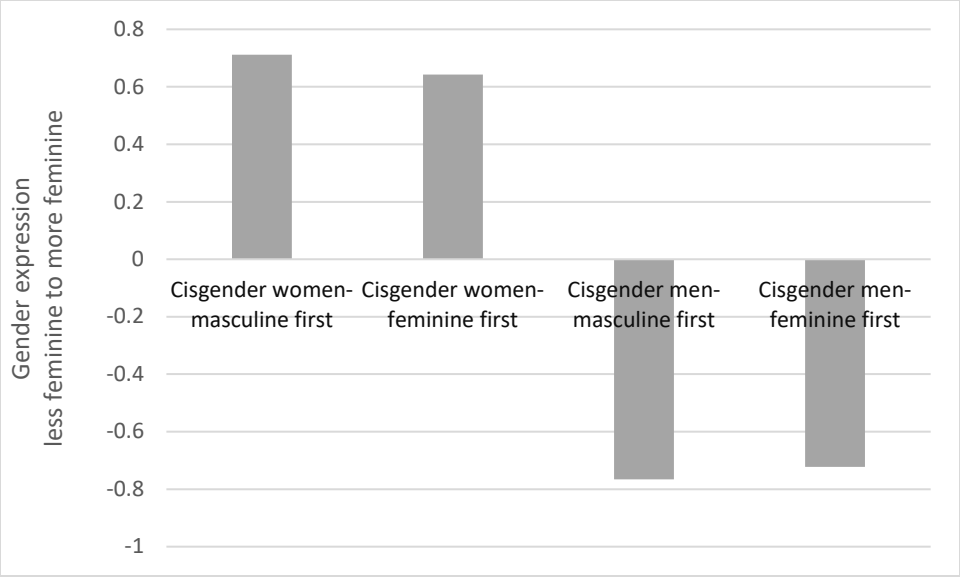
Figure 1. Interactions between scale presentation and gender in predicting gender expression, MTurk 2019



1a. Bipolar item-behavior



1b. Unipolar item-appearance masculine



1c. Unipolar item-appearance feminine

Notes. Sample is restricted to cisgender women and men due to small samples of gender minority respondents. In panel 1, 1=very feminine to 7=very masculine. In panels 2 and 3, negative values indicate less of the dimension (less masculine), and positive values indicate more of the dimension (more masculine).

Next, the association between self-appraised gender expression and gender identity is examined. Table 3 shows the predicted mean self-appraised gender expression from the (OLS) regression of gender expression on gender identity (self-appraised appearance and behavior for the bipolar and unipolar scales). The means follow expected patterns in terms of cisgender polarization: cisgender men have higher levels of self-appraised masculine appearance and behavior than all other groups, and cisgender women have lower levels of self-appraised masculine appearance and behavior than all other groups. The reverse is true for feminine appearance; in addition, mean feminine behavior is higher (although not significantly so) for transgender women compared to cisgender women.

An important difference emerges between the bipolar and unipolar measures of gender expression. With the unipolar measures, differences in gender expression emerge among those who are gender minorities (transgender women, transgender men, and gender nonbinary respondents): Transgender men evaluated themselves as more masculine than did transgender women and nonbinary respondents for appearance, and transgender women evaluated themselves as more feminine than did gender nonbinary respondents for feminine appearance and behavior. These differences are not present in the bipolar measures.[1] In this sense, the unipolar measures show better reliability in picking up distinctions in gender expression among those who are statistically rare in the sample.

[1] In order to examine whether this was due to the differences in sample size for gender minority respondents who received bipolar and unipolar measures, the models were re-estimated for respondents who received a particular set of unipolar response options (endpoint only labels-very, endpoint only-extremely, verbal labels-very, verbal labels-extremely). Many of the results were attenuated but still statistically significant (not shown).

Table 3. Mean self-appraised gender expression by gender identity, MTurk 2019

	Cisgender Women		Cisgender Men		Transgender Women		Transgender Men		Gender Nonbinary Respondents	
Appearance-bipolar	2.494	**b, d, e**	5.679	**a, c,** d, **e**	3.000	**b**	4.375	**a,** b	3.750	**a, b**
Behavior-bipolar	2.588	**b,** *c,* **d, e**	5.628	**a,** *c, d,* **e**	4.000	*a, b*	4.750	**a,** *b*	3.750	**a, b**
Appearance-masculine	-0.856	**b, c, d, e**	0.666	**a, c, e**	-0.268	**a, b,** d	0.424	**a,** c, *e*	0.024	**a, b,** *d*
Appearance-feminine	0.677	**b, d, e**	-0.744	**a, c, d, e**	0.506	**b, e**	0.067	**a, b**	-0.291	**a, b, c**
Behavior-masculine	-0.821	**b,** *c,* **d, e**	0.646	**a, c, d, e**	-0.392	*a,* **b**	0.034	**a, b**	-0.163	**a, b**
Behavior-feminine	0.645	**b,** *d,* **e**	-0.702	**a, c, d, e**	0.678	**b, e**	0.242	*a,* **b**	-0.175	**a, b, c**
N	2,261		2,196		20		23		57	

Notes. Significant differences noted above are at the level of p<.001 if bolded, p<.01 if not bolded, and p<.05 if italicized.
a=significantly different from cisgender women, b=significantly different from cisgender men, c=significantly different from transgender women, d=significantly different from transgender men, e=significantly different from genderqueer or nonbinary persons.
For the bipolar items, very feminine=1 and very masculine=7. The unipolar items are standardized so mean=0 and standard deviation=1, such that a positive score is more masculine on the masculine scales and more feminine on the feminine scales.

