Further Examination of the Factor Structure of the Male Role Norms Inventory-Short Form (MRNI-SF): Measurement Considerations for Women, Men of Color, and Gay Men

Resubmitted April 8th, 2017

Ryon C. McDermott
University of South Alabama

Ronald F. Levant
University of Akron

Joseph H. Hammer
University of Kentucky

Rosalie J. Hall
Durham University

Daniel K. McKelvey
Zachary Jones
University of South Alabama

Abstract

Using multigroup structural equation modeling in a large sample of online-survey respondents (*N* =6,744), the present study examined the reliability and dimensionality of the Male Role Norms Inventory-Short Form (MRNI-SF), a popular measurement of traditional masculinity ideology (TMI), and also tested measurement invariance between individuals that do and do not fit the White heterosexual male TMI reference group. Results indicated that (a) it is appropriate to model the MRNI-SF using either a bifactor or unidimensional model but not a second-order model, (b) the raw MRNI-SF total score is a suitable measure of the general TMI construct, (c) the raw self-reliance through mechanical skills and negativity toward sexual minorities subscale scores may be appropriate measures of their respective specific factors (akin to subscale factors), and (d) SEM or ipsatizing procedures should be used to model the five other specific factors, given the insufficient model-based reliability of their raw subscale scores. When comparing men to women, White men to Black and Asian men, and gay men to heterosexual men, the MRNI-SF demonstrated configural invariance and at least partial metric invariance (i.e., measured similar constructs). However, scalar and residuals invariance were only supported for Asian men compared to White men. Taken together, these findings suggest that a general TMI factor of the MRNI-SF is best represented by a bifactor model, even in individuals that do not fit the White heterosexual male TMI reference group, but the instrument may be tapping somewhat different constructs in women, Black men, and gay men.

Public Significance Statement: The Male Role Norms Inventory Short-Form is a popular and widely used measure of traditional masculinity ideology (TMI). Recent research suggests it is best measured through a structural equation modeling approach, but this may not be practical for most psychologists, particularly clinicians. The present findings provide important guidelines for the use and interpretation of the instrument’s raw scores, as well as considerations for measuring TMI in women, men of color, and gay men.

Keywords: Traditional Masculinity ideology, Male Role Norms Inventory-Short Form, Bifactor, race, gender

Further Examination of the Factor Structure of the Male Role Norms Inventory-Short Form (MRNI-SF): Measurement Considerations for Women, Racial Ethnic Minorities, and Gay men

In the past 30 years, counseling psychologists have made significant advancements in the measurement of masculinity-related constructs, such as gender role conflict (O'Neil 1981, 2008), conformity to masculine norms (Mahalik et al., 2003) and masculinity ideologies (Levant, Hall, & Rankin, 2013; Levant, Hall, Weigold, & McCurdy, 2015). Most notably, investigators have identified that men internalize masculinity ideologies, consisting of socially constructed beliefs about what men should think, feel, and do. Furthermore, although there are likely countless masculinity ideologies present in any given culture (Pleck, 1995), researchers have identified that certain ideologies are rooted in patriarchal, Western, heteronormative, and traditional perspectives of men (Levant & Richmond, 2016). Such ideologies are based on beliefs about men and women prevalent before the feminist deconstruction of gender roles in the 1960s (c.f., Levant & Richmond, 2007, 2016). This set of beliefs—commonly referred to as *traditional masculinity ideology* (TMI)—has been connected to a variety of negative interpersonal and intrapersonal correlates (Levant & Richmond, 2007; 2016).

Given the potential negative consequences of TMI, researchers have worked to refine the measurement of these belief systems. Several instruments for assessing TMI and related domains have been developed over the decades, such as the Brannon Masculinity Scale (BMS; Brannon & Juni, 1984), the Male Role Norms Scale (MRNS; Thompson & Pleck, 1986) the Male Role Norms Inventory (MRNI; Levant et al, 1992), and the Conformity to Masculine Role Norms Inventory; Mahalik et al., 2003). Although all of these instruments have had a profound, positive impact on the field, only one—the MRNI—has been repeatedly refined through advanced factor analytic procedures to identify the best ways to model both specific aspects of TMI and a broad TMI general factor. The responses to items from the most recent version of the MRNI, the MRNI short form (MRNI-SF; Levant et al., 2013), can be used to construct a total score and seven subscale scores: avoidance of femininity (AoF), negativity toward sexual minorities (NTSM), self-reliance through mechanical skills (SRMS), toughness (T), dominance (Dom), importance of sex (IoS), and restrictive emotionality (RE).

Numerous studies have used various iterations of the MRNI to examine TMI across a variety of demographic groups, including, men, women, people of color, and sexual minorities (Levant & Richmond, 2007, 2016). However, very few researchers have examined the generalizability of the TMI construct across these demographic groups. Considering that TMI is based on heterosexual, White, male, and Eurocentric perspectives of men and masculinity (Connell & Messerschmidt, 2005; Levant & Richmond, 2016), determining whether the MRNI-SF measures constructs in the same way across different demographic groups is critical. For example, although distinctive masculinities have been identified for different ethnic/racial groups (e.g., Rogers, Sperry, & Levant, 2015), theoreticians have posited that everyone in society must contend with the dominant (or hegemonic) masculine norms (Connell & Messerschmidt, 2005). However, it is not yet known whether persons from different demographic groups construe TMI in the same way – that is, attach the same meaning to the same scale score. Finding evidence for invariance would allow us to have confidence that the scores from different demographic groups could be reliably compared. In addition, although the MRNI-SF has been subjected to rigorous confirmatory factor analytic (CFA) investigations of its factor structure (Levant et al., 2013; Levant et al, 2015; Levant, Hall, Weigold, & McCurdy, 2016), advances in measurement research indicate that further important psychometric properties of the MRNI-SF may need to be examined. Accordingly, the present study extended prior research by modelling the factor structure of the MRNI-SF, calculating indices of model reliability and dimensionality, and testing measurement invariance between groups that fit the White, male, and Eurocentric perspectives reflected in TMIs (i.e., White heterosexual men), and groups that do not reflect those qualities (i.e., women, racial and ethnic minorities, and gay men).

Factor Structure of the MRNI-SF

The MRNI-SF has several advantages over the original MRNI, as well as other instruments measuring masculinity ideology. In addition to being significantly shorter, the MRNI-SF has demonstrated good fit in a confirmatory factor analysis (CFA) measurement model of the seven subfactors and more advanced models specifying simultaneous influences of the seven subfactors and a general TMI factor (Levant et al., 2013). Specifically, Levant and colleagues (2013) tested two competing measurement models for the MRNI-SF which included a broad TMI factor: a second-order factor structure and a bifactor structure. Comparisons of model fit statistics (i.e., via scaled chi-square difference tests) suggested a strong preference for the bifactor model over the second-order factor model (Levant et al., 2013; Levant et al., 2015).

This finding has important implications for the measurement of TMI and for interpreting the meaning of various MRNI-SF scores. In particular, a second-order factor model (see figure 1a) implies a hierarchical structuring of a broad TMI factor and narrower factors representing specific TMI domains. The higher-level factor (i.e., MRNI-SF total score) accounts for any observed relationships among the set of lower-order factors (i.e., MRNI-SF subscales) and ultimately the variance explained in each item by its respective latent variable (c.f., Chen, West, & Sousa, 2006). By contrast, a bifactor structure (see Figure 1b) imposes no hierarchy among the factors and suggests that the variability in responses to the items is potentially attributable both to the general factor (i.e., a broad TMI construct), as well as additional, unrelated variance contributed by one or more specific factors (i.e., the seven TMI domains). Therefore, a bifactor model implies that the variance of each item is an additive combination of variance explained, in part, by a general factor and a specific (i.e., group) factor (Kline, 2016), keeping in mind that it might be the case that some items have only variance attributable to a general factor or only attributable to a specific factor (Chen et al., 2006). That the bifactor model emerged as a best fit for the MRNI-SF in previous studies suggests individuals may have both an overall conception of what it means to be traditionally masculine and a separate understanding of specific aspects of TMI that exist independent of a general TMI factor.

Although Levant and colleagues' (2013, 2015) have taken some significant steps in examining the bifactor structure of MRNI-SF scores, several issues remain unaddressed. For example, previous bifactor model specifications of the MRNI-SF have varied considerably in the extent to which correlations are allowed among the specific factors. The specification for a classic bifactor model forces all factors to be completely orthogonal; thus partitioning item level variance into only two sources: the general factor and the item's intended specific factor (Reise, 2012). However, in the initial validation study of the MRNI-SF, Levant and colleagues (2013) allowed all of the specific factors to intercorrelate with each other, and in a later study (Levant et al., 2015), certain correlations were freed based on modification indices. If a bifactor structure is the best representation of the MRNI-SF's dimensionality, then it is important for investigators to identify the effects of different ways of specifying the model for future research and practice.

In addition to clarifying previous MRNI-SF bifactor specifications, recent recommendations for best practices in bifactor modeling (Reise, 2012; Rodriguez, Reise, & Haviland, 2016) suggest additional analyses could provide a more nuanced understanding of how the MRNI-SF functions and how observed scores should be used or interpreted. In particular, one advantage of a bifactor model compared to a second-order or common-factors model is that researchers can calculate ancillary bifactor indices to determine the most appropriate interpretation of an instrument's dimensionality and the model-based reliability of the total and subscale scores (see Hammer & Toland, 2016, for a video walkthrough).

