Published in final edited form as:

Sex Roles. 2017 June ; 76(11): 731–746. doi:10.1007/s11199-016-0657-3.

Description and Evaluation of a Measurement Technique for Assessment of Performing Gender

Paul J. Fleming^{1,2}, Kathleen Mullan Harris^{1,3}, and Carolyn Tucker Halpern^{1,4}

¹Carolina Population Center, University of North Carolina, Chapel Hill, NC, USA

²Department of Health Behavior & Health Education, University of Michigan, Ann Arbor, MI, USA

³Department of Sociology, University of North Carolina, Chapel Hill, NC, USA

⁴Department of Maternal and Child Health, University of North Carolina, Chapel Hill, NC, USA

Abstract

The influence of masculinity and femininity on behaviors and outcomes has been extensively studied in social science research using various measurement strategies. In the present paper, we describe and evaluate a measurement technique that uses existing survey items to capture the extent to which an individual behaves similarly to their same-gender peers. We use data from the first four waves of The National Longitudinal Study of Adolescent to Adult Health (Add Health), a nationally representative sample of adolescents (age 12–18) in the United States who were re-interviewed at ages 13–19, 18–26, and 24–32. We estimate split-half reliability and provide evidence that supports the validity of this measurement technique. We demonstrate that the resulting measure does not perform as a trait measure and is associated with involvement in violent fights, a pattern consistent with theory and empirical findings. This measurement technique represents a novel approach for gender researchers with the potential for expanding our current knowledge base.

Keywords

Measurement; Gender; Reliability; Validity; Add health; Longitudinal

In Sandra Lipsitz Bem's (1974) foundational paper, based on a sample of university students in the United States, she argued that masculinity and femininity are not oppositional but rather are two independent dimensions. Her work to develop a measure of "psychological androgyny" helped to usher in a new wave of theory and research about how gender is conceptualized and how it shapes individuals' lives (Taylor and Hall 1982). Since the publication of the Bem Sex Role Inventory (BSRI) in the *Journal of Clinical and Consulting Psychology* (Bem 1974), theoretical understandings of gender, including Bem's (1995) own

Correspondence to: Paul J. Fleming.

Compliance with Ethical Standards

Conflict of Interest The authors have no conflicts of interest to declare.

Human/Animals Rights We conducted research with secondary data collected from human participants.

Informed Consent The data we use comes from a dataset where the participants gave informed consent.

understanding, evolved away from viewing masculinity and femininity as static individual traits that characterize what an individual *is* to approaching gender as something fluid that individuals *do* (West and Zimmerman 1987).

Informed by the theoretical concepts of *doing gender* (West and Zimmerman 1987) and *gender performativity* (Butler 1990, 1993, 1997), we aim to describe and evaluate a measurement technique for gender research that focuses on individuals' performance of gender, rather than on personality traits or ideologies. Specifically, this technique measures the extent to which an individual behaves similarly to their same-gender peers. To evaluate the measurement approach, we draw upon several standard techniques to assess the reliability and validity of a measure (DeVellis 2012). Specifically, we (a) test our measure's reliability using split-half samples, (b) assess its content validity by examining its distribution and selected items, and (c) evaluate its construct validity by examining its longitudinal performance and association with violent fighting. The approach presented in the present paper represents a novel methodological tool to study gender performance and its associations with socio-behavioral outcomes in the United States.

The measurement technique that we describe is useful for gender researchers aiming to explore questions such as (a) how an individual's gender performance (i.e., the extent to which an individual behaves similarly to their same-gender peers) is correlated with other gender-related measures such as ideology or gender role conflict and (b) whether health and social outcomes known to be associated with binary gender (i.e., male vs. female) are actually associated with how an individual performs gender. Additionally, any researcher with access to socio-behavioral data could potentially use this measurement technique because the technique, by nature, captures the developmental, historical, and cultural context embodied in the data.

Bem's (1974, 1977) work on masculinity and femininity in the United States helped to establish the idea that masculinity and femininity are not mutually exclusive and are not dependent on one's biological sex. A key concept from Bem's work is that an individual's degree of masculinity is not dependent on their degree of femininity, and some individuals may be high on both dimensions (i.e., androgynous). The BSRI, as well as other similarly conceived measures such as the Personal Attributes Questionnaire (Spence and Helmreich 1978), assigns individuals both a femininity and a masculinity score based on personality characteristics that were perceived at a particular point in time (1970s) and in a particular cultural setting (e.g., U.S. universities) to be qualities more typical of males or more typical of females (Bem 1974). In the Bem Sex Role Inventory, individuals are asked to assess on a 7-point scale to what extent each of 60 personality characteristics describes them. Example characteristics include competitive (masculine), tender (feminine), cheerful (feminine), and analytical (masculine) (Bem 1974).

Theoretical conceptualization of masculinity and femininity has evolved substantially over the past decades. Although the BSRI is an important milestone in gender measurement and has been widely used (the original article has been referenced over 7000 times and continues to be used; e.g., Bos et al. 2012; Choi et al. 2009; Colley et al. 2009), many have questioned the contemporary utility of the BSRI given this theoretical change (Luyt 2015; Pleck 1981;

Holt and Ellis 1998; Leszczynski and Strough 2008; Twenge 1997). First, some have argued that a trait measure such as the BSRI implies that one's degree of masculinity or femininity is inherent and static rather than something that is fluid and shaped by changing social contexts (Luyt 2015; Pleck 1981). Second, because notions of masculinity and femininity vary by time and culture, an unchanging set of questions that distinguishes men and women at one point in time (e.g., 1974), or at a given age (e.g., young adults), or for a given population (e.g., U.S. university students) may not produce reliable or valid distinctions at other times and places, or for other populations (Holt and Ellis 1998; Leszczynski and Strough 2008; Strough et al. 2007; Twenge 1997). Third, though the assumption that an individual's traits can be both masculine and feminine does reflect the true complexity in males' and females' personalities, dualistic trait measures fail to acknowledge and capture the relational aspect of gender (Connell 1995). As prominent gender scholar Connell (1995, pp. 43–44) writes,

Masculinity and femininity are inherently relational concepts, which have meaning in relation to each other, as a social demarcation and a cultural opposition. This holds regardless of the changing content of the demarcation in different societies and periods of history ... It is gender-relations that constitute a coherent object of knowledge for science.

Thus, treating masculinity and femininity as comprised of distinct, independent sets of temporally static traits misses the point of examining the relational aspect of gender. Instead, measures for gender research should be examining this “social demarcation” with measures that are dynamic to “different societies and periods of history” (Connell 1995, pp. 43–44; also Luyt 2015).

Prominent gender theorists across the globe—including in the United States (Butler 1990; West and Zimmerman 1987), sub-Saharan Africa (Morrell et al. 2012), Australia (Connell 1995), and Latin America and the Caribbean (de Moya 2003; Gutmann 2006)—have transitioned from understanding gender as something that an individual *is* to something that individuals *do*. The distinction moves away from gender as a static set of traits to constellations of actions of individuals that change over time, and importantly, the social environments that ascribe meaning to those actions. As West and Zimmerman (1987, p. 126) write in their foundational paper on doing gender:

We contend that the “doing” of gender is undertaken by women and men whose competence as members of society is hostage to its production. Doing gender involves a complex of socially guided perceptual, interactional, and micropolitical activities that cast particular pursuits as expressions of masculine and feminine “natures.”

The idea is that men and women are motivated to act in a certain way to meet the social and cultural norms of appropriate and acceptable male and female behavior in a given society. This conceptualization of gender prioritizes an individual's behavior and how it compares to the behavior of other individuals in the society.

Butler, a U.S.-based philosopher and gender theorist, built upon the concept of *doing gender* by developing the idea of *gender performativity* (1990, 1993, 1997). Butler (1990, 1993,

1997) hypothesized that individuals construct a coherent gender identity through their repeated behaviors and interactions. Butler (1990, 1993, 1997) also writes about the *audience*, an individual's social environment that is the witness to one's performance. The audience seeks a coherent set of behaviors by which to easily characterize an individual (Butler 1990, 1993, 1997). Therefore, although an individual may not always coherently perform masculine or feminine behaviors, if he/she repeatedly performs certain behaviors that are concordant with dominant representations of a certain gender, she/he can more readily be characterized as masculine or feminine (Butler 1990, 1993, 1997). To learn how to perform, an individual observes the performances of other males and females, as well as reactions of the audience who are in their same social environment, to determine which behaviors are appropriate for males or females (Butler 1990, 1993, 1997).