Validity

To examine the concurrent validity of the gender expression items, the interaction between gender identity and expression is examined for its association with a measure of mental health (PHQ-9, in which a higher score indicates worse mental health; Tables 4 and 5). Tables 4 and 5 display a series of regressions of the outcome of interest on the interaction of gender identity with the following measures of gender expression: self-appraised appearance and behavior (Model 2), reflected appraisal of appearance and behavior, (Model 3), both self and reflected appraisals (Model 4), and Model 5 adds other sociodemographic controls to Model 4. These results are compared to the bivariate model regressing mental health on gender identity alone (Model 1). The sample is restricted to cisgender women and men as these two groups have sufficient sample size to examine the effects of gender expression by identity in predicting health.

Table 4 shows that including the interaction between gender identity and bipolar measures of gender expression (Models 2-4) improves the prediction of mental health compared to the bivariate association of mental health and gender identity (Model 1) (in terms of adjusted R-squared, in which higher value indicates more variance explained, and AIC and BIC, in which lower scores indicate better fit when comparing two models). Self- and reflected appraisals of behavior interact with gender identity in predicting mental health (Model 4), and the interaction between gender identity and self-appraised behavior remains significant once sociodemographic controls are introduced (Model 5). Table 5 shows that for the unipolar measures of gender expression, model fit in predicting mental health improves when including the interaction between gender identity and expression (Models 2-4) compared to gender identity alone (Model 1), and half of the interactions between gender identity and gender expression are significant in

Models 4 and 5 (self-appraised masculine behavior, feminine and masculine reflected appraisals of appearance, and reflected appraisals of feminine behavior).

Figure 2 illustrates how the unipolar measures offer a more nuanced interpretation of how gender expression and gender identity combine to influence mental health, focusing on self-appraised behavior. For the bipolar measures of gender expression, the negative interaction indicates that moving from "very feminine" to "very masculine" is associated with decreasing poor mental health for cisgender men and increasing poor mental health for cisgender women. With the unipolar items, we see that the association of masculinity of self-appraised behavior with poor mental health is significantly different for cisgender women and men, while femininity of self-appraised behavior is not. In other words, using the unipolar items allows for identification of particular dimensions of gender expression—masculinity and femininity—and demonstrates that, in particular, the intersection of gender identity and masculinity of self-appraised behavior is associated with mental health.

Supplementary analyses. In the supplementary appendix, the same sets of models are presented using self-rated health (SRH) as an outcome of interest (Supplementary Appendix Tables A.2 and A.3). As with the results for mental health, gender identity moderates the association of several measures of gender expression with SRH, and the unipolar scales have the potential to identify unique interactions between gender identity and masculinity and femininity that the bipolar scales cannot capture. In addition, model fit improves when including the interaction between gender identity and gender expression in predicting SRH, with the exception of the bipolar items and BIC as a measure of model fit.

Finally, operationalizations of gender nonconformity from previous research are used to examine whether the current measures replicate the substantive results from previous research.

We build on findings of previous studies that demonstrate that nonconformity in gender expression (reflected appraisals) is associated with worse mental health (Lowry et al. 2018) and worse SRH (Hart et al. 2019) for both cisgender women and men. By including the interaction between nonconformity and gender identity, this study shows that the association between nonconformity in reflected appraisals of appearance and mental health are stronger for cisgender men than women (Supplementary Appendix Table A.4), and the effect of gender nonconformity on SRH (A.5) is stronger for cisgender women than men.

Table 4. Poor mental health (PHQ-9) regressed on bipolar measures of gender expression and gender identity, MTurk 2019

	Model 1			Model 2			Model 3			Model 4			Model 5		
	Coeff.	SE		Coeff.	SE		Coeff.	SE		Coeff.	SE		Coeff.	SE	
Cisgender men (vs. women)	-0.034	0.025		0.950	0.096	***	0.881	0.099	***	0.935	0.100	***	0.673	0.104	***
Appearance-self				0.006	0.020					-0.009	0.028		0.006	0.028	
Behavior-self				0.073	0.019	***				0.063	0.029	*	0.060	0.028	*
Appearance-other							0.023	0.019		0.019	0.027		0.006	0.026	
Behavior-other							0.050	0.018	**	0.007	0.027		-0.014	0.027	
Men*appearance-self				-0.074	0.030	*				-0.079	0.041		-0.076	0.040	
Men*behavior-self				-0.142	0.029	***				-0.090	0.041	*	-0.089	0.040	*
Men*appearance-other							-0.039	0.029		0.037	0.041		0.048	0.040	
Men*behavior-other							-0.164	0.027	***	-0.084	0.038	*	-0.042	0.038	
Feminine (side of) scale first				0.015	0.024		0.012	0.024		0.014	0.024		0.017	0.023	
Self-appraisal first				-0.039	0.024		-0.039	0.024		-0.041	0.024		-0.030	0.023	
Age													-0.004	0.001	***
Ethnoracial identity															
White (reference)													--	--	
Latino													-0.053	0.054	
Black													0.125	0.040	**
Asian													0.053	0.042	
2 or more Races													-0.111	0.056	*
Other													0.040	0.106	
Education is less than bachelor's degree (vs. bachelor's or higher)													-0.048	0.024	*
Married (vs. not)													0.020	0.024	
Sexual minority (heterosexual is reference)													0.147	0.034	***
Constant	2.687	0.017	***	2.496	0.045	***	2.516	0.042	***	2.497	0.045	***	2.687	0.061	***
R-squared (adjusted)	0.001			0.113			0.107			0.120			0.172		
AIC	792.34			691.96			697.50			689.01			642.38		
BIC	801.94			730.36			735.89			746.60			743.16		