Riese (2012) provided examples of situations, for instance, in which a general factor may emerge in a bifactor structure. However, the total score could still be an *unreliable* measure of the general factor, and thus the raw total score for that instrument would primarily measure error and not the construct of interest. Likewise, subscale scores may or may not be sufficiently reliable measures of their corresponding specific factors. In other words, because a bifactor structure contains both general and specific sources of common variance, it is important to separate out the reliable variance in a composite score that can be attributed to either the general or specific factor (Rodriguez et al., 2016). Thus, to develop a more nuanced understanding of the model-based reliability of an instrument's total and subscale scores, researchers have strongly recommended calculating ancillary bifactor indices including the omega coefficients (Reise, 2012; Rodriguez et al., 2016). These indices provide valuable information as to whether the raw total score and raw subscale scores are pure and reliable measures of the intended construct (Rodriguez et al., 2016). For example, if specific factor variance significantly contaminated the raw total score for the MRNI-SF, then this suggests that it would be misleading to interpret the raw total score as a *pure* and *reliable* measure of the TMI general factor. Likewise, if the general factor variance significantly contaminated the raw Restrictive Emotionality subscale score, then this suggests it would be misleading to use this subscale score as a measure of the Restrictive Emotionality specific factor.

In addition to model-based reliability estimates, researchers have recommended that further diagnostic indices (e.g., explained common variance [ECV], individual item explained common variance [IECV], and percent of uncontaminated correlations [PUC]; c.f., Rodriguez et al., 2016) are warranted to provide more information about the dimensionality of an instrument. PUC has been shown to moderate the influence of ECV (i.e., the proportion of all common variance explained by the general factor) on parameter bias (Reise, Scheines, Widaman, & Haviland, 2013) and IECV (i.e., the item-level variation attributed to the general factor alone) allows researchers to determine the percent of item common variance attributable to a general dimension (Stucky & Edelen, 2014). For example, if the ECV and PUC were below recommended thresholds, this would suggest that the MRNI-SF is primarily multidimensional and conceptualizing the instrument as having a general dimension that can be measured by a total score would be contraindicated. In this case, calculating and interpreting a total score would lead to false conclusions that risk misleading scholars, clinicians, and our publics. In particular, using such a score to make clinical or policy decisions predicated on a person or population's "overall level of TMI" would be baseless.

Using subscale scores when the instrument is primarily unidimensional can also lead to problems. For example, Rodriguez, Reise, & Haviland (2016a) provided evidence that the majority of the supposedly-multidimensional instruments used in 50 recent studies were actually primarily unidimensional, and therefore produced subscale scores that failed to measure their intended subscale factors. *The Standards for Educational and Psychological Testing* state that "the improper use of tests… can cause considerable harm to test takers and other parties affected by test-based decisions" (American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 1). Thus, there are risks in continuing to misuse instruments that have a different underlying factor structure than those suggested by the creators of the instrument. Because researchers, to date, have not examined ancillary bifactor indices of reliability and dimensionality, questions still remain as to how to best model and score the MRNI-SF.

Invariance across Different Demographic Groups

In addition to examining ancillary bifactor indices, researchers investigating TMI may benefit from exploring whether there are measurement differences among specific groups. Indeed, investigators have examined different versions of the MRNI in samples of women and Black men (Levant, Majors, & Kelley, 1998) and Black and Latino men (Levant et al., 2003). A consistent finding among these studies is that men and racial-ethnic minorities tend to report higher levels of TMI compared to women and individuals from the racial-ethnic majority (c.f., Levant & Richmond, 2016). However, it is vital to remember that the TMI construct largely reflects White, male, heterosexual, and Eurocentric masculinities, and very little research has examined the generalizability of TMI in populations that differ from those reference groups.

It is important to determine whether the MRNI-SF is comparable across groups of individuals who do and do not fit the White, male, heterosexual, and Eurocentric reference group for the construct. Rather than simply comparing correlations or means of observed MRNI-SF scores between White heterosexual men and groups of women and minorities, multigroup structural equation modeling and testing for measurement invariance may illuminate between-group differences in the meaning, scaling, and precision of the instrument. Measurement invariance testing is a systematic way of determining which measurement model parameters are the same, and which differ, across defined groups (Kline, 2016; Vandenberg, 2002). Specifically, using multi-group structural equation modeling, researchers are able to determine various levels of measurement invariance. At the most basic level (i.e., *configural invariance*), a measurement model with a specific structure provides acceptable fit in all groups when no cross-group equality parameter constraints are specified. Configural invariance is a prerequisite for testing whether stronger forms of invariance hold. *Metric invariance* is stronger than configural invariance, and it is present when the factor loadings for observed variables (e.g., MRNI-SF items) are not significantly different across groups (Kline, 2016; Vandenberg & Lance, 2000). Once configural and metric invariance have been established, researchers can examine *scalar invariance*, in which the intercepts of the measured indicators are equivalent across groups. Both metric and scalar invariance are essential for testing stricter levels of invariance, such as the equivalence of residuals. Thus, if a measurement model demonstrates all types of invariance, (a) the basic factor structure in each group is similar (i.e., configural invariance), (b) items are measuring similar constructs in each group (i.e., metric invariance), (c) differences in observed scores are reflective of differences in the true construct (i.e., scalar invariance), and (d) the construct is being measured in each group with the same degree of precision (i.e., invariance of residuals) (Kline, 2016).

Although researchers have yet to examine measurement invariance of the MRNI-SF in racial and sexual minority groups, evidence suggests that the dominant White, European, and heterosexual TMI may be differentially internalized among specific cultural groups. Unlike White men, for instance, Black men likely develop their sense of manhood in ways that intersect with racial identity (e.g., Rogers et al., 2015). Cross-cultural research also suggests that some Asian men deemphasize physical toughness or avoidance of femininity in their conceptions of masculinity (c.f., Iwamota & Kaya, 2016). Regarding sexual orientation, a small but growing body of literature indicates that gay men may internalize certain aspects of TMI and exaggerate them in certain subcultures (c.f., Sanchez, 2016).

Because the MRNI-SF may be best represented by a bifactor structure, it may be possible to determine differences in the specific TMI domains, as well as an overall conception of TMI, between individuals that do and do not fit the White heterosexual male TMI reference group. For example, Levant and colleagues (2013) explored invariance in a bifactor model of the MRNI-SF across men and women. The authors discovered that men and women's MRNI-SF factor loadings were non-invariant for the general factor but were invariant for the specific factors. The authors interpreted these gender differences as suggesting that men's sense of self or identity may be engaged to a greater degree than women's when responding to questions about how much they agree or disagree with normative statements regarding men's behavior. Given the relative lack of research examining the MRNI-SF for measurement invariance between men and women or other groups, it is currently unclear whether the same pattern of results would be obtained in a different, larger sample of men and women, as well as when testing invariance between racial or sexual minority men compared to racial or sexual majority men.

The Present Study

To build upon recent research examining the psychometric properties of the MRNI-SF, additional investigation is warranted to examine (a) the dimensionality and model-based reliability of the MRNI-SF and (b) measurement invariance in populations which differ substantially from the White, European, male and heterosexual reference group from which TMI originates. Such analyses will provide more specific recommendations for how the structure of the MRNI-SF should be specified (e.g., unidimensional, bifactor, correlated factors) and whether raw scores of the MRNI-SF are reliable enough to warrant interpretation. Accordingly, the present study followed recent best-practice recommendations in examining bifactor structures by calculating reliability (i.e., omega coefficients) and dimensionality indices (i.e., ECV, IECV, and PUC) to determine the most appropriate way to model the MRNI-SF and interpret the raw MRNI-SF scores. In addition, the present study tested configural, metric, scalar, and residuals measurement invariance of the MRNI-SF bifactor structure, across groups of White heterosexual men compared to White heterosexual women, heterosexual men of color, and gay White men.

Based on previous research (i.e., Levant et al., 2013; Levant et al., 2015), we hypothesized that (a), a bifactor model would represent acceptable model fit, (b) a bifactor model would represent a better fit compared to a second-order hierarchical model, and (c) a bifactor model will demonstrate, at least, configural invariance between White heterosexual men and each comparison group. Because only one other study has examined different levels of measurement invariance across men and women (e.g., Levant et al., 2013), and because no published studies are available of the MRNI-SF omega reliabilities or dimensionality indices, no hypotheses regarding ancillary bifactor indices or metric, scalar, or residuals invariance were advanced.

Method

Procedures and Participants

The present sample was obtained by combining data gathered from six studies using the MRNI-SF between 2013 and 2015 examining the following: perspectives of intimate relationships (Study 1; *N* = 3,349), male reference group identity (Study 2; *N* = 1,439), gender role ideology development (Study 3; *N* = 1,231), sexual health (Study 4; *N* = 73), gender role discrepancy strain (Study 5; *N* = 165), and adult son's recollections of their fathers' expectations (Study 6; *N* = 357). Participants from study 1 were used in a previously published investigation that did not include the MRNI-SF (McDermott, Cheng, Lopez, McKelvey, & Bateman, 2016). Likewise, participants from Study 6 were used in a recently published investigation that did not include MRNI-SF responses (Levant, Gerdes, Alto, Jadaszewski, & McDermott, in press). However, in addition to the six primary studies, 70 men of color were pulled from the original MRNI-SF validation study (Levant et al., 2013) by permission to address the lack of racial diversity in the pooled sample.