Researchers have deployed a variety of techniques over the past century to measure different gender-related constructs (e.g., personality trait, gender ideology, gender role conflict). (For a review of such approaches, see Smiler and Epstein 2010.) But, only a relatively small body of work by gender researchers uses measures that are based in the theoretical notions of *doing gender or gender performativity* (Butler 1990, 1993, 1997; Connell 1995; Courtenay 2000; West and Zimmerman 1987). In a recent critique of gender research that focuses on this issue, Luyt (2015, p. 207) highlighted that gender "is understood as a social practice, made meaningful through shared understanding of dominant gender representation." Also, Tate et al. (2014) advocate conceptualizing gender identity as a spectrum with various types of behavioral expressions, rather than simply as a binary. Luyt (2015, p. 207) argues that measuring an individual's "social practice" (e.g., behaviors) and comparing that practice to the "dominant gender representation" (e.g., the most common social practices of others within the social environment) is a useful measurement strategy for quantitative gender research. In the following we review the measurement strategies used by gender researchers that are linked to the idea of doing gender.

Egan and Perry's (2001, p. 455) measure of gender identity with U.S. school children asked participants to assess their own *gender typicality* (i.e. "feeling that one's skills or interests are similar to those of same-sex others"), but this was a self-perception rather than a behavioral measure. The Conformity to Masculine Norms Inventory (Mahalik et al. 2003, p. 5) and Conformity to Feminine Norms Inventory (Mahalik et al. 2005) examine gendered behaviors in comparison to dominant ideals in that they assess the extent to which an individual "conforms or does not conform to the actions, thoughts, and feelings" of masculinity or femininity in the White, middle/upper-class heterosexual culture in the United States, as determined by a team of researchers who conducted a thorough literature review and focus groups with U.S. graduate students. These scales have been applied to various populations throughout the United States (Mahalik et al. 2007; Parent and Moradi 2010; Smiler 2006), but have the limitation of having a large number of survey items and having potentially limited applicability outside the cultural context specified (Mahalik et al. 2006; Smiler and Epstein 2010). Other scales that measure gender performance have tended to focus on diagnosing children who have non-normative behaviors for their gender with a psychological disorder (Smiler and Epstein 2010).

Another measurement strategy based on doing gender—and the focus of our paper—is to compare participants' reports of their behaviors to the responses of their same-gender peers within the same sample. Cleveland et al. (2001) used this strategy with data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a nationally representative longitudinal study of over 20,000 adolescents in grades 7–12 in the United States in 1994–95. (Note that Cleveland and colleagues drew upon by foundational work by Lippa and Connelly (1990), to measure *gender diagnosticity* based on U.S. college students' occupational preferences.) Using behavioral and attitudinal data from the first interview wave in 1995, they identified 16 survey items with the highest discrimination between males and females, and they generated predicted probabilities of whether each respondent was a male or female based on their responses to the 16 items. The resulting probability thus measured the extent to which an individual's reported behaviors were congruent with the reported behaviors of other male or female research participants. (They called this *sex-typed behaviors*).

Rather than a researcher proposing which behaviors were suitable or typical for men and women, Cleveland et al. (2001) used data from a nationally representative sample of 20,000 male and female adolescents to determine which behaviors actually discriminate between males and females. One advantage of this method, compared to using a scale like the Conformity to Masculine Norms Inventory (Mahalik et al. 2003), is that it does not reify norms of gendered behaviors that are anchored in a certain place and time when measures were developed, but rather captures individuals' behaviors relative to their peers who can be explicitly described in terms of sociodemographic, historical, and cultural characteristics. Although the Cleveland et al.'s measure only reflected one domain of gender performance (i.e., the extent to which an individual behaves similarly to their same-gender peers), it avoided imposing a preordained idea of what men's and women's behaviors *should* be. Depending on the intended purpose, a potential limitation of such a measurement technique is that it is based on the gender binary, and an individual's score is always relative to the sample population from which it is derived. This poses certain problems if the sample population used for measurement development includes gender identities beyond the gender binary or is not representative of the larger population to which a researcher wishes to generalize.

This empirically derived measure has been used in several studies demonstrating its sensitivity to environmental influence and its relevance for social and health outcomes in the United States (Cleveland et al. 2001; Udry and Chantala 2004, 2006). More recently, Mahalik et al. (2015) used this measure to determine that behaving similarly to male peers increased men's and women's risk for substance use in the United States. Taken together, this previous research demonstrates the potential predictive utility of empirically derived constructs based on behaviors that are important health and social outcomes.

We build on this previous research in several ways. First, we present a more thorough description of the methodological technique and present analyses demonstrating reliability and validity of this empirically derived measure of gender, whereas Cleveland et al. (2001); Udry and Chantala (2004, 2006), and Mahalik et al. (2015) did not. Second, for evaluation and descriptive purposes, we apply and assess the measurement technique at all four waves

of Add Health data, rather than at a single wave as was done by Cleveland et al., Udry and Chantala, and Mahalik et al., extending measurement from adolescence into adulthood and enabling us to observe developmental change in the measure. Third, we frame this measurement approach using the more recent theoretical concepts of *doing gender* and *gender performativity* (Butler 1990, 1993, 1997; Luyt 2015; West and Zimmerman 1987), which enables us to examine how gendered behaviors are associated with psychosocial and health outcomes across developmental periods.

Previous empirical literature and theory based on U.S. research suggests that violence perpetration is an important example of a behavioral expression of gender and, thus, in our paper we use violence perpetration to test the validity of this measure (Courtenay 2000; Fleming et al. 2015; Mahalik et al. 2003). In the United States, because being involved in a fight is more often a characteristic associated with men, it has been theorized that men use violent behaviors as a way to express a masculine gender identity (Courtenay 2000; Fleming et al. 2015). Women may perform certain behaviors to demonstrate a feminine identity but being involved in physical fights is not typically associated with female behavior. In fact, “Violence” is one of the major constructs within the Conformity to Masculine Norms Inventory (Mahalik et al. 2003), but it is not represented within the Conformity to Feminine Norms Inventory (Mahalik et al. 2005). Similarly, results from the Conformity to Masculine Norms Inventory among U.S. samples suggest that greater conformity to male-typical behaviors is associated with perpetrating violence for U.S. college males (Cohn and Zeichner 2006; Locke and Mahalik 2005) and a sample of men recruited from the internet of unknown country origin (Mahalik et al. 2007). Unfortunately, similar research using the Conformity to Feminine Norms Inventory has rarely explored fighting among U.S. women.

In the present paper, we aim to (a) describe the application of this measurement technique to existing data from Add Health at each wave of data collection (representing different developmental periods) to generate time-varying measures of gender performance, (b) assess the reliability of the derived measures, and (c) evaluate the validity of the derived measures. We first describe the steps taken to create this measure using existing items from Add Health. Then, we present findings from our assessment of the measure’s reliability by showing the creation of the measure using split-half samples and comparing the resulting split-half measures to the full-sample measure. Finally, we demonstrate the validity of this measure by presenting results that show that the measure performs as expected and is associated with violence perpetration. Specifically, we hypothesize that because *doing gender* is fluid (i.e., not a stable trait), correlations between longitudinally collected measures of doing gender across developmental periods will vary. We also present correlations between the measure at each wave and involvement in violent fights to demonstrate the construct validity of our measure because engaging in physical violence (or not) is considered a part of men’s (women’s) gender performance in the United States (Courtenay 2000; Fleming et al. 2015).

Method

Data

We use data from all four waves of Add Health (National Longitudinal Study of Adolescent to Adult Health). In 1994–95 (Wave 1), Add Health recruited a school-based sample representative of U.S. adolescents in grades 7–12 in the 1994–1995 school year; this initial sample included 20,745 adolescents. Add Health used a cluster sampling design at Wave 1 to randomly select schools and then randomly select individuals from within schools. The Wave 1 sample was followed up for surveys in 1996 (Wave 2, $n = 14,738$), 2001–02 (Wave 3, $n = 15,197$), and 2008–09 (Wave 4, $n = 15,701$). (Note that Wave 2 is slightly smaller due to the exclusion by design of Wave 1 participants who had graduated high school; Waves 3 and 4 attempted to follow all Wave 1 participants with response rates of 77 % and 80 %, respectively.) At Wave 4, the participants were between the ages of 24 and 32 years-old. For the analyses presented here, we use only those participants who were interviewed at all four waves and have the measure of gender performance for at least one wave ($n = 9417$).

All interviews were conducted in each participant's home with a 90-min CAPI/CASI survey on demographics, school performance, attitudes and beliefs, physical and emotional health, sexual and romantic relationships, and risky sexual and substance use behaviors, among others. Although Add Health collects biomarker data, our study only uses the survey data, including demographic, attitudinal, and behavioral items. For more details on the Add Health study design, see Harris (2013).