Notes: Coef.=coefficient, SE=standard error. ***p<.001, **p<.01, *p<.05. Sample (N=897) is restricted to cisgender women and men due to small samples of gender minority respondents. Regression is an ordinary least squares regression. Gender expression is measured as very feminine=1 and very masculine=7.

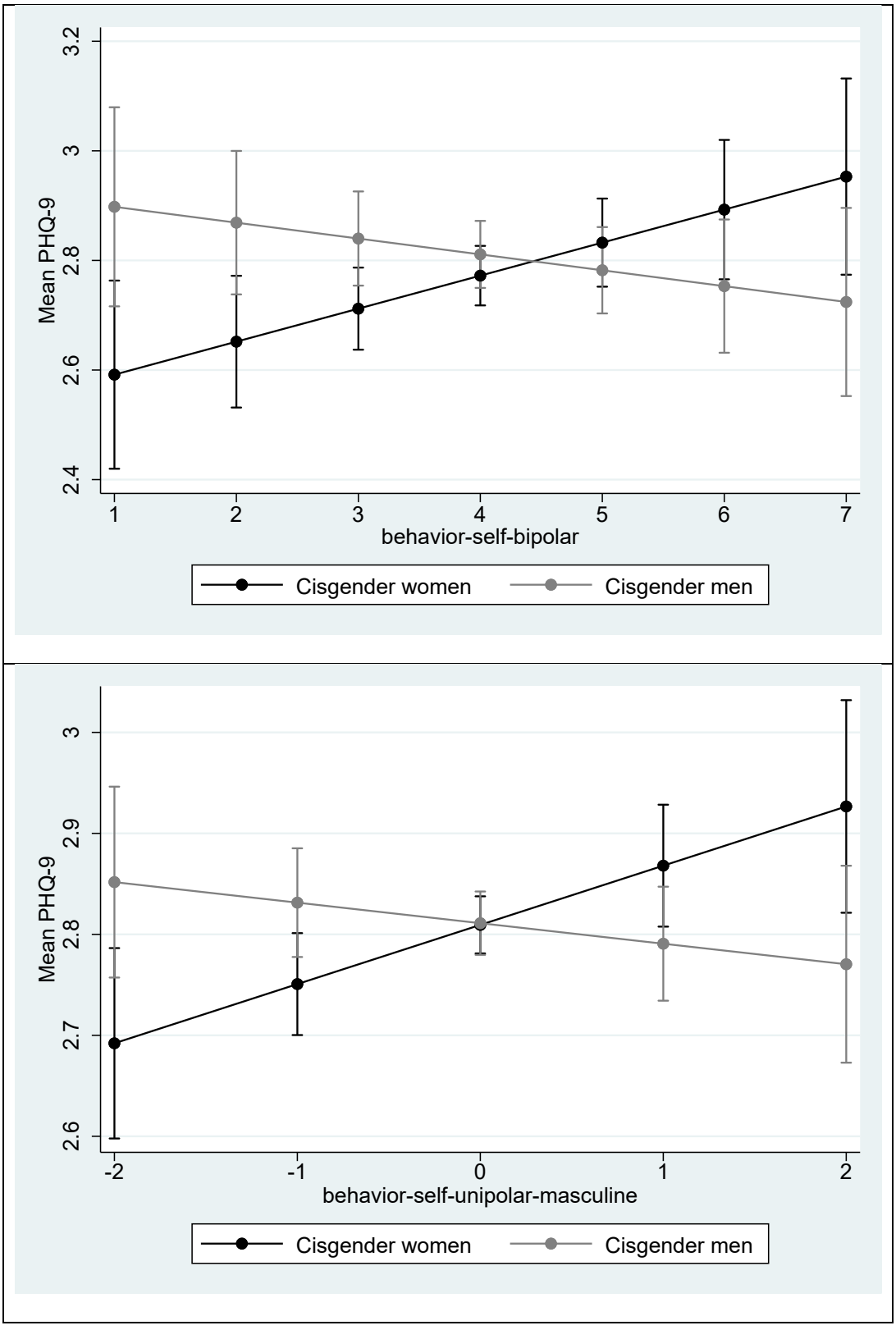
Table 5. Poor mental health (PHQ-9) regressed on unipolar measures of gender expression and gender identity, MTurk 2019