After IRB approval, data were collected across two large universities in the Midwest (Studies 1, 2, 3, and 6) and a mid-size university in the South East (Studies 4 and 5). Studies 1 and 2 involved sending an anonymous online survey to a random, representative sample of students recruited by e-mail. The remaining data for each study was primarily collected through student subject pools at each institution. In addition to a student subject pool, the survey link for Studies 2, 3, 4, and 6 was posted on the volunteer section of Craigslist across 10 major U.S cities capturing all regions of the country, as well as advertised through Facebook.

The pooled sampled consisted of 6,744 men and women. Of this number, the majority (approximately 71%) were recruited from the university populations and approximately 16% of the sample were recruited from the internet. As a result of a technical error, in which participants from study 2 were randomly not exposed to the recruitment origin question, approximately 13% of the sample did not indicate their recruitment origin. Due to the inability to determine who actually saw the link to the online surveys, as well as which students actually opened the e-mail recruitment message, a true response rate could not be calculated for the two studies that used e-mail recruitment. However, approximately 25% of the targeted college student population responded to the e-mail recruitments for those surveys. The pooled sample was diverse in age (*M* = 25.06, *SD* = 9.52), and, although ages ranged from 18 to 87, 83% of the sample fell between 18 and 29 years of age, and the median age was 22. The sample was generally balanced between men (56%) and women (44%) but was primarily heterosexual (72%) and cisgender (99%). Participants reported a variety of highest achieved educational levels: no high school diploma (1%), high school diploma (56%), Associate degree (5%), Bachelor degree (28.9%), Master's (3.8%), or specialist or doctorate (2%).[1] The total sample was also diverse in race and ethnicity: White (61%), Asian (19%), Black (10%), and Latino (7%), and other (3%) (i.e., multiracial or a group not captured by the standard White, Black, Asian, and Latino categories).

Instrument

Male Role Norms Inventory - Short Form (MRNI-SF). The MRNI-SF (Levant, Hall, & Rankin, 2013) is a 21-item version of the Male Role Norms Inventory-Revised (MRNI-R; Levant et al., 2010) designed to measure endorsement of beliefs associated with TMI. Participants taking the MRNI-SF rate their agreement with statements indicating beliefs about appropriate male behaviors (e.g., " Men should be the leader in any group.") using a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). The MRNI-SF generates a total score and seven subscale scores of three items each: Avoidance of femininity (AoF; “Men should watch football games instead of soap operas”), negativity toward sexual minorities (NTSM; “Homosexuals should never marry”), self-reliance through mechanical skills (SRMS; “A man should know how to repair his car if it should break down.”), toughness (T; “When the going gets tough, men should get tough”), dominance (Dom; “Men should be the leader in any group”), importance of sex (IoS; “Men should always like to have sex”) and restrictive emotionality (RE; “A man should never admit when others hurt his feelings”). Significant correlations with relevant other latent factors provided concurrent validity evidence for the MRNI-SF specific latent factors (Levant et al., 2016). Validity of the general TMI factor was supported by latent correlations with: (a) Male Role Attitudes Scale; (b) general factor of Conformity to Masculine Norms Inventory-46; (c) higher-order factor of Gender Role Conflict Scale; and (d) Personal Attributes Questionnaire-Masculinity Scale (Levant et al., 2016). Internal consistencies for the present study were commensurate with previous studies (e.g., Levant et al., 2013) for AoF (.87), NTSM (.89), SRMS (.88), T (.76), Dom (.87), IoS (.86), RE (.79), and the total MRNI-SF score (.93).

Primary Analysis Plan

Our primary analyses consisted of two parts. First, we used structural equation modeling to examine the factor structure of the MRNI-SF in the total pooled sample. Specifically, a series of measurement models were tested to confirm that a seven-factor MRNI-SF was appropriate and to identify the best approach for modeling a general TMI latent variable by comparing the fit of unidimensional, second-order, and bifactor models. Second, we used multi-group SEM to determine if the best fitting model from our previous analysis was invariant across different groups. All analyses were performed using Mplus Version 7.31 (Muthén & Muthén, 1998-2015), FIML to handle missing values, and a maximum likelihood estimator with robust standard errors (MLR) to address normality violations.

Analytic approach for single-group models. In evaluating each individual model, we used the following fit indices and recommended cutoffs (Kline, 2016): the Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI) (values close to .95 indicate a good fit for both the CFI and TLI); the Root Mean Square Error of Approximation (RMSEA) with 90% confidence intervals [CI] (values of .06 or less indicate a good fit), and the Standardized Root-Mean-Square Residual (SRMR; values of .08 or less indicate a good fit). The chi-square test statistic was also reported (a non-significant value indicates a good fit); however, it was interpreted with caution given the extremely large sample size (Kline, 2016). For comparative model testing, we used chi-square difference tests. Because the chi-square statistic was scaled to accommodate non-normality, any chi-square difference tests that were performed were corrected according to the procedure developed by Satorra and Bentler (c.f., Satorra & Bentler, 2001) and described on the Mplus website (https://www.statmodel.com/chidiff.shtml). For comparisons between non-nested models, we examined the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), with lower AIC and BIC values indicating a better fitting model (Kline, 2016).

If a bifactor model represented a better fit than a second-order model, we calculated several ancillary bifactor indices, some of which were missing from previous MRNI-SF research but are recommended as essential practices in bifactor modeling (Rodriguez et al., 2016). These were: (a) the percent of explained common variance (ECV; Reise, Moore, & Haviland, 2010) associated with the general and each specific factor; (b) Percent of Uncontaminated Correlations (PUC; Reise et al.,2013); (c) Individual Explained Common Variance (IECV Stucky & Edelen, 2014); and (d two specialized model-based reliability coefficients known as Coefficient Omega Hierarchical and Coefficient Omega Hierarchical Subscale (Reise, 2012). When Percent of Uncontaminated Correlations (PUC) values are higher than .80, low general ECV values are less indicative of measurement parameter bias; when PUC values are lower than .80, general ECV values > .60 and Omega Hierarchical > .70 suggest that the presence of some multidimensionality is not severe enough to disqualify the interpretation of the instrument as primarily unidimensional (p. 22, Reise et al., 2013). In other words, a PUC value greater than .80 suggests that modeling an instrument as unidimensional is likely appropriate, even if the ECV value is lower than the aforementioned .60 threshold. The more general index of Coefficient Omega (ω) measures the proportion of total score variance in a set of indicators (e.g., the MRNI-

SF items) that can be attributed to all common factors, thus ω estimates true score variance (and excludes error variance). Coefficient Omega Hierarchical (ωH; McDonald, 1999) is an adaptation of Coefficient Omega that measures the proportion of subscale score variance that can be attributed to a single general factor after accounting for specific (i.e., subscale) factors. Coefficient Omega Hierarchical Subscale (ωHS) is a version of ωH that measures the proportion of subscale variance that is uniquely due to one specific factor, after controlling for the general factor.

Analytic approach for multi-group models. We used multi-group SEM to test different forms of invariance at the measurement level (Cheung & Lau, 2012; Kline, 2016). Invariance is traditionally tested by examining differences in the chi-square statistic across a series of nested models in which parameters are constrained to be equal across groups; however, researchers have identified that the model chi-square difference test often yields a statistically significant result even with very modest chi-square changes, especially in large samples (Cheung & Rensvold, 2002). As an alternative, simulation studies suggest that changes in the comparative fit index (ΔCFI) of less than .01 may be reliable indicators of different forms of measurement invariance (Cheung & Rensvold, 2002). However, statisticians have noted that the ΔCFI is questionable when groups have unequal sample sizes, and that CFI values have no known sampling distribution (Cheung & Rensvold, 2002). The latter limitation is especially problematic because it means that changes in CFI cannot be tested for statistical significance, and thus the .01 value is a general “rule of thumb” (Cheung & Lau, 2012).

In a response to these criticisms, Cheung and Lau (2012) proposed and demonstrated a direct-model comparison approach to invariance testing using bias corrected bootstrap confidence intervals. Unlike the nested model comparisons, Cheung and Lau’s technique avoids the pitfalls of the chi-square difference test and the ΔCFI because it does not compare nested models. Instead, assuming the model passes an initial configural invariance test, different forms of invariance are examined in the same model systematically by performing a bootstrap analysis of differences between groups on specific parameters (Cheung & Lau, 2012). The procedure creates 1000 bootstrap samples and derives the high and low confidence intervals for each parameter (e.g. differences across groups on factor loadings or intercepts). If zero falls within the conference interval, then the difference between the groups are not statistically significant, and those parameters are considered to be invariant across groups. Cheung and Lau (2012) demonstrated several advantages of the bias corrected bootstrap confidence interval approach, including the ability to determine where non-invariance exists within a model with ease. Moreover, the bootstrap procedure allows researchers to systematically test different items as the referent (i.e., which items are constrained to 1 to scale the latent variable; see Cheung & Lau, 2012 for a discussion of referent item selection). Thus, the present study involved separate bootstrap analyses constraining the metric to 1 for all possible combination of items (3 possible referents for the specific factor items and 21 possible referents for the general factor items), and, consistent with Cheung and Lau's (2012) recommendations, only items that were consistently invariant across different referents were considered invariant.