Description of Measurement Technique

The present measurement technique assesses the degree to which males and females behave in ways that are similar to the behaviors of their same-gender peers. We define behaviors broadly: they include both actions performed by the individual (e.g., exercising) and states of being (e.g., getting sad). How a respondent presents themselves in a survey is similarly part of *doing gender*; thus, responses to questions about “getting sad,” or similar states of being questions, are behavioral expressions of their gender (Jachyra et al. 2014). To apply this technique at each interview wave, we use a multi-step process similar to the one described by Cleveland et al. (2001). Our measure differs from Cleveland et al.'s measure in that it draws from a larger pool of survey items, is developed independently at each of the four waves, and is interpreted as a measure of having behaviors consistent with same-gender peers rather than with Masculinity/Femininity.

For this technique, we identified the variables that best discriminate between males and females and then used those variables in a logistic regression model to create a predicted probability of being a male or female for each individual. We show the procedure we followed in Fig. 1. First, for each of the four waves, we began with a list of all variables derived from the complete set of questions asked of the participant during the in-home survey. Then, we excluded variables that had more than 300 missing observations (where most were due to survey skip patterns). This number was selected iteratively to ensure a maximum number of variables were included without excluding too many participants from analyses. We next excluded variables that were (a) generally only possible for only one

gender (e.g., menstruation or playing football), (b) demographic (e.g., race), or (c) passive (e.g., reports about how a friend or parent behaves). For the remaining variables, we conducted t-tests for continuous variables and chi-square tests for categorical variables to compare means and frequencies, respectively, between males and females. We then ranked the p-values to identify the approximately 50 variables that had the most statistically significant differences between males and females. Using this set of variables as predictors, we modeled the variable “sex” (1 = male, 2 = female) using logistic regression.

$$\text{logit}(\theta) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_{50} x_{50}$$

where θ is the dichotomous variable *sex* and each x is one of the 50 final variables.

Using manual backwards stepwise regression, we removed variables one-by-one that explained the least amount of variance in sex. We removed variables until the remaining variables were significant at $p < .001$. Using this criterion for each wave, 25 variables remained at Wave 1, 27 variables remained at Wave 2, 23 variables remained at Wave 3, and 22 variables remained at Wave 4. For each wave, we used the final set of 20+ variables and the same model above to create predicted probabilities for each participant. This produced a model-based predicted probability that the individual observation is a *male*. A predicted probability score of .99 would indicate a .99 probability that the individual is a male (and 1 % chance they are female), whereas a score of .01 would indicate a 1 % chance of being a male (and a 99 % chance of being a female).

As we will demonstrate in the Results section, the resulting measure is highly skewed. Therefore, to account for skewness we created a rank measure for males and females at each wave. We ranked males and females separately at each wave based on their probability score such that a rank of 1 indicated the individual with the highest probability score (i.e., for males, this indicates the highest degree of similarity to other males’ behaviors; for females, the highest degree of similarity to other females’ behaviors). Subsequently, we divided each male’s rank at each wave by the number of males with valid sampling weights at that respective wave and then did the same for females. We subtracted the number from 1 in order to end up with a percentile where a higher value indicates greater similarity to the behaviors of male peers (for males) or greater similarity to the behaviors of females (for females).

For example, a male with a probability score of .95 at Wave 1, which suggests high similarity to other males’ behaviors during adolescence, would, when ranked in relation to all his peers’ scores, receive a rank of 1000 (out of 7000 total individuals). We then calculated his percentile: $1 - (1000/7000) = .85$. This means that this individual is in the 85th percentile for similarity to other males’ behaviors. The same process is followed for females; a percentile score of .95 would indicate a high degree of similarity to other females’ behaviors and .05 would be a low degree of similarity to females’ behaviors. The resulting percentile measure has a mean of .50 for both males and females. Because the percentile is better suited for bivariate and multivariate analyses, given that it accounts for the skewness of the probability measure, we use the percentile to conduct the construct validity checks.

Analytic Tests Used to Assess Percentile Measure

All of our analyses, including the creation of the probability and percentile measures, were conducted using SAS version 9.4. We use several strategies to assess the reliability and validity of the probability and percentile measure, specific analyses and hypotheses are described alongside results in the next section.

Results

See Table 1 for demographic characteristics of our analytic sample. The demographic characteristics are similar to overall characteristics of the samples at each wave (Harris 2009). Each item selected for our measure at each wave is presented in Table 2.

Reliability and Validity of Probability Measure

We use multiple methods to assess the reliability and validity of our measure. Test-retest reliability of a measure is the extent to which a measure “yields the same results on repeated trials” (Carmines and Woods 2005a, p. 361). To assess the reliability of this measure, we used the *surveyselect* procedure in SAS 9.4 to randomly split the nationally representative sample into two halves at each wave and created the probability measure in split samples. We compared the items selected and the distribution of the probability measure from the full sample at each wave to the probability measure created from split-half samples at each wave. We hypothesized that there would be no major differences between the full sample measures and split-half sample measures for items selected and distributions. To create the measure in each of the eight split-half samples (two samples at each of the four waves), we performed the same steps described in the Method section that we used with the full sample. We stopped eliminating variables in the backwards stepwise regression when the number of independent variables matched the number from the full sample (25 for Wave 1, 27 for Wave 2, 23 for Wave 3, and 22 for Wave 4). Table 3 indicates that the measures created with the split-half samples were nearly identical to the measure created with the full samples. Comparing the final lists of selected variables from the split-half samples for the measure to the full sample, there were between zero and three different variables selected. In the cases where there were differences, analyses in the split-half samples had not selected the variables included in the full sample measure that explained the smallest amount of variance in sex (see Table 3). The greatest difference between the mean in the split-half samples for males and females and the mean in the full sample was .01.

Figure 2 illustrates the similarity between the measure in the Wave 1 split-half samples and the full sample. (Note that graphs for Waves 2, 3, and 4 looked nearly identical, available upon request.) The graph demonstrates that the measure has nearly identical distributions when comparing the split samples to the full sample at each wave. The consistency between the split-half samples and the full sample indicates that this measure yields nearly identical results “on repeated trials” (Carmines and Woods 2005a, p. 361). It also indicates that the survey items that make up the measure are not selected by chance, but rather tap into a systematic pattern that is reliable across sub-samples.

Construct validity of a measure is the extent to which “an indicator of some abstract concept measures what it purports to measure” (Carmines and Woods 2005b, p. 933). First, we assessed construct validity by examining the distribution and mean at each wave to determine if the measure was performing as expected. Theoretical perspectives of *doing gender* would suggest that females perform similarly to other females and males perform similarly to other males (Butler 1990; West and Zimmerman 1987). Thus, females should be clustered together on our measure and males should be clustered together. We found, as hypothesized, that the distribution for males is left-skewed and for females is right skewed at each wave, as seen in Fig. 2. The majority of males and females are concentrated near the poles, with decreasing numbers of individuals having a probability closer to the other gender. Additionally, we found that the mean probability at each wave was approximately .75 for males and .25 for females, as seen in Table 4. This indicates that the measure is performing as expected given that males have a higher probability of being male than females and the averages tend to be roughly equidistant from the total mean of .50.

Assessing content validity requires determining whether our measure represents the intended concept accurately and fully (Carmines and Woods 2005b). The intended concept is to obtain a measure of the extent to which an individual behaves similarly to their same-gender peers. Our methodological process—constraining possible items to only be behavioral items and selecting items that were most significantly discriminant between males and females—identified the available behavioral variables that were most significantly different between genders. A limitation of this approach is that we were limited by the range of variables available in Add Health and that the selected variables may have been spuriously selected rather than an actual signifier of one’s gender. However, we reviewed each of the 95 items individually that was selected to confirm that responses could plausibly be an individual *doing gender*. We concluded that each item could indeed be an expression of one’s gender. For example, females reported higher “Frequency of sunbathing in the summer” than did males, and this is a behavior that could communicate information about one’s gender performance to members of their social environment.

One way to assess construct validity is to assess whether the measure performs as expected by determining the extent to which our measure is associated with other concepts that are theorized to have a relationship with our measure (convergent validity) and is not associated with concepts it is not theorized to have a relationship with (discriminant validity) (Carmines and Woods 2005b). Given that this measurement technique produces a broad measure of how an individual is doing gender (i.e., the extent to which their behaviors are similar to their same-gender peers), it presents challenges for identifying constructs to test convergent and discriminant validity because *doing gender* shapes almost every facet of life. Thus, we would expect the resulting measure to be associated with a wide range of outcomes, which presents challenges for assessing convergent validity.