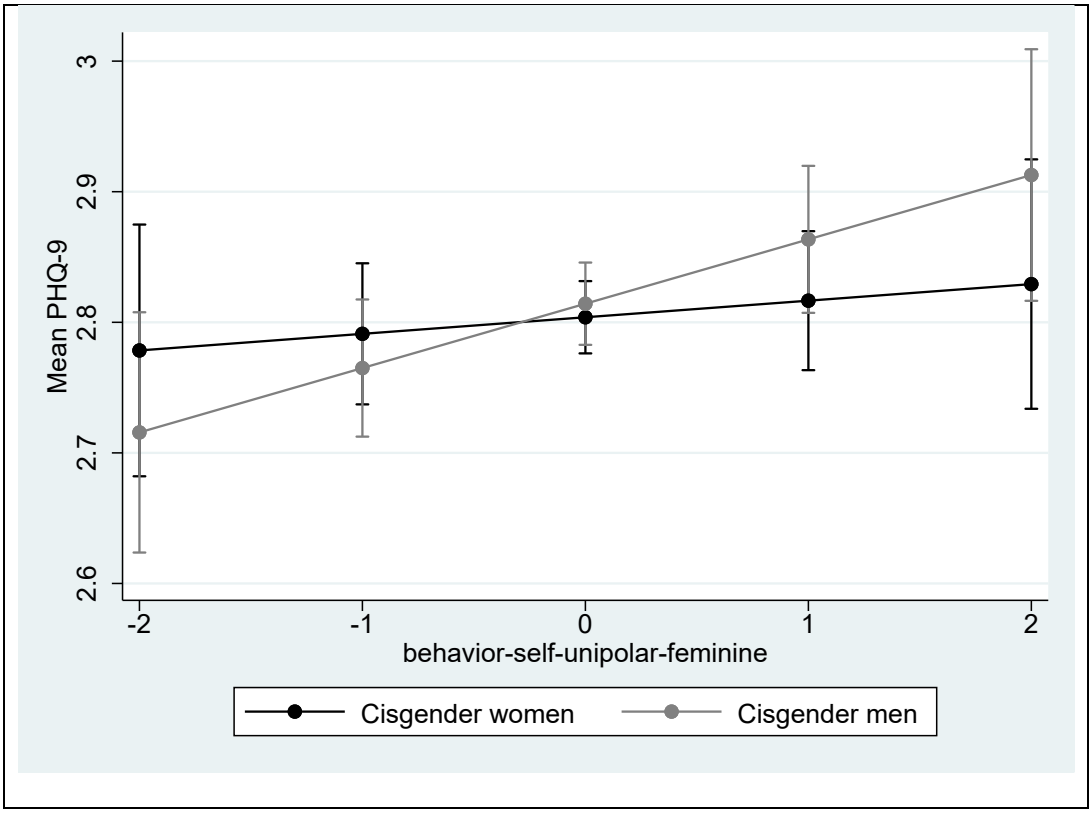
	Model 1		Model 2			Model 3			Model 4			Model 5		
	Coeff.	SE	Coeff.	SE		Coeff.	SE		Coeff.	SE		Coeff.	SE	
Cisgender men (vs. women)	-0.015	0.013	0.014	0.021		0.005	0.020		0.009	0.021		-0.002	0.021	
Appearance-masculine-self			0.080	0.018	***				0.016	0.022		0.005	0.021	
Appearance-feminine-self			0.029	0.018					0.026	0.024		0.017	0.024	
Behavior-masculine-self			0.144	0.018	***				0.073	0.025	**	0.059	0.024	*
Behavior-feminine-self			0.008	0.018					0.000	0.024		0.013	0.023	
Appearance-masculine-other						0.110	0.018	***	0.089	0.022	***	0.072	0.022	***
Appearance-feminine-other						0.034	0.019		0.013	0.025		0.007	0.024	
Behavior-masculine-other						0.121	0.018	***	0.063	0.025	*	0.066	0.025	**
Behavior-feminine-other						0.005	0.019		0.005	0.024		0.007	0.024	
Men*appearance-masculine-self			-0.060	0.026	*				0.020	0.032		0.027	0.032	
Men*appearance-feminine-self			0.106	0.026	***				0.019	0.033		0.021	0.033	
Men*behavior-masculine-self			-0.153	0.026	***				-0.099	0.034	**	-0.079	0.034	*
Men*behavior-feminine-self			0.129	0.025	***				0.052	0.033		0.037	0.033	
Men*appearance-masculine-other						-0.128	0.026	***	-0.123	0.032	***	-0.100	0.032	**
Men*appearance-feminine-other						0.097	0.025	***	0.079	0.033	*	0.075	0.032	*
Men*behavior-masculine-other						-0.088	0.026	***	-0.022	0.035		-0.021	0.034	
Men*behavior-feminine-other						0.138	0.025	***	0.090	0.033	**	0.071	0.033	*
Feminine (side of) scale first			0.021	0.012		0.018	0.012		0.018	0.012		0.019	0.012	
Self-appraisal first			-0.008	0.012		0.006	0.012		0.000	0.012		-0.001	0.012	
Response options														
Endpoint labels very (reference)			--	--		--	--		--	--		--	--	
Endpoint labels extremely			-0.005	0.017		-0.002	0.017		-0.003	0.017		-0.002	0.016	
Verbal labels extremely			0.017	0.017		0.016	0.017		0.017	0.017		0.019	0.017	
Verbal labels very			0.018	0.017		0.017	0.017		0.017	0.017		0.017	0.016	
Age												-0.004	0.001	***
Ethnoracial identity														
White (reference)												--	--	
Latino												0.019	0.028	
Black												0.024	0.020	
Asian												0.045	0.022	*
2 or more Races												-0.041	0.025	
Other												0.123	0.061	*