Of note, although Cheung and Lau's (2012) approach is novel and promising, scholars generally recommend against relying solely on one invariance testing technique to make decisions (Kline, 2016). The present study used a combination of ΔCFI values of .01 or less and bias-corrected bootstrap confidence intervals to supplement the chi-square difference test for determining measurement invariance across groups.

Results

Preliminary analyses

Prior to conducting our primary analyses, we examined data for missing values, univariate outliers, and assumptions of normality. The specifics of these analyses are available in the online supplementary files. In summary, the number of participants with missing values, univariate, and multivariate outliers were minimal, but MRNI-SF scores were positively skewed. Table 1 displays the raw correlations, means, and standard deviations of MRNI-SF subscale and total scores in the total sample.

Single-Group Measurement Models

We used SEM analyses with a robust estimator (i.e., the MLR estimator available in Mplus, v. 7.31) that corrects the chi-square and standard error values for non-normality. In addition, we used Full Information Maximum Likelihood (FIML) estimation to address missing data on MRNI-SF items.

Common Factors Model. As has been recommended in previous bifactor analyses (e.g., Chen et al. 2006), we first tested a common factors model to ensure that the underlying structure for the MRNI-SF items had seven narrower dimensions as anticipated. If a common factors model with seven freely co-varying factors corresponding to the various intended TMI domains showed serious misfit, the planned further analyses to determine the best way to model a broad MRNI-SF factor (i.e., representing general TMI) would need to be modified (and would differ from the results of previous studies). Although the chi-square test statistic was significant, χ^2 (168, N =6,744) = 2,665.92, p < .001 (indicating that the model was not a perfect fit), the remaining indices suggested acceptable fit, CFI = .960, TLI = .951, RMSEA = .047 (90% CI = .045, .049), and SRMR = .034.

Modeling general TMI. After confirming that a seven-factor MRNI-SF was appropriate for further analyses, we examined three different approaches for modeling a general TMI latent factor: a unidimensional structure, a second-order model, and a bifactor model (see Chen et al., 2006 for an in-depth discussion of the differences between second-order and bifactor models). For the unidimensional model, we specified the 21 items as loading on the overall TMI factor. For the second-order model, we specified a higher-order TMI latent variable with paths leading to each of seven lower-order TMI domains and the covariation between disturbance terms for each of the lower-order TMI domains constrained to zero. For the bifactor model, we tested three different variations based on previous MRNI-SF research: (a) an oblique bifactor model, (b) a completely orthogonal bifactor model, and (c) a modified bifactor model. Each model specified that the general and specific factors were orthogonal to each other (which is critical to a bifactor model; Reise, 2012), but to provide a complete exploration of the possible bifactor structures of the MRNI-SF used in previous research, we varied the degree to which the specific factors were orthogonal to each other a priori. Specifically, the oblique bifactor model (Levant et al., 2013) allowed all covariances between specific factors to be freely estimated, which was the specification used in the original validation study of the MRNI-SF. By contrast, the completely orthogonal model constrained all covariances between specific factors to zero, based on the general recommendations of Reise (2012) of a pure bifactor model. Finally, the modified bifactor model, based on the specification of Levant and colleagues (2015), freed the covariances between T and SRMS and between Dom and NT.

As illustrated in Table 2, although the chi-square test for each model was significant, the second-order and bifactor models all yielded acceptable CFI, TLI, RMSEA, and SRMR values. However, the unidimensional model evidenced extremely poor fit. The oblique bifactor model (based on Levant et al., 2013) yielded the strongest fit overall, followed by the modified bifactor (based on Levant et al., 2015) and the completely orthogonal bifactor model. However, the oblique bifactor model yielded a non-positive definite matrix, as evidenced by a negative error variance for one of the items. Chen and colleagues (2006) noted that the presence of technical errors can occur in bifactor models due to problems that are masked in a second-order model; however, this may also indicate that the oblique bifactor model might not be trustworthy (e.g., Kline, 2016). By contrast, the modified bifactor model and the orthogonal (i.e., pure) bifactor model converged without any technical errors and evidenced acceptable fit, suggesting that either approach may be appropriate for further use.

Because a second-order model is nested within a bifactor model, we used a scaled chi-square difference test to determine if the more parsimonious second order model was a significantly worse fit than the bifactor models. For each comparison, the second-order model had a significantly larger scaled chi-square, smaller CFI, TLI values, and larger RMSEA and SRMR values than the bifactor model. Thus, the bifactor model, regardless of the specification of correlations between specific factors, represented the most appropriate way of modeling an overall TMI general factor (see Table 3).[2]

Ancillary Bifactor Indices

We calculated the Explained Common Variance attributable to the general factor and to each of the seven specific factors from the bifactor model (see Table 4). Following best-practice recommendations for the use of bifactor modeling for variance partitioning diagnostic purposes (e.g., Reise, 2012; Rodriguez et al. 2016), we constrained the covariances to zero between each of the seven TMI specific factors to obtain pure (i.e., uncontaminated by shared variance) measurements of each ancillary bifactor diagnostic measure. The general factor ECV was .58, indicating that 58% of the common variance across the 21 items was due to the general factor. The remaining 42% of the common variance is due to the set of seven specific factors, with SRMS and NT accounting for the largest share of that collective specific factor variance (11% and 9%, respectively).

Three findings further inform the dimensionality of the MRNI-SF. First, according to Reise and colleagues (2013), because the PUC value (.90) was greater than .70, even though the general ECV values (.58) was slightly below the .60 threshold, this does not necessarily indicate that modeling the MRNI-SF as a unidimensional instrument would lead to substantial measurement parameter bias (i.e., biased item factor loadings). Second, the average Individual Explained Common Variance (IECV) coefficient for the general TMI factor for the 21 items ranged from .33 to .91. The average IECV of .59 suggested that, on average, items measured the general factor to a slightly stronger degree than they measured the intended specific factor. Third, the average relative measurement parameter bias (see Rodriguez et al., 2016) across items was 4%, which falls well below the upper limit (10-15%) posited by Muthén, Kaplan, and Hollis (1987). Examined another way, the difference between an item’s standardized loading in a unidimensional solution (i.e., all MRNI-SF items specified to load on a single factor) and its general factor loading in the bifactor solution was no more than an average of $\Delta\beta = .02$. In summary, despite the poor fit of a unidimensional solution for the MRNI-SF, these three findings collectively suggest that, while the MRNI-SF contains significant multidimensionality, it may be possible to model the general TMI latent factor in the context of a unidimensional solution. In other words, it appears that negligible measurement parameter bias is introduced by modeling the general TMI latent factor using a simpler unidimensional solution, rather than the more accurate—but statistically complex—bifactor solution.

In addition, the model-based reliability coefficients of Coefficient Omega Hierarchical (ωH) and Coefficient Omega Hierarchical Subscale (ωHS) were calculated. While no definitive benchmarks for evaluating ωH and ωS exist at the time of this writing, Reise, Bonifay, and Haviland (2013) state that "tentatively, we can propose that a minimum would be greater than .50, and values closer to .75 would be much preferred" (p.137). Thus, ωH > .75 would typically indicate that the MRNI-SF's total score predominantly reflects a single general factor despite the presence of multidimensionality across items. Normally, this would signify that it is permissible to interpret the MRNI-SF total score as a sufficiently reliable and appropriate measure of the general construct of TMI.

Regarding the subscales, ωS < .50 would indicate that the majority of that subscale's variance is due to the general factor and that negligible unique variance is due to that specific factor. In other words, that subscale score's reliability is substantially inflated (i.e., confounded) by the general factor and does not reliably measure the narrower subdomain construct that the subscale was designed to measure.

Table 4 summarizes the ω, ωH, and ωHS coefficients for the bifactor solution for the MRNI-SF. The general factor achieved an ωH > .75, and 91% of the reliable variance (i.e., ωH divided by ω) in the MRNI-SF total score was due to the general factor, which means that the general TMI factor is the primary influence on raw total score variation. Thus, model-based reliability estimates provided evidence supporting the use of the raw MRNI-SF total score to represent the general TMI construct.

In contrast to the five other specific factors, the SR (ωHS = .56) and NT (ωHS = .52) specific factors accounted for 63% and 58%, respectively, of their corresponding subscale's true score variance (i.e., ωHS divided by ω). These results suggest that the raw SR and NT subscale scores primarily measure their intended subdomain construct but also re-measure, to some degree, the general TMI construct (which is not desirable). The other specific factors accounted for less reliable variance and were better measures of the general TMI construct than of their intended subdomain constructs. Because no specific factor met the preferred ωHS > .75 benchmark in any of the seven subsamples, we suggest researchers use caution when considering the use of raw subscale scores as measures of the subdomain constructs, although their use as latent variables in a SEM context is appropriate.