In an attempt to demonstrate discriminant validity, we examined correlations between the measures at each wave to ensure our measure was not performing like a trait. If our measure were capturing a trait, the correlations between the measures at each wave would be very high and stable due to the fact that traits should not change over time (Cobb-Clark and Schurer 2012). We hypothesized that our measure is *not* a trait and thus should vary across

developmental periods (i.e., measures at each wave should *not* be highly correlated with each other). Correlations between probability scores at each wave were moderate when using the whole sample of males and females; the Pearson correlation coefficient between the probability score at Wave 1 and Wave 2 is $r(8920) = .65$, between Waves 1 and 3 is $r(8620) = .56$, and between Waves 1 and 4 is $r(8614) = .53$. But, when the correlations are computed only among males, or only among females, we see in Table 4 that they are low. Moreover, among both males and females, the correlation coefficient is lower as the time between waves is longer. These results support discriminant validity because they demonstrate that what the measure is capturing is not a trait that is static over time, but rather is a performance of gender that changes across developmental periods.

To assess one aspect of convergent validity, we tested the measure's associations with being involved in a violent fight in the past 12 months. Based on this prior research (Cohn and Zeichner 2006; Locke and Mahalik 2005; Mahalik et al. 2007) and theory that proposes that men perform violent behaviors as part of their expression of a masculine gender identity but not necessarily an expression of femininity (Courtenay 2000; Fleming et al. 2015; Mahalik et al. 2003, 2005), we hypothesized that there would be cross-sectional, significant associations between being involved in fights and our percentile measure for men and a significant but weaker correlation for women. As such, men who behave most similarly to their male peers will demonstrate the highest likelihood of having been involved in a fight in the past 12 months and men who behave dissimilarly to male peers will be less likely to be involved in a fight. For women, those who behave less similarly to female peers will be most likely to report engaging in fights but those who behave most similarly to female peers will be less likely to engage in fights.

We examined the bivariate association at each wave between the percentile measure and being involved in a serious fight (dichotomous). Being involved in a serious fight was measured by asking, "In the past 12 months, how often did you get into a serious physical fight?" Response options were "never," "once," and "more than once," and we created a dichotomous measure where the two categories were 0 (*never*) and 1 (*once or more than once*). This variable was selected from the available variables in the existing data. We used bivariate logistic regression to examine the associations with being involved in a serious fight as the dependent variables and the percentile measure as the independent variable; all analyses used survey commands in SAS 9.4 to account for the complex survey sampling design. The fighting variable was ineligible for selection for our measure because there were more than 300 missing observations due to design skips—one of our exclusion criteria. (Note that fighting was not measured at Wave 3.)

Table 5 shows that, at each wave (W) with the fighting variable, compared to men who had not been in a fight in the past 12 months, men who had been in a fight had a significantly higher mean percentile at each wave 6 (Wave 1: .61 vs .41, $t(128) = 18.92$, $p < .0001$; Wave 2: .60 vs .47, $t(128) = 10.88$, $p < .0001$; Wave 4: .63 vs .50, $t(128) = 6.27$, $p < .0001$). Thus, the more similar male's behaviors were to their male peers, the more likely they were to have engaged in fighting within the past 12 months. Women who reported participating in a serious fight in the past 12 months had a lower mean percentile than those who did not report being in a fight (Wave 1: .42 vs .53, $t(128) = -7.57$, $p < .0001$; Wave 2: .45 vs .53,

$t(128) = -5.40, p < .0001$; Wave 4: .43 vs. .51, $t(128) = -2.20, p = .0300$). Therefore, women who fought behaved less similarly to female peers than did women who have not been in a serious fight. Overall, we see support for our hypothesis that an individual's overall gender performance (i.e. similarity to same-gender peers' behaviors) is associated with their likelihood of fighting.

Discussion

We described and evaluated a measurement technique that assesses the extent to which an individual behaves similarly to their same-gender peers. We provided evidence that it is both reliable and valid. We demonstrated the reliability by performing the same procedures on split-half samples for each wave, which resulted in measures nearly identical to the measure derived from the full sample. We also found support for the validity of the measure. The skewed distribution for both males and females, low correlations between waves, and different means for males and females at each wave all suggest that the measure performs as expected. Finally, we also correlated our measure separately for males and females with being involved in a serious fight which previous literature suggests is an expressions of gender in the United States, or *doing gender* (Courtenay 2000; Fleming et al. 2015; Mahalik et al. 2003). We found, as expected, that our measure did not perform as a trait (discriminant validity). Rather, males who behaved more similarly to their male peers, as well as females who behaved less similarly to their female peers, were more likely to engage in fighting (convergent validity).

The evidence for the reliability and validity of this measurement technique opens up new possibilities for research to explain how *doing gender* patterns a variety of important health and social outcomes throughout stages of the life course. For example, when we observe global differences in men's and women's smoking habits (Jha et al. 2002), our measure will allow us to better understand smoking behaviors by exploring within-gender differences (e.g., Are men who behave similarly to other men more likely to smoke than men who behave dissimilarly?), between gender differences (e.g., Are women who report behaviors that are overall less similar to women's behaviors as likely to smoke as men?), and change over time (e.g., Do changes in how men *do gender* correspond to changes in smoking patterns?). Such research will help build upon the foundation of gender research that Sandra Bem pioneered.

Our measure uses existing items from Add Health. By excluding variables that were (a) generally unique to one gender, (b) demographic, or (c) passive, we were able to use the method to select items that best discriminated between men and women to create our measure. The items selected represent aspects of individuals' behaviors or preferences, and they can be considered to be reflections of how they represent their gender in their interactions with their social environment (Luyt 2015). Given the conceptualization of gender as a series of behaviors and states of being (Butler 1993; Connell 1995; Courtenay 2000; West and Zimmerman 1987), these items allow us to obtain a quantitative measure of *doing gender* that can be used in bivariate and multivariate analyses.

This measurement technique could potentially be applied to any large-scale dataset with a diverse set of behavioral questions. Upon applying this technique, the resulting measure would always compare the individual's behaviors to the behaviors in the sample (in our case, a nationally representative sample of U.S. adolescents in 1994–95 who are followed into young adulthood). A strength of this measurement technique is that it is sufficiently dynamic to accommodate a change in reference group to different contexts. For example, a measure could be created for each region of the United States (e.g., Northeast, Midwest, South, West) or for each ethno-racial group (e.g., White/Caucasian, African-American/Black, Native American/Pacific Islander, Asian, and Hispanic). Or, using the European Social Survey, researchers could develop a measure for each participating European country that compares individuals to nationally representative populations from their country. Applications of this method in other datasets should ideally use sample populations that are representative of the larger population being studied.

This measurement approach may have limited applications to specific sub-samples of gender non-conforming populations. Additionally, researchers would need to conceptually consider which sample or sub-sample populations would be appropriate reference groups (e.g., it may be inappropriate to use a multi-national sample or a sample of individuals with a wide age range as the referent group). In our case, the age ranges captured developmental periods (e.g., adolescence, emerging adulthood, and young adulthood), but it would be up to each researcher to determine which age range is appropriate for their use of the measure. Any use of this method with other populations or datasets would still need to verify the measure's reliability and validity, and may need to assess the construct validity differently based on the specific cultural context and historical time. This future research could add depth and nuance to our understanding of how gender influences socio-behavioral outcomes in diverse contexts and historical periods.

Our results demonstrate that the measure varies over developmental periods, suggesting that trait methods for measuring gender may be inappropriate because changing environments and life experiences impact gender expressions. Measuring gender ideology and gender-role stress are useful for understanding the extent to which societal gender norms are internalized by individuals, and the measurement technique we present has the potential to provide an additional dimension by which to examine individuals' performance of their gender (Luyt 2015). Additionally, because scales commonly used in gender research, such as the BSRI, are infrequently utilized in large datasets, creating a measure using existing data allows for gender research on various health and social outcomes in nationally representative datasets.

The measurement technique we describe is unique as an individual measure for gender research in that it taps into the social constructionist perspective that individuals are *doing gender* in their daily lives (West and Zimmerman 1987, p. 125). Although we see behavioral differences in the aggregate between males and females, the approach provides support for the gender similarities model (Hyde 2005) because it demonstrates that there is substantial overlap between the behaviors of individual males and individual females. Additionally, by focusing on gender as a set of behaviors, actions, and states of being, our measure avoids *gender essentialism* (i.e., gender is a natural trait for men and women; Bohan 1993; Courtenay 2000; Luyt 2015; O'Neil 2010). The differences between men and women are

primarily influenced by a changing system of gender norms that creates different behavioral norms for men and women (Bem 1995; Connell 1987; Courtenay 2000). Therefore, when we examine the extent to which an individual behaves similarly to his/her same-gender peers, we are observing the effect of gender norms as a social structure within which the individual behaves.