Education is less than bachelor's degree (vs. bachelor's or higher)													-0.048	0.012	***
Married (vs. not)													-0.015	0.012	
Sexual minority (heterosexual is reference)													0.129	0.017	***
Constant	2.681	0.009	***	2.830	0.020	***	2.831	0.020	***	2.836	0.020	***	2.976	0.029	***
R-squared (adjusted)	0.0001			0.186			0.198			0.202			0.241		
AIC	3354.66			2651.40			2599.84			2589.21			2426.12		
BIC	3366.98			2743.74			2692.19			2730.80			2623.12		

Note: Coef.=coefficient, SE=standard error. ***p<.001, **p<.01, *p<.05. Sample (N=3,485) is restricted to cisgender women and men due to small samples of gender minority respondents. Regression is an ordinary least squares regression. Regression is an ordinary least squares regression. No other differences among response options. Gender expression unipolar items are standardized so mean=0 and standard deviation=1, such that a positive score is more masculine on the masculine scales and more feminine on the feminine scales.

Figure 2. Mean PHQ-9 of Gender Expression of Appearance by Identity, Mturk 2019





Notes. Sample is restricted to cisgender women and men due to small samples of gender minority respondents. In panel 1, 1=very feminine to 7=very masculine. In panels 2 and 3, negative values indicate less of the dimension (less masculine), and positive values indicate more of the dimension (more masculine).

Associations among operationalizations of gender expression

The data collected for this study also offer an opportunity to show the associations among the different measures of gender expression in order to prioritize among them when survey space is at a premium. As shown in Table 6 (top panel), the correlations between appearance and behavior and self and reflected appraisals are very strong among the bipolar items. Similarly, the correlations between appearance and behavior and self and reflected appraisals are very strong within levels of masculinity and femininity for the unipolar items (bottom panel). For example, the correlation is .85 for masculine appearance and behavior (self-appraised), .91 for masculine appearance (self and reflected appraisal), and .84 for appearance-self and behavior-reflected masculine appraisal.

Table 6. Correlations among gender expression items, MTurk 2019

Bipolar items		1	2	3	4				
Appearance-self-bipolar	1	1.00							
Behavior-self-bipolar	2	0.89	1.00						
Appearance-other-bipolar	3	0.94	0.89	1.00					
Behavior-other-bipolar	4	0.87	0.94	0.88	1.00				
Unipolar items		1	2	3	4	5	6	7	8
Appearance-self-masculine	1	1.00							
Appearance-self-feminine	2	-0.67	1.00						
Behavior-self-masculine	3	0.85	-0.63	1.00					
Behavior-self-feminine	4	-0.63	0.84	-0.69	1.00				
Appearance-other-masculine	5	0.91	-0.67	0.84	-0.63	1.00			
Appearance-other-feminine	6	-0.66	0.91	-0.62	0.83	-0.68	1.00		
Behavior-other-masculine	7	0.84	-0.63	0.92	-0.68	0.85	-0.62	1.00	
Behavior-other-feminine	8	-0.63	0.84	-0.68	0.91	-0.64	0.84	-0.69	1.00

Discussion

While much research focuses on the survey measurement of gender identity, a key feature of the variation with which gender is lived, experienced, and has implications for life chances is in terms of one's gender expression: how gender is displayed through appearance and enacted through behavior (GenIUSS 2014; Lorber 1994; NASEM 2022; Spence 2011). This study examines the reliability and validity of the survey measurement of gender expression in a 2-by-5-by-2 factorial experiment administered through MTurk that varies the question order, type of response scale, and the order of gender presentation in the response scale. The results provide some information on how different versions of the gender expression items function in surveys and indicate paths forward for future research and survey practice, as well as add nuance to substantive findings about the association between gender expression and health.

The different versions of the unipolar items—endpoint labeled with "very" at the highest point, endpoint-labeled with "extremely," verbal labels with "very," and verbal labels with "extremely"—did not significantly differ from one another in predicting the distribution of gender expression. This indicates that this feature of question context did not seem to impact the distribution of gender expression, indicating that results from prior studies with differing (unipolar) response scales are comparable.

Two of the four unipolar items showed question context effects in terms of question order (masculine appearance and behavior). In addition, question context effects were found for gender presentation of the response scale, which varies by gender identity for the unipolar items and one bipolar item (behavior). Although Figure 1 shows that these differences are substantively small, the findings with respect to gender presentation context effects varying by gender identity are important for a few reasons. First, at least with respect to cisgender men and women, the results indicate that presenting the masculine (side of the) scale first leads to gender polarization (e.g.,

women report being less masculine and men more masculine when masculine is presented first compared to second). Thus, the version of the question that presents the feminine (side of the) scale first appears to lead cisgender respondents to rate themselves less normatively, which aligns with the idea that respondents considered a broader range of responses before selecting an answer (Bradburn et al. 2004; Garbarski et al. 2015). In addition, the results are important for survey implementation, in that the scale(s) can be presented the same regardless of the respondent's gender, as they are in self-administered questionnaires like the YRBS.[2]