Multi-Group Measurement Invariance

After examining the total sample, we created several comparison groups to identify differences between participants who did and did not fit the White heterosexual male TMI reference group. Specifically, after removing individuals who did not report key demographic information in the full sample, 1,939 heterosexual White men, 853 heterosexual White women, 222 heterosexual Black men, 506 heterosexual Asian men, and 404 gay White men were selected.[3] Demographic information, correlations between MRNI-SF total and subscale scores, means, standard deviations, factor loadings, and model fit indices for each of these groups are presented in the online supplementary materials. To provide a precise depiction of how each item functioned in relation to the general TMI factor and its sole *intended* specific factor, we used the more conservative orthogonal bifactor model specification for each invariance analysis. Additional details of these analyses are available in the online supplementary materials. Table 5 displays the fit statistics of each invariance model.

White Heterosexual Men compared to White Heterosexual Women. A configural invariance model, in which both groups were estimated simultaneously with no cross-group equality constraints, provided an acceptable fit (see Table 5). Next, using the configural model as a baseline, we followed traditional measurement invariance testing procedures and created a nested model in which the factor loadings for the general and specific factors were constrained to be equal between men and women (i.e., metric invariance). The nested model was a significantly worse fit than the configural model, as evidenced by the scaled chi-square difference test, $\Delta\chi^2$ (34) = 173.31, p = .008, but the nested model also evidenced a marginally acceptable change in CFI (ΔCFI = .007).

Highlighting the potential source of the measurement non-invariance, the bootstrapping procedure suggested that all items on the specific factor were invariant, but all 21 items were non-invariant for the general factor. These results were consistent with the partial measurement invariance model identified by Levant and colleagues (2013). Thus, we tested a partial metric model freeing the cross-group equality constraints on *all* the factor loadings of the general factor. The partial metric model provided an acceptable fit that was not significantly different from the configural model, $\Delta\chi^2$ (14) = 6.96, p = .936, ΔCFI = 0. Thus, the partial metric model was retained for testing scalar invariance.

Using the partial metric invariance model as a base, we tested a scalar invariance model by constraining the intercepts to be equivalent between men and women. The scalar invariance model provided marginally acceptable fit as a whole. However, the scaled chi-square difference test indicated that the scalar model was a significantly worse fit than the partial metric invariance model, $\Delta\chi^2$ (21) = 525.84 p < .001, the ΔCFI was .026, and the bootstrap confidence interval of the differences in intercepts between men and women indicated that all 21 MRNI-SF items were non-invariant. Therefore, the MRNI-SF did not demonstrate scalar invariance, because men's intercept values were significantly greater than women's intercept values across the board. The lack of scalar invariance indicated that further testing of residuals invariance was inappropriate (Kline, 2016).

White Heterosexual Men compared to Black Heterosexual Men. Using the same procedures employed for assessing measurement invariance between men and women, we first tested a configural invariance model for White heterosexual men compared to Black heterosexual men, which evidenced acceptable fit (see Table 5). Of note, the configural invariance model also evidenced a negative error variance. As mentioned previously, such Heywood cases (c.f., Kline, 2016) are common in bifactor models and can be corrected by re-specifying the model (e.g. Chen et al., 2006). Specifically, because the negative error variance was non-significant and relatively small (-.65, p = .70), we followed the recommendations of Muthén (2007) and constrained this value to zero to resolve the non-positive definite matrix.

Next, we examined a metric invariance model. The chi-square difference test comparing the metric invariance model to the configural invariance model was non-significant ($\Delta\chi^2$ [34] = 41.78, *p* = .168), and the ΔCFI was .001. The bootstrap procedures indicated that two items may be non-invariant for the general factor only: one from RE and one from DO. However, the evidence was inconclusive; because these items yielded non-invariance in only 2 out of the 21 referent possibilities (see online supplemental materials). We thus used a full metric model to form the base for scalar invariance. The scaled-chi-square difference test suggested that the scalar model was a significantly worse fit than the metric model ($\Delta\chi^2$ [21] = 129.83, *p* < .001); however, the ΔCFI was .004. Despite the marginally acceptable change in CFI, all but four of the 21 item intercepts yielded significant differences between White men and Black men, as evidenced by the bootstrap procedure (see online supplementary materials). Thus, scalar invariance was not supported and further testing was inappropriate.

White Heterosexual Men compared to Asian Heterosexual men. A configural invariance model and a metric invariance model yielded acceptable fit (see Table 5). However, the scaled chi-square difference test suggested that the metric invariance model was a significantly worse fit than the configural invariance model, $\Delta\chi^2$ (34) = 86.59, *p* < .001, despite an acceptable ΔCFI of .002. The bootstrap procedure revealed that four items were most consistently non-invariant on the general factor: item M4 (“Men should watch football games instead of soap operas”), item M8 (“A man should prefer watching action movies to reading romantic novels”), item M16 (“Men should be detached in emotionally charged situations”), and item M17 (“It is important for a man to take risks, even if he might get hurt”). A scaled chi-square difference test indicated that a partial metric invariance model, in which the four non-invariant items were freed, was not an equivalent fit compared to the configural model, $\Delta\chi^2$ (30) = 62.80, *p* < .001, although the ΔCFI was .001. Thus, the evidence for measurement invariance was equivocal. Given that two out of three measurement invariance tests supported a partial metric invariance model; however, we proceeded to examine possible scalar invariance.

Using the partial metric model as a base, the scaled chi-square difference test indicated that the scalar model was a worse fit than the partial metric model, $\Delta\chi^2$ (19) = 120.045, *p* < .001. Interestingly, despite a significant chi-square difference, the ΔCFI was marginally acceptable (.004), and the bootstrap confidence interval tests indicated that only eight of the 21 intercepts were non-invariant. Asian men yielded larger intercepts compared to White men for six of the eight items. Because only eight of the 21 intercepts were non-invariant (i.e., less than half; c.f., Cheung & Lau, 2012), we examined a partial scalar model freeing the non-invariant intercepts. The partial scalar model provided acceptable fit; however, the partial scalar model evidenced a worse fit compared to the partial metric model, $\Delta\chi^2$ (13) = 34.411, *p* < .001, despite also yielding a ΔCFI of .001. Thus, the evidence for or against retaining the partial scalar model was somewhat equivocal, though two out of three procedures supported a partial scalar model. We therefore examined a residuals invariance model as an exploratory analysis.

The residuals invariance model provided acceptable fit. The chi-square difference test, however, suggested the residual model was a worse fit than the partial scalar model, $\Delta\chi^2$ (25) = 120.18, *p* < .001. The bootstrap procedure suggested a partial residual model, because only 4 out of the 21 residuals were non-invariant: item M1 (“Homosexuals should never marry”), item M8 (“A man should prefer watching action movies to reading romantic novels”), item M10 (“Boys should prefer to play with trucks rather than dolls”), and item M15 (“A man should never admit when others hurt his feelings”). After freeing four non-invariant residuals identified by the bootstrap procedure, the partial residuals model was still a worse fit than the partial scalar model, $\Delta\chi^2$ (21) = 67.46, *p* < .001, but the ΔCFI of .001 was minimal. Thus, the evidence for residuals invariance was, again, somewhat equivocal, with two out of three tests providing support.

Heterosexual White men compared to Gay White men. A configural invariance model provided and a metric invariance model provided acceptable fit. The scaled chi-square difference test approached a non-significant difference between the configural model and the metric model, $\Delta\chi^2$ (34) = 49.02, *p* = .046. The change in CFI was also zero, but the bootstrap procedure revealed non-invariant items. One item (M12; “A man should always be the boss”) was non-invariant on the specific factor and two items were consistently non-invariant on the general factor: item M9 (“Men should always like to have sex”) and M19 (“When the going gets tough, men should get tough”). Accordingly, we examined a partial metric model freeing the non-invariant items. There was no difference in CFI or scaled chi-square between the partial metric and the configural model, $\Delta\chi^2$ (31) = 32.21, p = .407.

Next, the scalar model yielded acceptable fit; however, the scaled chi-square difference test indicated that this model was a worse fit than the partial metric model, $\Delta\chi^2$ (18) = 94.79, p < .001. Although the ΔCFI was acceptable (.004), the bootstrap confidence interval tests suggested that 14 of the 21 intercepts were non-invariant. Therefore, scalar invariance was not supported, and the preponderance of non-invariant items suggested that testing a partial-scalar invariance or residual invariance would be inappropriate (e.g., Kline, 2016). For all 14 items, heterosexual men evidenced significantly higher intercept values than gay men (see online supplementary materials).

Discussion

Measuring traditional masculinity ideology (TMI) is critical for researchers and clinicians working with men. The present study provided an in-depth examination of the psychometric properties and factor structure of the short form of a widely-used measure of TMI, the Male Role Norms Inventory (MRNI-SF). Specifically, we confirmed the bifactor structure of the instrument for modeling a general TMI variable and seven narrower domains of TMI, and we calculated several ancillary bifactor indices consistent with recent best-practice recommendations (i.e., Rodriguez et al., 2016) but previously unexamined for the MRNI-SF. Additionally, we tested the MRNI-SF for measurement invariance across groups which do and do not reflect the White, male, Eurocentric, and heterosexual aspects of TMI. Although no hypotheses were advanced regarding higher levels of measurement invariance (i.e. metric, scalar, or residuals) or the specific values of each ancillary bifactor index, we hypothesized that (a) the bifactor model would represent an acceptable fit, (b) a bifactor model would represent a better fit compared to a second-order model, and (c) a bifactor model would demonstrate, at least, configural invariance.