Our findings are specific to the U.S. context of adolescents and young adults. We found that the relationship between the percentile measure and violent behavior was stronger for males than for females. Stronger results for men are likely due to the fact that men are socially compelled to act in ways that are perceived as masculine. Vandello and Bosson (2013, p. 101) call this *precarious manhood*; they found among U.S. men that manhood is “hard won, easily lost, and that requires continual public demonstrations of proof,” such as violence perpetration (also see Vandello et al. 2008). Although there has been some loosening of the strict definition of masculinity, most men in the United States, and across cultural contexts, are still socially pressured to act masculine by demonstrating their strength, often through shows of force or violence (Barnes et al. 2012; Connell 1995; Courtenay 2000; Gilmore 1990). Men who behave dissimilarly from other men may be less concerned with portraying a masculine outward identity. Therefore, they may feel less of a need to prove their masculine status through engagement in fights. Men who do behave similarly to other males may be less willing to accept challenges to their masculinity and therefore are more willing to engage in violent behaviors to demonstrate their masculinity to peers (Vandello and Bosson 2013). Although further exploration is needed, it seems that women who behave dissimilarly to female peers than female peers may pick up on these male dynamics and are thus more likely to perpetrate violence. Though similarities exist across cultural contexts (Gilmore 1990; Mahalik et al. 2006), more research would be needed to examine whether or not these same dynamics and relationships apply to non-U.S. settings.

In conclusion, measurement technique focused on the theoretical concept of doing gender has the potential to give gender researchers leverage when exploring the role of gender in a variety of outcomes of interest. The strength of this measurement technique is that it fits with contemporary gender theory which emphasizes doing gender and gender performativity (Butler 1990, 1993, 1997; West and Zimmerman 1987) and has the potential to be applied to samples that vary across time, culture, and developmental periods. By utilizing new, reliable, and valid measurement techniques within datasets with rich information on other outcomes, social scientists can produce methodologically rigorous research that builds upon Bem’s foundational research and enhances our understanding of the role of gender in our social world.

Acknowledgments

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (<http://www.cpc.unc.edu/addhealth>). We are grateful to the Carolina Population Center for training support (T32 HD007168) and for research support (R24 HD050924). Paul J. Fleming was supported by the UNC STD/HIV training grant (T32AI007001) and subsequently by a NIDA training grant (T32 DA023356).

References

- Barnes CD, Brown RP, Tamborski M. Living dangerously: Culture of honor, risk-taking, and the nonrandomness of “accidental” deaths. *Social Psychological and Personality Science*. 2012; 3(1): 100–107.
- Bem SL. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*. 1974; 42(2):155–162. [PubMed: 4823550]
- Bem SL. On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*. 1977; 45(2):196–205. [PubMed: 850004]
- Bem SL. Dismantling gender polarization and compulsory heterosexuality: Should we turn the volume down or up? *Journal of Sex Research*. 1995; 32(4):329–334.
- Bohan JS. Essentialism, constructionism, and feminist psychology. *Psychology of Women Quarterly*. 1993; 17(1):5–21.
- Bos H, Goldberg N, Van Gelderen L, Gartrell N. Adolescents of the U.S. National Longitudinal Lesbian Family Study: Male role models, gender role traits, and psychological adjustment. *Gender & Society*. 2012; 26(4):603–638.
- Butler, J. *Gender trouble*. London: Routledge; 1990.
- Butler, J. *Bodies that matter: On the discursive limits of sex*. London: Routledge; 1993.
- Butler, J. *Excitable speech: A politics of the performative*. London: Routledge; 1997.
- Carmines, EG., Woods, JA. Reliability assessment. In: Kempf-Leonard, K., editor. *Encyclopedia of social measurement*. Vol. 3. Oxford: Elsevier; 2005a. p. 361–365.
- Carmines, EG., Woods, JA. Validity assessment. In: Kempf-Leonard, K., editor. *Encyclopedia of social measurement*. Vol. 3. Oxford: Elsevier; 2005b. p. 933–937.
- Choi N, Fuqua DR, Newman JL. Exploratory and confirmatory studies of the structure of the Bem Sex Role Inventory Short Form with two divergent samples. *Educational and Psychological Measurement*. 2009; 69(4):696–705.
- Cleveland HH, Udry JR, Chantala K. Environmental influences on sex-typed behaviors and attitudes of male and female adolescents. *Personality and Social Psychology Bulletin*. 2001; 27(12):1587–1598.
- Cobb-Clark DA, Schurer S. The stability of big-five personality traits. *Economics Letters*. 2012; 115(1):11–15.
- Cohn A, Zeichner A. Effects of masculine identity and gender role stress on aggression in men. *Psychology of Men & Masculinity*. 2006; 7(4):179–190.
- Colley A, Mulhern G, Maltby J, Wood AM. The short form BSRI: Instrumentality, expressiveness and gender associations among a United Kingdom sample. *Personality and Individual Differences*. 2009; 46(3):384–387.
- Connell, RW. *Gender and power: Society, the person and sexual politics*. Cambridge: Polity Press; 1987.
- Connell, RW. *Masculinities*. Berkeley: University of California Press; 1995.
- Courtenay WH. Constructions of masculinity and their influence on men’s well-being: A theory of gender and health. *Social Science & Medicine*. 2000; 50(10):1385–1401. [PubMed: 10741575]
- de Moya EA. Versiones y subversiones de la masculinidad en la cultura dominicana. [Versions and subversions of masculinity in Dominican culture]. *Perspectivas Psicológicas*. 2003; 3:184–190.
- DeVellis, R. *Scale development: Theory and applications*. Thousand Oaks: Sage; 2012.
- Egan SK, Perry DG. Gender identity: A multidimensional analysis with implications for psychosocial adjustment. *Developmental Psychology*. 2001; 37(4):451. [PubMed: 11444482]
- Fleming PJ, Gruskin S, Rojo F, Dworkin SL. Men’s violence against women and men are inter-related: Recommendations for simultaneous intervention. *Social Science & Medicine*. 2015; 146:249–256. [PubMed: 26482359]
- Gilmore, DD. *Manhood in the making: Cultural concepts of masculinity*. New Haven: Yale University Press; 1990.
- Gutmann, MC. *The meanings of macho: Being a man in Mexico City*. Berkeley: University of California Press; 2006.

- Harris, KM. The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007–2009. Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill; 2009.
- Harris, KM. The Add Health Study: Design and accomplishments. 2013. Retrieved from <http://www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIIV.pdf>
- Holt CL, Ellis JB. Assessing the current validity of the Bem Sex-Role Inventory. *Sex Roles*. 1998; 39(11–12):929–941.
- Hyde JS. The gender similarities hypothesis. *American Psychologist*. 2005; 60(6):581–592. [PubMed: 16173891]
- Jachyra P, Atkinson M, Gibson BE. Gender performativity during interviews with adolescent boys. *Qualitative Research in Sport, Exercise and Health*. 2014; 6(4):568–582.
- Jha P, Ranson MK, Nguyen SN, Yach D. Estimates of global and regional smoking prevalence in 1995, by age and sex. *American Journal of Public Health*. 2002; 92(6):1002–1006. [PubMed: 12036796]
- Leszczynski JP, Strough J. The contextual specificity of masculinity and femininity in early adolescence. *Social Development*. 2008; 17(3):719–736.
- Lippa RA, Connelly S. Gender diagnosticity: A new Bayesian approach to gender-related individual differences. *Journal of Personality and Social Psychology*. 1990; 59(5):1051–1063.
- Locke BD, Mahalik JR. Examining masculinity norms, problem drinking, and athletic involvement as predictors of sexual aggression in college men. *Journal of Counseling Psychology*. 2005; 52(3): 279–283.
- Luyt R. Beyond traditional understanding of gender measurement: The gender (re) presentation approach. *Journal of Gender Studies*. 2015; 24(2):1–20.
- Mahalik JR, Locke BD, Ludlow LH, Diemer MA, Scott RP, Gottfried M, et al. Development of the conformity to masculine norms inventory. *Psychology of Men & Masculinity*. 2003; 4(1):3–25.
- Mahalik JR, Morray EB, Coonerty-Femiano A, Ludlow LH, Slattery SM, Smiler A. Development of the Conformity to Feminine Norms Inventory. *Sex Roles*. 2005; 52(7–8):417–435.
- Mahalik JR, Lagan HD, Morrison JA. Health behaviors and masculinity in Kenyan and US male college students. *Psychology of men & Masculinity*. 2006; 7(4):191–202.
- Mahalik JR, Burns SM, Syzdek M. Masculinity and perceived normative health behaviors as predictors of men's health behaviors. *Social Science & Medicine*. 2007; 64(11):2201–2209. [PubMed: 17383784]
- Mahalik JR, Lombardi CM, Sims J, Coley RL, Lynch AD. Gender, male-typicality, and social norms predicting adolescent alcohol intoxication and marijuana use. *Social Science & Medicine*. 2015; 143:71–80. [PubMed: 26344125]
- Morrell R, Jewkes R, Lindegger G. Hegemonic masculinity/masculinities in South Africa: Culture, power, and gender politics. *Men and Masculinities*. 2012; 15(1):11–30.
- O'Neil JM. Is criticism of generic masculinity, essentialism, and positive-healthy-masculinity a problem for the psychology of men? *Psychology of Men & Masculinity*. 2010; 11(2):98–106.
- Parent MC, Moradi B. Confirmatory factor analysis of the Conformity to Feminine Norms Inventory and development of an abbreviated version: The CFNI-45. *Psychology of Women Quarterly*. 2010; 34(1):97–109.
- Pleck, JH. *The myth of masculinity*. Cambridge: MIT Press; 1981.
- Smiler AP. Conforming to masculine norms: Evidence for validity among adult men and women. *Sex Roles*. 2006; 54(11–12):767–775.
- Smiler, AP., Epstein, M. Measuring gender: Options and issues. In: Chrisler, JC., McCreary, DR., editors. *Handbook of gender research in psychology: Gender research in general and experimental psychology*. New York: Springer; 2010. p. 133–157.
- Spence, JT., Helmreich, R. *Masculinity and femininity: Their psychological dimensions, correlates, and antecedents*. Austin: University of Texas Press; 1978.
- Strough J, Leszczynski JP, Neely TL, Flinn JA, Margrett J. From adolescence to later adulthood: Femininity, masculinity, and androgyny in six age groups. *Sex Roles*. 2007; 57(5–6):385–396.