One benefit of the unipolar items is that they show distinctions among the gender minority population in ratings of self-appraised gender expression. In addition, the unipolar items also show more nuance with respect to concurrent validity in predicting health outcomes among cisgender respondents, that is, illustrating particular dimensions of masculinity or femininity as impacting health. Thus, the unipolar items are promising with respect to data collection efforts in general population surveys. General population surveys serve a range of purposes for which all research questions cannot be predicted a priori, such that incorporating a unipolar version of gender expression will allow for a multifaceted assessment of how gender—identity and expression--has implications for well-being and life chances across a variety of domains, similar to what the concurrent validity analyses show. In addition, the bipolar items were conceptually and empirically problematic in cognitive interviews with queer women (Garbarski and LaVergne 2020), such that decoupling masculinity and femininity was recommended among sexual and gender minority populations. The unipolar items would allow these populations to see

[2] Overall, the association between gender expression and gender identity with mental and physical health does not vary depending on which (side of the) scale is presented first (additional analyses available upon request), such that the effects described here are shifts in means of gender expression and do not impact the validity of the gender expression items with respect to mental and physical health.

themselves represented in data collection efforts and lead to consistency across both large-scale general population data collection efforts and surveys focused on particular sociodemographic groups (e.g., LGBTQ-specific studies). Finally, the unipolar items allow for flexibility in operationalizing gender expression, such as including those who are high on both dimensions (masculinity and femininity) or low on both.

The results of this study also guide decisions about which measures to include when survey space is at a premium. The self and reflected appraisals and questions about appearance and behavior are highly correlated, such that including each would be redundant in studies that are interested in gender expression as one dimension of gender and not of primary research interest. With respect to mental and physical health, the validity analyses indicate that intersection between gender identity and reflected appraisals impact mental and physical health more so than the self-appraisals when all are included in the same models (Tables 4-5, Supplementary Appendix Tables A.2-A.5). These findings motivate the continued exploration of dimensions of reflected appraisals, such as appearance and behavior, impacting mental and physical health.

This study focused on examining three features of the survey measurement of gender expression at the exclusion of other features that could be manipulated. Future research should continue to examine differences across features such as: parsing or combining appearance and behavior, order of appearance and behavior, order of the sections of the survey, various definitions of appearance and behavior, temporality and reference periods, mode, visual presentation in self-administered modes, response option ordering, and sociodemographic characteristics. These steps will continue to improve our understanding of both the measurement

of gender expression and nuanced lived experiences beyond the current landscape of presumed dualities and categorical differences.

An important limitation of the current study is that a sample from Amazon Mechanical Turk does not constitute a nationally representative sample of the U.S. However, although non-probability samples generated through methods like MTurk tend to look different from the general population in their sociodemographic characteristics (Antoun et al. 2016), most of the differences in outcomes of interest across probability and nonprobability samples are considerably reduced after controlling for measurable sociodemographic characteristics (Levay, Freese, and Druckman 2016). In addition, systematic differences between a non-probability sample and the population are less problematic for an experiment, given the internal validity of experiments through random assignment of respondents to experimental treatments. Indeed, several studies report that estimated treatment effects in online convenience samples are similar to those observed in probability-based studies (Mullinix et al. 2015; Weinberg, Freese, and McElhattan 2014). Nevertheless, continued research using both probability-based and non-probability-based sampling is necessary.

An additional limitation is that the sample size of transgender and nonbinary respondents is too small to conduct validity analyses (interaction by gender expression and identity) as well as a more thorough exploration of the differences in the distribution of gender expression among gender minority respondents. Future studies are needed to intentionally recruit gender minority respondents in order to fully examine the boundaries and contours of gender expression for transgender and nonbinary respondents.

Conclusion

Although gender expression is a distinct facet of gender and useful for understanding a range of inequalities in well-being by gender, survey measurement of gender expression has received comparably little research focus. This study examines various features of the survey measurement of gender expression: question order, types of response scales, and order of gender presentation in the response scales. The results of the study suggest both benefits and drawbacks to both the bipolar and unipolar items. Based on the results of this study and other conceptual and empirical considerations, the unipolar items are preferable, but the results do not necessarily obviate the need for continued research. Overall, the results of this study suggest certain considerations and possible refinements for measuring gender expression in surveys, with implications for practitioners and researchers who are interested in capturing the full scope of how gender is lived and experienced with respect to health, well-being, and life chances.

Acknowledgments

This research was supported by [blinded]. A previous version of this work was presented at the annual meeting of the American Association for Public Opinion Research. The author would like to thank [blinded] for research assistance. Opinions expressed therein are those of the author.

Appendix A. Gender Expression Experimental Factors

I. Factor 1—Question order

Description that precedes the first question:

"Feminine" and "masculine" are words used to describe qualities traditionally associated with women and men respectively. However, people can have both feminine and masculine features of their appearance and behavior.

A. Self appraisal first

Q1. In general, how would you describe your appearance, style, and dress?

Q2. In general, how do you think people would describe your appearance, style, and dress?

Q3. In general, how would you describe how you walk, talk, sit, stand, and gesture?

Q4. In general, how do you think people would describe how you walk, talk, sit, stand, and gesture?

B. Reflected appraisal first

Q1. In general, how do you think people would describe your appearance, style, and dress?

Q2. In general, how would you describe your appearance, style, and dress?

Q3. In general, how do you think people would describe how you walk, talk, sit, stand, and gesture?

Q4. In general, how would you describe how you walk, talk, sit, stand, and gesture?

II. Factor 2—Response scale type

One of these sets of response scales is randomly assigned to follow each of the four questions listed under Factor 1.