In support of our first two hypotheses, a bifactor structure yielded acceptable fit in the total sample and each sub-group (i.e., heterosexual White women, heterosexual White men, heterosexual Black and Asian men, and gay White men). A bifactor model also evidenced a statistically superior fit compared to a unidimensional model and a second-order model, a finding that is consistent with recent factor-structure studies of the MRNI-SF (Levant et al. 2013; 2015; 2016). Furthermore, the present findings help to clarify the specification of orthogonality constraints on a bifactor model of the MRNI-SF, considering that Reise (2012) has argued that a bifactor model should be completely orthogonal, but the instrument's specific factors have been modeled in previous research as oblique (Levant et al., 2013) or as partially oblique (Levant et al., 2015). Our results suggest that the MRNI-SF can be modeled as a completely orthogonal or partially oblique model without any technical errors. However, the original specification of the instrument (i.e., a completely oblique bifactor model) may produce technical errors. Researchers should also keep in mind that a full orthogonal bifactor model provides the most "pure" variance partitioning effects, which is essential for measurement diagnostic purposes (e.g., Reise et al., 2016).

Internal Structure and Reliability

Because the orthogonal bifactor model also provided adequate fit to the MRNI-SF, it became possible to examine more precisely the degree of multidimensionality versus unidimensionality of the instrument, as well as the reliability of the MRNI-SF's general and specific factors. Ancillary bifactor indices revealed that, although the unidimensional structure evidenced poor fit, modeling the instrument using a unidimensional solution may not adversely affect the 21 items' ability to measure the general TMI construct. Said another way, whether the MRNI-SF is modeled using a unidimensional or bifactor solution, item factor loadings on the general factor are of a similar magnitude, suggesting a lack of relative measurement parameter bias (see Rodriguez et al., 2016, p. 145). Furthermore, the model-based reliability of the MRNI-SF total score reached recommended levels (see Reise et al., 2013, p. 137), suggesting that the MRNI-SF's total score primarily reflects a single general TMI factor. In sum, these results suggest that it may be permissible for researchers and clinicians to interpret the raw MRNI-SF total score as a sufficiently reliable and appropriate measure of the general construct of TMI.

Ancillary bifactor indices also indicated that most of the raw subscale scores do not capture enough reliable, unique variance (beyond that accounted for by the general TMI factor) to justify their calculation or interpretation outside of a bifactor framework. In other words, our findings suggest that the raw subscale scores primarily re-measure the general TMI factor, rather than the narrow subdomain construct the subscale score was designed to measure. Professional standards indicate that subscale scores must show distinctiveness and reliability as a prerequisite to their use in research and practice (AERA et al., 2014, p. 27). Thus, interpreting the raw MRNI-SF subscale scores as if they are meaningfully measuring their specific ideologies may be misleading. Nevertheless, it is important to remind the reader that, just because a raw subscale score should not be used to measure a given construct does not disqualify the use of the corresponding latent factor score in the context of bifactor SEM models. Indeed, in a recent convergent validity study of the MRNI-SF, Levant et al., (2016) identified that four of the seven specific factors (two were not examined) were significant predictors of theoretically similar constructs in a structural model with an oblique bifactor measurement component. Consistent with the present findings emphasizing the importance of the TMI general factor, Levant et al also found that the TMI general factor evidenced the most robust relationships with convergent validity measures.

That being said, two raw subscale scores (for self-reliance though mechanical skills and negativity toward sexual minorities) consistently achieved the minimum (but not the preferred) model-based reliability thresholds suggested by Reise and colleagues (2013) in the present study. Therefore, these two raw subscale scores may account for enough reliable, unique variance such that their use in future research may be warranted. Users should bear in mind, however, that these raw subscale scores are partially measuring the intended subdomain construct but also partially re-measuring the general TMI construct.

Measurement Invariance

Different patterns of measurement invariance emerged suggesting some between-group differences in the meaning (i.e., metric invariance) scaling (i.e., scalar invariance) and precision (i.e., invariance of residuals) of MRNI-SF scores. Consistent with our hypotheses, configural invariance was supported for each group comparison, indicating that the MRNI-SF is accurately represented by a general TMI factor and the seven TMI domain specific factors across each group. Configural invariance is the least restrictive level of measurement invariance (Kline, 2016), so it is not surprising that our results also yielded mixed support for the MRNI-SF with respect to more restrictive-levels of invariance according to three major criteria: a non-significant chi-square difference test, a change of CFI less than .01, and non-significant between-group bootstrapped differences on each parameter of interest.

Full metric invariance was supported only for heterosexual White men compared to heterosexual Black men, but partial metric invariance was supported for all other group comparisons. These findings indicate that, although the MRNI-SF appears to be capturing both general and specific factors of TMI based on White, male, Eurocentric, and heterosexual cultural values, those factors may represent somewhat different constructs in other cultural groups. Of note, most comparisons evidenced non-invariant factor loadings on the general factor, suggesting that the major differences across cultural groups appear to be on the general TMI construct and not the specific aspects of TMI.

In general, only a few items on the TMI factor produced non-invariant factor loadings across race and sexual orientation groups, suggesting that the between-group differences on the meaning of overall TMI may be relatively trivial. However, when comparing men to women, it is noteworthy that *all* of the items on the general factor were non-invariant, and that this was the only group comparison to yield such a result. Our findings are consistent with Levant and colleagues' (2013) results and may indicate a potential out-group homogeneity effect for gender (see Rubin & Badea, 2007 for a review). Specifically, one possibility for future research is that men and women have a shared understanding of specific aspects of masculinity, because messages about narrow aspects of masculinity are highly prevalent in the broader media and culture. However, a general conceptualization of masculinity (i.e., one’s personal ideology about men overall) may be more abstract and created through personal experiences that vary based on whether one is in the in-group (men) or the out-group (women). Such personal experiences may fundamentally change the meaning of overall TMI for men compared to women, in that a man’s perception of overall masculinity may be more nuanced than women’s.

Several items also produced non-invariant factor loadings on the general TMI factor when comparing heterosexual White men to heterosexual Asian men; however, there were still enough invariant loadings to suggest partial metric invariance of the general TMI construct. Although there were more similarities than difference in overall TMI between White and Asian men in the present sample, four items on the general TMI factor were consistently non-invariant. The content of these items reflected the importance of avoiding feminine behaviors, acting tough, and being stoic in the face of emotional situations. However, there were no differences on the factor loadings for the specific factors corresponding to these four items. Our results, therefore, indicate that Asian men and White men may share a similar conception of specific aspects of TMI, but, when combined into an overall concept of TMI, cultural differences may become apparent. Our findings are consistent with cross-cultural previous conclusions that Asian cultures do not emphasize hegemonic aspects of masculinity, particularly avoidance of femininity (see Iwamoto and Kaya, 2016 for a review). However, our results are inconsistent with a finding that another popular measure of TMI, the Conformity to Masculine Role Norms Inventory-46 (CMNI-46; Parent & Moradi, 2011), was largely non-invariant between White and Asian men (Hsu & Iwamoto, 2014). The CMNI-46 and MRNI-SF overlap in some instances but measure TMI in different ways. Specifically, the MRNI-SF taps perspectives of what men “should” be and do from a third-person perspective (e.g., “Men should be detached in emotional situations”), whereas the CMNI-46 measures conformity to TMIs from a first-person perspective (e.g., “I never share my feelings”). Therefore, the present findings raise the possibility that the item reference (i.e., other versus self) may influence the meaning of TMI between White and Asian men. However, given that partial invariance (i.e., metric, scalar, and residuals) was supported for Asian men in the present sample, additional research is needed to determine how subtle differences impact, if at all, the correlates of TMIs in Asian and White samples.

When comparing White heterosexual men to White Gay men, one item on the general factor and two items on the specific factor produced non-invariant factor loadings. The content of these items reflected hegemonic male norms of dominance, toughness, and importance of sex. However, because partial metric invariance was met in each instance, our findings suggest that heterosexual and gay men predominately share the same conceptions of TMI measured by the MRNI-SF. These findings are consistent with assertions that many gay men include traditional, patriarchal (sometimes exaggerated) perspectives in their definitions of masculinity (see Sánchez, 2016 for review).

Contrary to our results regarding the potential meaning of TMI (i.e., metric invariance), White heterosexual women, Black heterosexual men, and White gay men all failed tests of scalar invariance when compared to White heterosexual men. Our results suggested that groups that do not fit the traditional TMI reference group, with the exception of Asian men, may exhibit a differential-additive response style (Cheung & Rensvold, 2000). In other words, the differences between item intercepts among these groups may be due to additive systemic influences (such as cultural worldviews) which impact the way individuals respond to specific items but not the meaning of each construct (Kline, 2016). Indeed, it is not surprising that White heterosexual men's intercepts were higher than women's, considering that men are socialized directly to think, feel, and behave in ways consistent with TMI (Levant & Richmond, 2016). Black heterosexual men's intercepts were also significantly higher than their White peers in the present study, consistent with TMI theories emphasizing Black men's exaggerated adherence to traditional male roles as a possible coping mechanism against systemic inequality (e.g., Majors & Billson, 1993). Likewise, White heterosexual men's intercepts were higher than their gay comparison group, and, considering that heterosexism and TMI are related constructs (Levant & Richmond, 2016), it is possible that this could be due to pre-existing values about gay men.