- Tate CC, Youssef CP, Bettergarcia JN. Integrating the study of transgender spectrum and cisgender experiences of self-categorization from a personality perspective. *Review of General Psychology*. 2014; 18(4):302–312.
- Taylor MC, Hall JA. Psychological androgyny: Theories, methods, and conclusions. *Psychological Bulletin*. 1982; 92(2):347–366.
- Twenge JM. Changes in masculine and feminine traits over time: A meta-analysis. *Sex Roles*. 1997; 36(5–6):305–325.
- Udry JR, Chantala K. Masculinity-femininity guides sexual union formation in adolescents. *Personality and Social Psychology Bulletin*. 2004; 30(1):44–55. [PubMed: 15030642]
- Udry JR, Chantala K. Masculinity-femininity predicts sexual orientation in men but not in women. *Journal of Biosocial Science*. 2006; 38(6):797–809. [PubMed: 16618376]
- Udry JR, Li RM, Hendrickson-Smith J. Health and behavior risks of adolescents with mixed-race identity. *American Journal of Public Health*. 2003; 93(11):1865–1870. [PubMed: 14600054]
- Vandello JA, Bosson JK. Hard won and easily lost: A review and synthesis of theory and research on precarious manhood. *Psychology of Men & Masculinity*. 2013; 14(2):101–113.
- Vandello JA, Bosson JK, Cohen D, Burnaford RM, Weaver JR. Precarious manhood. *Journal of Personality and Social Psychology*. 2008; 95(6):1325–1339. [PubMed: 19025286]
- West C, Zimmerman DH. Doing gender. *Gender & Society*. 1987; 1(2):125–151.

Fig. 1.
Steps to create measure of gender performance for Wave 1

Fig. 2.

Comparing distribution of respondents on Wave 1 measure, comparing between males and females, and overall sample and each split-half sample

Table 1Characteristics of analytic sample ($n = 9417$)

Characteristics	Men $n = 4277$	Women $n = 5140$	Difference test
Race/Ethnicity (Wave IV) ^a			Rao-Scott $\chi^2(df, p)$
White	67.6 %	68.1 %	4.9 (4), .30
Black/AA	14.8 %	15.8 %	
Asian/PI	3.8 %	3.2 %	
AmerIndian	1.5 %	1.1 %	
Hispanic	12.3 %	11.8 %	
Education (Wave IV)			Rao-Scott $\chi^2(df, p)$
Less than high school	10.6 %	7.2 %	72.5 (4) <.01
High school graduate	20.4 %	13.9 %	
Some college	40.7 %	44.4 %	
College graduate or higher	27.5 %	34.5 %	$t(df, p)$
Mean Age at Wave 1 (SE)	15.1 (.12)	15.0 (.11)	-3.52 (128), <.01
Mean Age at Wave 2 (SE)	16.0 (.12)	15.9 (.12)	-3.31 (128), <.01
Mean Age at Wave 3 (SE)	21.5 (.12)	21.3 (.12)	-4.10 (128), <.01
Mean Age at Wave 4 (SE)	28.0 (.12)	27.8 (.11)	-3.85 (128), <.01 $t(df, p)$
Mean Probability Measure at Wave 1 (SE)	.70 (.01)	.30 (.01)	-52.4 (127), <.01
Mean Probability Measure at Wave 2 (SE)	.75 (.01)	.25 (.01)	-55.1 (128), <.01
Mean Probability Measure at Wave 3 (SE)	.77 (.01)	.23 (.01)	-74.9 (128), <.01
Mean Probability Measure at Wave 4 (SE)	.73 (.01)	.26 (.00)	-70.0 (128), <.01 Rao-Scott $\chi^2(df, p)$
Fighting in past 12 months at Wave 1	42.9 %	22.5 %	286.4 (1), <.01
Fighting in past 12 months at Wave 2	27.2 %	13.6 %	180.3 (1), <.01
Fighting in past 12 months at Wave 4	8.5 %	2.4 %	113.1 (1), <.01

Using longitudinal weights, clustering at primary sampling unit and by individual, and stratification by region.

^aThese categories are mutually exclusive, for more details on how this variable was derived, see Udry et al. 2003

Table 2

Final variable lists by WAVE

#	Question (Response Min/Max)	Min/ Max	Boys' Mean (SD)	Girls' Mean (SD)	Boy-Girl Mean/Boy SD	Wald Chi-sq. contribution to model
WAVE1 – Final Variable List (sorted by contribution to the model)						
1	Frequency of Crying (0 = Never; 4 = Every day)	0/4	.16 (.43)	.64 (.81)	–1.13	320.95
2	Frequency of playing an active sport (0 = Not at all; 3 = 5 or more times)	0/3	1.70 (1.13)	1.07 (1.08)	.56	227.93
3	Got into a physical fight (0 = Never; 2 = More than once)	0/2	.62 (.79)	.32 (.62)	.38	145.53
4	How do you think of yourself in terms of weight? (1 = Very underweight; 5 = Very overweight)	1/5	2.99 (.77)	3.32 (.80)	–.43	118.76
5	How much do you feel that your friends care about you? (1 = Not at all; 5 = Very much)	1/5	4.11 (.82)	4.35 (.79)	–.29	91.10
6	Hours per week playing video/computer games (0–99 h)	0/99	4.22 (8.15)	1.35 (3.85)	.35	89.85
7	What do you think your chances are of getting an STD? (1 = Very high; 5 = No chance)	1/5	4.16 (.95)	4.40 (.87)	–.25	78.88
8	How many hours do you spend working for pay (0–140 h)	1/40	8.35 (12.42)	7.04 (11.00)	.11	72.65
9	Have you ever received an out-of-school suspension from school? (0 = No; 1 = Yes)	0/1	.37 (.48)	.21 (.41)	.34	71.93
10	Frequency of poor appetite (0 = Never/Rarely; 3 = Most of the time)	0/3	.35 (.62)	.60 (.78)	–.41	70.81
11	Frequency of wearing a helmet while cycling (0 = Never; 4 = Always)	0/5	1.38 (2.10)	2.08 (2.36)	–.33	58.79
12	Hours per week listening to the radio (0–99 h)	0/99	15.75 (19.45)	18.54 (20.71)	–.14	58.01
13	Frequency of doing work around the house (0 = Not at all; 3 = 5 or more times)	0/3	1.92 (.91)	2.15 (.87)	–.26	55.80
14	Upset by difficult problems (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.59 (1.03)	2.31 (.96)	.27	39.07
15	How much do you feel adults care about you? (1 = Not at all; 5 = Very much)	1/5	4.30 (.87)	4.44 (.79)	–.17	36.76
16	Frequency of moodiness (0 = Never; 4 = Every day)	0/4	1.09 (.96)	1.50 (1.06)	–.43	31.80
17	You have a lot to be proud of (1 = Strongly agree; 5 = Strongly disagree)	1/5	1.65 (.70)	1.78 (.75)	–.18	28.79
18	Have you taken a pledge to remain a virgin until marriage? (0 = No; 1 = Yes)	0/1	.09 (.29)	.16 (.37)	–.24	28.11
19	Frequency of exercise (0 = Not at all; 3 = 5 or more times)	0/3	1.61 (1.09)	1.66 (1.02)	–.04	27.90
20	Rely on gut feelings to make decisions (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.92 (1.14)	3.08 (1.12)	–.14	22.04