1) Bipolar
   Very feminine, mostly feminine, somewhat feminine, equally feminine and masculine, somewhat masculine, mostly masculine, very masculine

2) Unipolar-Endpoints Labeled-Very

   Not at all feminine 1 2 3 4 5 very feminine

   Not at all masculine 1 2 3 4 5 very masculine

3) Unipolar-Endpoints Labeled-Extremely

   Not at all feminine 1 2 3 4 5 extremely feminine

   Not at all masculine 1 2 3 4 5 extremely masculine

4) Unipolar-Verbal Labels-Extremely

Not at all, a little, somewhat, very, extremely feminine

Not at all, a little, somewhat, very, extremely masculine

5) Unipolar-Verbal Labels-Very

Not at all, a little, somewhat, very feminine

Not at all, a little, somewhat, very masculine

III. Factor 3: Response Scale Order

The third experimental factor varies whether the feminine or masculine (side of the) scale is presented first. For example, under Factor 2, the feminine (side of the) scale is presented first.

References

Aguinis, Herman, Isabel Villamor and Ravi S. Ramani. 2020. "MTurk Research: Review and Recommendations." *Journal of Management* 0(0):0149206320969787. doi: 10.1177/0149206320969787.

Antoun, Christopher, Chan Zhang, Frederick G. Conrad, and Michael F. Schober. 2016. Comparisons of Online Recruitment Strategies for Convenience Samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods* 28(3):231 246.

Austin, S. Bryn, Najat J. Ziyadeh, Jerel P. Calzo, Kendrin R. Sonneville, Grace A. Kennedy, Andrea L. Roberts, Jess Haines, and Emily A. Scherer. 2016. "Gender expression associated with BMI in a prospective cohort study of US adolescents." *Obesity* 24(2): 506-515.

Beckstead, Jason W. 2014. "On Measurements and Their Quality. Paper 4: Verbal Anchors and the Number of Response Options in Rating Scales." *International journal of nursing studies* 51(5):807-14.

Bem, Sandra Lipsitz. 1993. *The Lenses of Gender: Transforming the Debate on Sexual Inequality*. New Haven and London: Yale University Press.

Bird, Chloe E., and Patricia P. Rieker. 2008. *Gender and Health: The Effects of Constrained Choices and Social Policies.* Cambridge University Press.

Bittner, Amanda, and Elizabeth Goodyear-Grant. 2017. "Sex isn't Gender: Reforming Concepts and Measurements in the Study of Public Opinion. *Political Behavior* 39(1):1019–1041 https://doi.org/10.1007/s11109-017-9391-y

Bradburn, Norman M, Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design--for Market Research, Political Polls, and Social and Health Questionnaires*.San Francisco: John Wiley & Sons.

Cassino, Dan. 2020. "Moving Beyond Sex: Measuring Gender Identity in Telephone Surveys." *Survey Practice* 13 (1).

Connell, Raewyn. 2005. *Masculinities.* Cambridge: Polity.

Constantinople, Anne. 1973. "Masculinity-Femininity: An Exception to a Famous Dictum?" *Psychological Bulletin* 80(5):389.

Cooley, Charles H. 1902. *Human Nature and the Social Order*. New York: Scribner's.

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.

Dobson, Keith S., and Kerry J Mothersill. 1979. "Equidistant Categorical Labels for Construction of Likert-Type Scales." *Perceptual and Motor Skills* 49(2):575-80.

Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity (SOGI). 2016a. Current Measures of Sexual Orientation and Gender Identity in Federal Surveys. https://nces.ed.gov/FCSM/pdf/current_measures_20160812.pdf. Accessed 22 December 2018

Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity (SOGI). 2016b. Evaluations of Sexual Orientation and Gender Identity Survey Measures: What Have We Learned? https://nces.ed.gov/FCSM/pdf/Evaluations_of_SOGI_Questions_20160923.pdf. Accessed 22 December 2018.