The scalar invariance test also revealed that Asian men had higher intercept values on six of the eight non-invariant items, with item content largely reflecting restrictive emotionality and toughness domains. The remaining two non-invariant intercepts reflected avoidance of femininity norms, and these intercept values were larger for White participants. These results are consistent with previous findings emphasizing the importance of emotional restraint in Asian cultures (Wong, Nguyen, Wang, Chen, & Steinfeldt, 2012), which may facilitate more endorsement of emotional control TMI. Some evidence also suggests that Asian men may be less likely to define their masculinity in opposition to femininity than White men (e.g., Chua & Fujino, 1999), which may suppress their endorsement of avoidance of femininity. However, although these individual item intercepts were non-invariant, the fact that partial scalar invariance was supported by two of the three invariance procedures suggests that the MRNI-SF subscale and total scores may generally reflect comparable scaling between Asian and White men. Indeed, we also found evidence of partial residual invariance, indicating the MRNI-SF may also measure TMI with the same degree of precision among White and Asian men. Future research is warranted to continue exploring the TMI construct in Asian men.

Limitations and Directions for Future Research

The present findings should be interpreted with respect to several key limitations. Most notably, although the sample was large and drawn from both community and college sources, it was still a convenience sample, and it is possible that participants may have self-selected. Furthermore, the sample lacked sufficient diversity to examine other racial or ethnic groups. Additional research is needed using more sophisticated sampling procedures to gather a representative sample of the United States population. Relatedly, the percent of community participants was relatively low, and, we lacked important information to determine how many participants in the internet samples were currently attending a university. It is possible that some of the observed similarities in the meaning of TMI across racial categories may have been due to the socializing effects of attending college in the United States. Measuring race through a categorical variable may have also obscured important within-group variability (e.g. racial identity or socioeconomic status). Likewise, although the sample was large, there were insufficient numbers to test race and gender interactions (i.e., White men vs White women, or White women vs. Black men), indicating a further need to examine invariance of the MRNI-SF across additional cultural groups. Lastly, it is important to remember that measurement invariance only means that the constructs tapped by the instrument appeared to be generally similar across different groups, but it is still possible that each group has certain culturally defined characteristics of masculinity which are not measured by the MRNI-SF.

Implications for the Use of the MRNI-SF

Despite the aforementioned limitations, the present findings offer several recommended and contraindicated uses of the MRNI-SF. First, it seems permissible to calculate and interpret the raw MRNI-SF total score as a measure of the general TMI construct. Second, it may be permissible to calculate and interpret the raw SR and NT subscale scores as imperfect yet still potentially useful measures of their intended subdomain constructs. Users who choose to use these two raw subscale scores must remind themselves and their readers that these scores are contaminated to some degree by the general TMI factor, which can complicate interpretation. Third, the use of raw subscale scores for the other five subscales is contraindicated. Fourth, it is a best practice to use SEM to model the general and specific factors of the MRNI-SF in the context of a bifactor solution, as this allows the precise measurement of orthogonal factor scores for all eight constructs. Of note, the use of a CFA bifactor solution allows for more precise investigation of how the general and specific masculinity constructs uniquely relate to external criteria. This can help answer questions about the incremental validity of the specific factors over and above the powerful general TMI factor. Fifth, when SEM is not available, users can use an ipsatization approach to partial out the unique variance due to the specific factors from the variance due to the general TMI factor. Ipsatization involves subtracting each respondent’s score on each of the 21 MRNI-SF items by that respondent’s mean MRNI-SF total score, resulting in 21 ipsatized item scores whose values represent deviations from that respondent’s mean TMI (Greer & Dunlap, 1997). For example, a positive score for a given ipsatized item would indicate that respondents scored higher on that item relative to their average. However, users should be warned that interpretation of raw ipsatized subscale score—calculated by taking the mean or sum of ipsatized items for that subscale—is more nuanced and complicated. For example, the finding that an ipsatized subscale SR score is correlated with, say, stress, would indicate that respondents who endorse self-reliance beliefs to a stronger degree *relative to other traditional masculine ideology beliefs* tended to report greater stress.

Lastly, our measurement invariance results suggest that gender, race, and sexual orientation categories are important in the measurement of TMI. In particular, we recommend using caution when interpreting raw mean differences of the MRNI-SF scores when comparing White men to women, Black men, or Gay White men. Because item intercepts represent the zero point of each construct, mean differences on the MRNI-SF may not reflect true differences in TMI but rather culturally influenced response patterns that artificially inflate or deflate the scaling of TMI. Response weights (c.f., Kline, 2016) may be needed to address scalar non-invariance by weighting the means for different groups when comparing between White heterosexual men and racial or sexual minority men. However, because configural and (at least) partial metric invariance were supported for all groups, the bifactor modeling of the instrument generally appears to be an appropriate method of measuring TMI across the cultural identities assessed in the present study.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Brannon, R., & Juni, S. (1984). A scale for measuring attitudes about masculinity. *Psychological Documents*, *14*(1). (Document No. 2612)

Connell, R. W., and Messerschmidt, J. W. (2005). Hegemonic masculinity: Rethinking the concept. *Gender & Society*, *19*, 829-859.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189–225. https://doi.org/10.1207/s15327906mbr4102_5

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, *15*(2), 167–198. https://doi.org/10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*(1), 1–27. https://doi.org/10.1177/014920639902500101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Chua, P., & Fujino, D. C. (1999). Negotiating new Asian-American masculinities: Attitudes and gender expectations. *The Journal Of Men's Studies*, *7*(3), 391-413. doi:10.3149/jms.0703.391

Greer, T., & Dunlap, W. P. (1997). Analysis of variance with ipsative measures. *Psychological Methods, 2*, 200–207

Hammer, J. H., & Toland, M. D. (2016, November). *Bifactor analysis in Mplus*. [Video file]. Retrieved from http://sites.education.uky.edu/apslab/upcoming-events/

Hsu, K., & Iwamoto, D. K. (2014). Testing for measurement invariance in the Conformity to Masculine Norms-46 across White and Asian American college men: Development and validity of the CMNI-29. *Psychology of Men & Masculinity*, *15*(4), 397–406. https://doi.org/10.1037/a0034548

Iwamoto, D. K., & Kaya, A. (2016). Asian American men. In Y. J. Wong, S. R. Wester, Y. J. Wong, S. R. Wester (Eds.), *APA handbook of men and masculinities* (pp. 285-297). Washington, DC, US: American Psychological Association. doi:10.1037/14594-013

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York NY: Guilford.

Levant, R. F., Hall, R. J., & Rankin, T. J. (2013). Male Role Norms Inventory-Short Form (MRNI-SF): Development, confirmatory factor analytic investigation of structure, and measurement invariance across gender. *Journal of Counseling Psychology*, *60*, 228–238. 10.1037/a0031545

Levant, R. F., Hall, R. J., Weigold, I., & McCurdy, E. R. (2015). Construct distinctiveness and variance composition of multi-dimensional instruments: Three short-form masculinity measures. *Journal of Counseling Psychology, 62*, 488-502.

Levant, R. F., Hall, R. J., Weigold, I., & McCurdy, E. R. (2016). Construct validity evidence for the male role norms inventory-short form: A structural equation modeling approach using the bifactor model. *Journal of Counseling Psychology, 62* (3), 488-502. doi:10.1037/cou0000092

Levant, R. F., Hirsch, L. S., Celentano, E., Cozza, T. M., Hill, S., MacEachern, M., . . .Schnedeker, J. (1992). The male role: An investigation of contemporary norms. *Journal of Mental Health Counseling, 14*, 325–337.

Levant, R. F., Gerdes, Z. T., Alto, K. M., Jadaszewski, S. & McDermott, R. C. (in press). Development and evaluation of the Fathers' Expectations About Sons' Masculinity Scale (Short Form). *Psychology of Men & Masculinity.*

Levant, R. F., Majors, R. G., & Kelley, M. L. (1998). Masculinity ideology among young African American and European American women and men in different regions of the United States. *Cultural Diversity and Mental Health*, *4*(3), 227–236.

Levant, R. F., & Richmond, K. (2007). A review of research on masculinity ideologies using the Male Role Norms Inventory. *The Journal of Men's Studies, 15*, 130-146. doi:10.3149/jms.1502.130

Levant, R. F., & Richmond, K. (2016). The gender role strain paradigm and masculinity ideologies. In Y. J.Wong & S. R.Wester (Eds.), *APA handbook on men and masculinities* (pp. 23–49). Washington, DC: American Psychological Association. doi:10.1037/14594-002

Levant, R. F., Richmond, K., Majors, R. G., Inclan, J. E., Rossello, J. M., Heesacker, M., … Sellers, A. (2003). A multicultural investigation of masculinity ideology and alexithymia. *Psychology of Men & Masculinity*, *4*(2), 91–99. https://doi.org/10.1037/1524-9220.4.2.91

Majors, R., & Billson, J. M. (1993). *Cool pose*: *The dilemmas of black manhood in America.* New York, NY: Touchstone Books/Simon & Schuster.