#	Question (Response Min/Max)	Min/Max	Boys' Mean (SD)	Girls' Mean (SD)	Boy-Girl Mean/Boy SD	Wald Chi-sq. contribution to model
5 = Strongly disagree)						
21	Trying to gain/lose/maintain weight? (1 = Lose; 4 = Nothing)	1/4	2.40 (.95)	2.08 (1.11)	.33	19.42
22	You never get sad (1 = Strongly agree; 5 = Strongly disagree)	1/5	3.47 (1.00)	3.81 (.89)	-.34	18.60
23	How likely is it that you will go to college (1 = Low; 5 = High)	1/5	3.98 (1.22)	4.27 (1.09)	-.24	15.97
24	You felt you were just as good as other people (0 = Never/Rarely; 3 = Most of the time)	0/3	1.96 (1.01)	1.79 (1.01)	.16	15.31
25	Frequency wearing a seatbelt in the car (0 = Never; 4 = Always)	0/4	2.95 (1.24)	3.22 (1.10)	-.22	10.12
WAVE2 – Final Variable List (sorted by contribution to the model)						
1	Frequency of crying (0 = Never; 4 = Every day)	0/4	.20 (.46)	.74 (.79)	-1.18	196.98
2	Frequency of sunbathing in the summer (1 = Frequently; 4 = Never)	1/4	3.46 (.85)	2.74 (1.14)	.85	186.07
3	Frequency of playing an active sport (0 = Not at all; 3 = 5 or more times)	0/3	1.68 (1.12)	1.04 (1.07)	.57	175.63
4	How you think of yourself in terms of weight (1 = Very underweight; 5 = Very overweight)	1/5	3.01 (.73)	3.32 (.77)	-.43	137.54
5	Have you ever driven a car (0 = No; 1 = Yes)	0/1	.84 (.36)	.77 (.42)	.20	70.54
6	Frequency of doing work around the house (0 = Not at all; 3 = 5 or more times)	0/3	1.93 (.88)	2.18 (.84)	-.29	68.28
7	Likely to use sunscreen (1 = Very likely; 3 = Unlikely)	1/3	2.63 (.65)	2.34 (.79)	.45	66.39
8	You like to take risks (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.30 (1.01)	2.62 (1.07)	-.32	62.15
9	Frequency of poor appetite (0 = Never/Rarely; 3 = Most of the time)	0/3	.36 (.62)	.64 (.79)	-.44	58.91
10	Difficult problems make you very upset (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.55 (1.06)	2.22 (1.00)	.31	53.04
11	Hours per week playing video/computer games (0–99 h)	0/95	3.85 (7.41)	1.33 (3.76)	.34	52.82
12	How much do you feel that your friends care about you? (1 = Not at all; 5 = Very much)	1/5	4.17 (.83)	4.42 (.80)	-.30	51.03
13	Past 12 months, how often get into a serious physical fight (0 = Never; =5 or more times)	0/3	.33 (.63)	.16 (.43)	.28	39.67
14	You will graduate from college (1 = Almost no chance; 5 = Almost certain)	1/5	3.73 (1.21)	4.05 (1.12)	-.26	34.44
15	You received testing/treatment for an STI/AIDS in past year	0/1	.04 (.20)	.08 (.27)	-.19	26.94

#	Question (Response Min/Max)	Min/Max	Boys' Mean (SD)	Girls' Mean (SD)	Boy-Girl Mean/Boy SD	Wald Chi-sq. contribution to model
(0 = No; 1 = Yes)						
16	Past 12 months, how often deliberately damage property that wasn't yours (0 = Not at all; 3 = 5 or more times)	0/3	.25 (.60)	.11 (.38)	.23	26.84
17	Times used sunlamp or a tanning bed in your life (0 = Never; 1 = 1 time; 5 = >20 times)	0/5	.21 (.82)	.75 (1.53)	-.65	26.54
18	You like yourself just the way you are (1 = Strongly agree; 5 = Strongly disagree)	1/5	1.75 (.81)	2.09 (.98)	-.42	18.31
19	You live without much thought for the future (1 = Strongly agree; 5 = Strongly disagree)	1/5	3.43 (1.11)	3.69 (1.03)	-.24	16.75
20	Number of past thirty days chewed tobacco (0–30 days)	0/30	1.23 (5.07)	.07 (.93)	.23	12.46
21	Frequency of wearing a helmet while cycling (0 = Never; 4 = Always)	0/5	1.67 (2.22)	2.33 (2.38)	-.30	11.67
22	You felt you were just as good as other people (0 = Never/Rarely; 3 = Most of the time)	0/3	2.00 (1.00)	1.88 (.99)	.12	9.59
23	Frequency wearing a seatbelt in the car (0 = Never; 4 = Always)	0/4	2.94 (1.23)	3.24 (1.08)	-.25	9.50
24	How honestly answered questions (1 = Not honestly at all; 4 = Completely honestly)	1/4	3.28 (.87)	3.48 (.72)	-.23	9.08
25	You felt lonely (0 = Never/Rarely; 3 = Most of the time)	0/3	.40 (.66)	.53 (.74)	-.20	6.73
26	Frequency of moodiness (0 = Never; 4 = Every day)	0/4	1.12 (.90)	1.51 (.97)	-.44	6.73
27	You are emotional (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.56 (.98)	2.14 (.94)	.43	6.71
WAVE3 – Final Variable List (sorted by contribution to the model)						
1	You were sad, during the past 7 days (0 = Never/Rarely; 3 = Most of the time)	0/4	.66 (.63)	1.26 (.68)	-.96	327.33
2	In past 7 days, how many times doing work around the house (0 = Not at all; 7 = 7 or more times)	0/7	3.82 (2.26)	5.02 (2.13)	-.53	157.24
3	What do you think of yourself in terms of weight? (1 = Very underweight; 5 = Very overweight)	1/5	3.15 (.78)	3.50 (.79)	-.45	124.22
4	In past 7 days, how many times did you participate in gymnastics, weight lifting (0 = Not at all; 7 = 7 or more)	0/7	1.39 (1.95)	.50 (1.22)	.45	119.33
5	I can do a good job stretching the truth when I talk to people (1 = Not true; 5 = Very true)	1/5	2.85 (1.36)	2.22 (1.27)	.47	86.09
6	Hours per week playing video/computer games (0–168 h)	0/168	6.39 (10.54)	3.18 (6.57)	.30	82.87
7	How many times engage in a hobby (e.g., play cards, arts and crafts, musical, etc.) (0 = Not at all; 7 = 7 or more)	0/7	2.91 (2.40)	2.43 (2.13)	.20	67.20
8	In past 7 days, how many times did you rollerblade/ski/	0/7	.59 (1.39)	.61 (1.38)	-.01	66.70