Garbarski, Dana. 2016. "Research in and prospects for the measurement of health using self-rated health." *Public Opinion Quarterly* 80: 977-997.
Garbarski, Dana, and Dana LaVergne. 2020. "The Measurement of Sexual Attraction and Gender Expression: Cognitive Interviews with Queer Women." Pp. 193-217 In *Understanding Survey Methodology*, edited by Phillip S. Brenner. Springer.
Garbarski, Dana, Nora Cate Schaeffer, and Jennifer Dykema. 2019. The Effects of Features of Survey Measurement on Self-Rated Health: Response Option Order and Scale Orientation. *Applied Research in Quality of Life,14(2),* 545-560. https://ecommons.luc.edu/soc_facpubs/18/
Garbarski, Dana, Nora Cate Schaeffer and Jennifer Dykema. 2015. "The Effects of Response Option Order and Question Order on Self-Rated Health." *Quality of Life Research* 24(6):1443-53.
Gordon, Allegra R., and Ilan H. Meyer. 2007. "Gender nonconformity as a target of prejudice, discrimination, and violence against LGB individuals." *Journal of LGBT health research* 3(3): 55-71.
Hart, Chloe Grace, Aliya Saperstein, Devon Magliozzi, and Laurel Westbrook. 2019. "Gender and health: Beyond binary categorical measurement." *Journal of Health and Social Behavior* 60(1):101-118.
Kroenke, Kurt, and Robert L. Spitzer. 2002. "The PHQ-9: A new depression diagnostic and severity measure." *Psychiatric Annals* 32:509–15.
Krosnick, Jon A., and Stanley Presser. 2010. "Question and Questionnaire Design." Pp. 263-314 in Marsden, Peter V. and James D. Wright, Eds., *Handbook of Survey Research*. Bingley, UK: Emerald Group Publishing.
Levay, Kevin E., Jeremy Freese, and James N. Druckman. 2016. "The Demographic and Political Composition of Mechanical Turk Samples." *SAGE Open* 6(1): 1–17.
Lorber, Judith. 1994. *Paradoxes of Gender*. New Haven and London:Yale University Press.
Lowry, Richard, Michelle M. Johns, Allegra R. Gordon, S. Austin, Leah E. Robin, and Laura. K. Kann. 2018. "Nonconforming Gender Expression and Associated Mental Distress and Substance Use among High School Students." *JAMA Pediatrics*. doi: 10.1001/jamapediatrics.2018.2140.
Magliozzi, Devon, Aliya Saperstein and Laurel Westbrook. 2016. "Scaling Up: Representing Gender Diversity in Survey Research." *Socius: Sociological Research for a Dynamic.* doi: 10.1177/2378023116664352.
Martin, Patricia Y. 2003. "Said and Done" Versus "Saying and Doing" Gendering Practices, Practicing Gender at Work. *Gender & Society* 17(3):342-66.
Mead, George H. 1934. *Mind, Self, and Society*. Chicago: University of Chicago Press.
Miller, Lisa R., and Eric A. Grollman. 2015. "The Social Costs of Gender Nonconformityfor Transgender Adults: Implications for Discrimination and Health." *Sociological Forum* 30(3):809–31.
Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2): 109–138.
National Academies of Sciences, Engineering, and Medicine 2022. *Measuring Sex, Gender Identity, and Sexual Orientation*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26424.

Ridgeway, Cecilia, and Shelley Correll. 2006. “Consensus and the Creation of Status Beliefs.” *Social Forces* 85(1):431–53.
Risman, Barbara J. 2018. *Where the Millennials Will Take Us: A New Generation Wrestles with the Gender Structure*. Oxford: Oxford University Press.
Saperstein, Aliya, and Laurel Westbrook. 2020. “Categorical and Gradational: Alternative Survey Measures of Sex and Gender.” *European Journal of Politics and Gender*
Schaeffer, Nora Cate, and Jennifer Dykema. 2020. "Advances in the Science of Asking Questions." *Annual Review of Sociology* 46:37-60.
SMART. 2009. Best practices for asking questions about sexual orientation on surveys. Created by the Sexual Minority Assessment Research Team (SMART). L. Badgett and N. Goldberg (Eds.). Los Angeles, CA: The Williams Institute. http://williamsinstitute.law.ucla.edu/wp-content/uploads/SMART-FINAL-Nov-2009.pdf.Accessed 19 December 2018.
Smyth, Jolene D., and Kristen Olson. 2020. "Male/Female is not enough: Adding measures of masculinity and femininity to general population surveys." Pp. 247-275 In *Understanding Survey Methodology*, edited by Philip S. Brenner Springer, Cham, 2020.
Spence, Janet T. 2011. “Off with the Old, on with the New.” *Psychology of Women Quarterly* 35(3):504-09. doi: 10.1177/0361684311414826.
Springer, Kristen W., and Dawne M. Mouzon. "“Macho men” and preventive health care: Implications for older men in different social classes." *Journal of Health and Social Behavior* 52, no. 2 (2011): 212-227.
The GenIUSS Group. 2014. Best practices for asking questions to identify transgender and other gender minority respondents on population-based surveys (Created by the Gender Identity in U.S. Surveillance (GenIUSS) Group). J. L. Herman (Ed.). Los Angeles, CA: The Williams Institute. https://williamsinstitute.law.ucla.edu/wp-content/uploads/geniuss-report-sep-2014.pdf. Accessed 22 December 2018.
Tourangeau, Roger, Mick P. Couper and Frederick G. Conrad. (2013). “Up Means Good": The Effect of Screen Position on Evaluative Ratings in Web Surveys. *Public Opinion Quarterly* 77(S1),69-88. doi: 10.1093/poq/nfs063.
Tourangeau, Roger., Rips, Lance J., and Kenneth Rasinski. 2000. *The Psychology of Survey \ Response*. Cambridge: Cambridge University Press.
Weinberg, Jill, Jeremy Freese, and David McElhattan. 2014. “Comparing Data Characteristics and Results of an Online Factorial Survey Between a Population-Based and a Crowdsource-Recruited Sample. *Sociological Science* 1(19):292-310.
West, Candace. and Don H. Zimmerman. 1987. "Doing Gender." *Gender & Society* 1(2),25-51.
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. doi: 10.1177/0891243215584758.
Westbrook, Laurel, and Kristen Schilt. 2014. “Doing Gender, Determining Gender: Transgender People, Gender Panics, and the Maintenance of the Sex/Gender/Sexuality System.” *Gender &Society* 28(1):32–57.
Zucker, Kenneth J., Janet N. Mitchell, Susan J. Bradley, Jan Tkachuk, James M. Cantor, and Sara M. Allin. 2006. “The Recalled Childhood Gender Identity/Gender Role Questionnaire: Psychometric Properties.” *Sex Roles* 54(7):469-83. doi: 10.1007/s11199-006-9019-x.