#	Question (Response Min/Max)	Min/Max	Boys' Mean (SD)	Girls' Mean (SD)	Boy-Girl Mean/Boy SD	Wald Chi-sq. contribution to model
racquet sports or aerobics (0 = Not at all; 7 = 7 or more)						
9	You like to take risks (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.31 (.99)	2.77 (1.08)	-.47	60.02
10	Have you used legal performance enhancing substances for athletes (i.e. creatine) (0 = No; 1 = Yes)	0/1	.15 (.36)	.01 (.12)	.38	59.98
11	In past 7 days, how many times did you participate in strenuous team sport (0 = Not at all; 7 = 7 or more)	0/7	.93 (1.65)	.19 (.82)	.45	54.30
12	Do you own a handgun? (0 = No; 1 = Yes)	0/1	.14 (.35)	.04 (.21)	.28	52.69
13	Past 12 months, how often deliberately damage property that wasn't yours (0 = Not at all; 3 = 5 or more times)	0/3	.17 (.47)	.05 (.25)	.25	43.12
14	In past 7 days, how many times hang with friends or talk on the telephone for more than 5 min? (0 = Not at all; 7 = 7 or more)	0/7	4.28 (2.41)	4.38 (2.35)	-.04	39.02
15	Number of past thirty days chewed tobacco (0–30 days)	0/30	1.38 (5.68)	.07 (.99)	.23	33.65
16	Have you ever been expelled from school (0 = No; 1 = Yes)	0/1	.11 (.31)	.04 (.20)	.20	29.37
17	Past 12 months, how often take part in physical fight where your group against another group (0 = Not at all; 3 = 5 or more times)	0/3	.19 (.51)	.04 (.20)	.30	28.37
18	Have you ever paid someone to have sex with you? (0 = No; 1 = Yes)	0/1	.05 (.22)	.01 (.07)	.21	26.38
19	Have you ever played games for money or taken part in another type of gambling for money? (0 = No; 1 = Yes)	0/1	.51 (.50)	.32 (.47)	.38	25.92
20	In past 7 days, how many times did you walk for exercise (0 = Not at all; 7 = 7 or more)	0/7	1.36 (2.18)	1.81 (2.16)	-.21	24.05
21	In past 7 days, how many times did you bike/skate/dance/skateboard (0 = Not at all; 7 = 7 or more)	0/7	1.56 (2.00)	1.10 (1.68)	.23	23.74
22	In past 7 days, how many times did you watch TV in the past seven days (0 = Not at all; 7 = 7 or more)	0/7	5.49 (2.14)	5.33 (2.18)	.07	21.03
23	How important is being faithful is for a successful marriage? (1 = Not important; 10 = Very important)	1/10	9.58 (1.30)	9.80 (.94)	-.17	16.83
WAVE 4 – Final Variable List (sorted by contribution to the model)						
1	Have you ever used chewing tobacco at least 20 times in your entire life? (0 = No; 1 = Yes)	0/1	.20 (.40)	.01 (.11)	.47	204.74
2	Have you ever been arrested? (0 = No; 1 = Yes)	0/1	.41 (.49)	.17 (.38)	.49	141.66
3	When you go outside on a sunny day for more than one hour, how likely are you to use sunscreen or sunblock? (1 = Very likely; 3 = Unlikely)	1/3	2.65 (.66)	2.32 (.83)	.50	128.80
4	I have a vivid imagination (1 = Strongly agree;	1/5	2.22 (.96)	2.48 (1.00)	-.28	123.17

#	Question (Response Min/Max)	Min/ Max	Boys' Mean (SD)	Girls' Mean (SD)	Boy-Girl Mean/Boy SD	Wald Chi-sq. contribution to model
5 = Strongly disagree						
5	I don't talk a lot (1 = Strongly agree; 5 = Strongly disagree)	1/5	3.20 (1.12)	3.60 (1.08)	-.35	117.98
6	I sympathize with others' feelings (1 = Strongly agree; 5 = Strongly disagree)	1/5	1.99 (.75)	1.66 (.68)	.44	84.33
7	In the past seven days, how many times did you participate in gymnastics, weight lifting, or strength training? (0 = Not at all; 7 = 7 or more times)	0/7	1.27 (1.91)	.54 (1.27)	.38	79.66
8	Hours per week playing video/computer games (0–105 h)	0/105	4.45 (9.45)	1.93(5.70)	.27	55.88
9	In past 7 days, how many times did you walk for exercise (0 = Not at all; 7 = 7 or more times)	0/7	1.73 (2.35)	2.15 (2.25)	-.18	55.36
10	I have frequent mood swings (1 = Strongly agree; 5 = Strongly disagree)	1/5	3.50 (1.06)	3.09 (1.10)	.38	51.38
11	How often do you pray privately? (0 = Never; 7 = More than once a day)	0/7	3.55 (2.63)	4.56 (2.38)	-.38	48.08
12	During typical summer week, how many hours do you spend in the sun during the day? (0–99 h)	0/99	17.52 (19.64)	10.56 (12.78)	.35	47.82
13	Have you ever been in the military? (0 = No; 1 = Yes)	0/1	.12 (.32)	.03 (.16)	.28	45.43
14	In the past 7 days, how many times did you participate in strenuous team sports such as football, soccer, basketball, lacrosse, rugby, field hockey, or ice hockey? (0 = Not at all; 7 = 7 or more times)	0/7	.49 (1.19)	.11 (.58)	.32	41.99
15	In the past 7 days, you felt too tired to do things. (0 = Never or rarely; 3 = Most or all of the time)	0/3	.72 (.71)	1.01 (.82)	-.40	39.88
16	I worry about things (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.56 (1.04)	2.09 (.89)	.44	39.20
17	Compared to other people your age, how intelligent are you? (1 = Moderately below; 6 = Extremely above)	1/6	4.01 (1.05)	3.86 (.99)	.15	30.61
18	I am not interested in other people's problems (1 = Strongly agree; 5 = Strongly disagree)	1/5	3.38 (.98)	3.76 (.90)	-.39	27.68
19	I like to take risks (1 = Strongly agree; 5 = Strongly disagree)	1/5	2.79 (.98)	3.18 (.98)	-.40	25.99
20	In the past 24 h, have you participated in vigorous activity long enough to work up a sweat, get your heart thumping, or get out of breath? (0 = No; 1 = Yes)	0/1	.49 (.50)	.33 (.47)	.33	23.39
21	I get stressed out easily (1 = Strongly agree; 5 = Strongly disagree)	1/5	3.46 (.96)	3.02 (1.05)	.46	22.28
22	I am not really interested in others. (1 = Strongly agree; 5 = Strongly disagree)	1/5	3.69 (.81)	3.93 (.73)	-.30	18.61

Table 3

Differences in final selected variables for split halves and original full dataset for each wave

Wave-Sample	Men ^a <i>M</i> (<i>SD</i>)	Women ^a <i>M</i> (<i>SD</i>)	# different from full sample	Item not included from full sample; ^b	Replaced with:	All variables in final model <i>p</i> <
W1-Sample 1	.70 (.24)	.30 (.25)	0	N/A	N/A	.02
W1-Sample 2	.71 (.24)	.31 (.25)	1	#25	Fearfulness	.01
W2-Sample 1	.75 (.24)	.27 (.26)	2	#26	-Is doing everything right	.10
				#19	-Hours listening to radio	
W2-Sample 2	.74 (.25)	.26 (.26)	2	#27	-Couldn't shake blues	.04
				#25	-Bothered by things that don't usually	
W3-Sample 1	.75 (.27)	.24 (.23)	3	#22	-Trying to change weight?	.01
				#21	-Participate in vigorous activities	
				#20	-Watch videos	
W3-Sample 2	.75 (.27)	.25 (.23)	1	#22	-Go to exercise center	.01
W4-Sample 1	.72 (.28)	.26 (.23)	3	#22	-Had health insurance	.01
				#20	-Used steroids	
				#19	-Is life of the party	
W4-Sample 2	.73 (.27)	.27 (.24)	2	#22	-How important is faith	.01
				#21	-I don't worry about things in the past	

^aRange of this measure is .00–1.00^bSee Table 1 for corresponding survey questions from full sample at each wave

Table 4

Descriptive statistics and correlations for probability measure of being male for each wave (w) and separated by men and women

Waves	<i>n</i>	Mean ^a (SE)	Pearson Correlation Coefficients			
			W1	W2	W3	W4
W1 Overall	9101	.50 (.01)	–			
W2 Overall	9203	.50 (.01)	.65	–		
W3 Overall	8905	.50 (.01)	.56	.61	–	
W4 Overall	8906	.50 (.01)	.53	.59	.61	–
			W1 (M)	W2 (M)	W3 (M)	W4 (M)
W1 Men	4136	.70 (.01)	–			
W2 Men	4162	.75 (.01)	.35	–		
W3 Men	4031	.77 (.01)	.17	.17	–	
W4 Men	3999	.73 (.01)	.17	.22	.23	–
			W1 (F)	W2 (F)	W3 (F)	W4 (F)
W1 Women	4965	.30 (.01)	–			
W2 Women	5041	.25 (.01)	.37	–		
W3 Women	4874	.23 (.01)	.18	.21	–	
W4 Women	4907	.26 (.00)	.17	.17	.18	–

Using longitudinal weights, clustering at primary sampling unit and by individual, and stratification by region

^aRange of this measure is .00–1.00

Table 5

Differences in mean of percentile measure at each wave for fighting

	Men				Women			
		%	Mean (SE) ^a	t(df)	p		Mean	t(df)
W1 - Fighting, past 12 months	- Yes	42.9	.61 (.01)	18.92 (128)	<.0001		.42 (.01)	-7.57 (128)
	- No	57.1	.41 (.01)				.53 (.01)	
W2 - Fighting, past 12 months	- Yes	27.2	.60 (.01)	10.88 (128)	<.0001		.45 (.02)	-5.40 (128)
	- No	72.8	.47 (.01)				.53 (.01)	
W4 - Fighting, past 12 months	- Yes	8.5	.63 (.02)	6.27 (128)	<.0001		.43 (.03)	-2.20 (128)
	- No	91.5	.50 (.01)				.51 (.01)	

Using longitudinal weights, clustering at primary sampling unit and by individual, and stratification by region

^aRange is .00–1.00