Mahalik, J. R., Locke, B. D., Ludlow, L. H., Diemer, M. A., Scott, R. J., Gottfried, M., & Freitas, G. (2003). Development of the Conformity to Masculine Norms Inventory. *Psychology of Men & Masculinity, 4,* 3-25. doi:10.1037/1524-9220.4.1.3

McDermott, R. C., Cheng, H., Lopez, F. G., McKelvey, D., & Bateman, L. S. (2016). Dominance orientations and psychological aggression in college student relationships: A Test of an Attachment Theory-Guided Model. *Psychology Of Violence,* doi:10.1037/vio0000061

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LawrenceErlbaum Associates. doi:10.1111/j.2044-8317.1981.tb00621.x

Muthén, L. K. (2007). Negative residual Variance [online discussion group]. Retrieved from http://www.statmodel.com/discussion/messages/9/572.html?1461050621

Muthén, B., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Parent, M. C., & Moradi, B. (2011). An abbreviated tool for assessing conformity to masculine norms: Psychometric properties of the Conformity to Masculine Norms Inventory-46. *Psychology of Men & Masculinity*, *12*(4), 339–353. https://doi.org/10.1037/a0021904

Pleck, J. H. (1995). The gender role strain paradigm: An update. In R. F.Levant & W. S.Pollack (Eds.), *A new psychology of men* (pp. 11–32). New York, NY: Basic Books.

Reise, S. P. (2012). Invited paper: The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5–26. https://doi.org/10.1177/0013164412449831

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rogers, B. K., Sperry, H. A., & Levant, R. F. (2015). Masculinities among African American men: An intersectional perspective. *Psychology Of Men & Masculinity, 16*(4), 416-425. doi:10.1037/a0039082

Rubin, M., & Badea, C. (2007). Why do people perceive ingroup homogeneity on ingroup traits and outgroup homogeneity on outgroup traits?. *Personality And Social Psychology Bulletin, 33*(1), 31-42. doi:10.1177/0146167206293190

Sánchez, F. J. (2016). Masculinity issues among gay, bisexual, and transgender men. In Y. J. Wong, S. R. Wester, Y. J. Wong, S. R. Wester (Eds.), *APA handbook of men and masculinities* (pp. 339-356). Washington, DC, US: American Psychological Association. doi:10.1037/14594-016

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183–206). New York, NY: Routledge/Taylor & Francis Group.

Thompson, E. H., & Pleck, J. H. (1986). The structure of male norms. *American Behavioral Scientist*, *29*, 531-543.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158. https://doi.org/10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Wong, Y. J., Nguyen, C. P., Wang, S., Chen, W., Steinfeldt, J. A., & Kim, B. K. (2012). A latent profile analysis of Asian American men's and women's adherence to cultural values. *Cultural Diversity And Ethnic Minority Psychology*, *18*(3), 258-267. doi:10.1037/a0028423

Footnotes

[1] Because there was no “some college” option, it is possible that many participants selected “high school diploma” as their highest level of education, but they were actually currently attending college.

[2] Because a second-order model is nested within a bifactor model, a logical question arises as to whether a second-order model with the same modifications would be a worse fit compared to the modified bifactor model. Thus, a second-order model with correlations between disturbance terms of the lower-order factors was tested against the modified bifactor model. The scaled chi-square difference test was significant, indicating the modified second-order model was a worse fit compared to the less parsimonious modified bifactor.

[3] The study sample of Latino heterosexual men (*N* = 153) did not meet the minimum sample size recommendations of 200 or more suggested by Kline (2016). Thus, we excluded comparisons between White heterosexual men and Latino heterosexual men. Because differences could be identified between gay and heterosexual men, it was critical to exclude gay men from any other group comparisons to more accurately locate the source of any between-group differences.

Table 1

MRNI-SF subscale and total score interrcorrelations, means, and standard deviations for the total sample.

	1	2	3	4	5	6	7	8
1. AoF	---	.54**	.47**	.55**	.61**	.65**	.58**	.84**
2. NTSM		---	.31**	.35**	.60**	.39**	.41**	.68**
3. SRMS			---	.57**	.33**	.44**	.39**	.69**
4. Tough				---	.41**	.55**	.55**	.77**
5. Dom					---	.53**	.55**	.74**
6. IoS						---	.56**	.78**
7. RE							---	.75**
8. MRNI								---
Mean	2.68	2.04	4.28	3.76	1.77	2.57	2.14	2.75
SD	1.65	1.56	1.69	1.58	1.19	1.52	1.20	1.12

Note. *N* = 6,744. AoF = Avoidance of Femininity, NTSM = Negativity towards Sexual Minorities, SRMS = Self-Reliance through Mechanical Skills, Tough = Toughness, Dom = Dominance, IoS = Importance of Sex, RE = Restrictive Emotionality, MRNI = Male Role Norms Inventory, *M* = Mean, *SD* = Standard Deviation.

***p* < .001.

Table 2

Measurement model indices of fit for combined sample of White Heterosexual Men and Women, Heterosexual Black and Asian men, and White Gay men.

Model	Chi-Square	*Df*	Scaling Correction Factor	CFI	TLI	RMSEA	RMSEA 90% CI	SRMR	AIC	BIC
Common-Factor	2665.921*	168	1.3406	.960	.951	.047	.045, .049	.034	448405.728	448978.306
Unidimensional	22750.751*	189	1.4448	.643	.603	.133	.132, .135	.097	477659.020	478088.454
Second-Order	3909.665*	181	1.3466	.941	.931	.055	.054, .057	.054	450070.475	450554.440
Bifactor-Oblique	1532.642*	148	1.3233	.978	.969	.037	.036, .039	.021	446899.995	447608.901
Bifactor-Orthogonal	3233.703*	168	1.3344	.951	.939	.052	.050, .054	.051	449147.035	449719.613
Bifactor-Modified	2149.509*	166	1.3255	.969	.960	.042	.041, .044	.030	447552.570	448145.597

Note. *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; and SRMR = Standardized Root Mean Square Residual. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

* *p* < .001.

Table 3

Model comparison tests between Nested and Non-nested models with Acceptable Fit

Model Comparisons	Scaled Chi-Square Difference	Δ*df*	Conclusion
Second-Order vs. Bifactor oblique	2230.4572***	33	Retain Bifactor
Second-Order vs. Bifactor orthogonal	631.3407***	13	Retain Bifactor
Second-Order vs. Bifactor modified	1528.7453***	15	Retain Bifactor

Note. *df* = degrees of freedom; CFI = Δ = delta (i.e., change between nested and comparison models). AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion. Negative change values for AIC or BIC indicate that the nested model was a worse fit compared to the comparison model.

*** *p* < .001.

Table 4

Explained Common Variance and Model-Based Reliability Estimates for the MRNI-SF

	General Factor	Restricted Emotionality	Self-Reliance	Negativity Toward Sexual Minorities	Avoidance of Femininity	Importance of Sex	Dominance	Toughness
Omega	.96	.80	.89	.90	.88	.86	.88	.78
Omega Hierarchical	.87	.01	.03	.02	.01	.01	.01	.01
Omega Hierarchical Subscale	---	.27	.56	.52	.17	.28	.37	.30
ECV	.58	.03	.11	.09	.03	.05	.04	.06

Note. ECV = Explained Common Variance.

Table 5.

Indices of fit for each Measurement invariance model

Group comparisons	Invariance Model	χ^2	*df*	CFI	TLI	RMSEA	RMSEA 90% CI	SRMR
Men vs. Women								
	Configural	1416.697*	336	.946	.932	.048	.045 .051	.053
	Metric	1595.107*	370	.938	.930	.049	.046 .051	.070
	Partial-metric	1414.377*	350	.946	.936	.047	.044 .049	.053
	Scalar	1950.212*	371	.920	.910	.055	.053 .058	.099
	Partial Scalar	---	---	---	---	---	---	---
	Residuals	---	---	---	---	---	---	---
	Partial Residual	---	---	---	---	---	---	---
White vs. Black								
	Configural	1335.395*	337	.955	.944	.052	.049 .055	.051
	Metric	1388.322*	371	.954	.948	.050	.048 .053	.053
	Partial-metric	---	---	---	---	---	---	---
	Scalar	1509.606*	392	.95	.946	.051	.049 .054	.063
	Partial Scalar	---	---	---	---	---	---	---
	Residuals	---	---	---	---	---	---	---
	Partial Residual	---	---	---	---	---	---	---
White vs. Asian								
	Configural	1486.496*	336	.952	.941	.053	.050 .056	.053
	Metric	1570.384*	370	.950	.944	.052	.049 .054	.057
	Partial-metric	1544.179*	366	.951	.944	.051	.049 .054	.055
	Scalar	1656.231*	385	.947	.943	.052	.049 .055	.057
	Partial Scalar	1585.044*	379	.950	.945	.051	.048 .054	.056
	Residuals	1710.911*	404	.946	.944	.051	.049 .054	.058
	Partial Residual	1644.810*	400	.949	.946	.050	.048 .053	.058
Heterosexual vs. Gay								

	Configural	1285.755*	336	.948	.935	.049	.046 .052	.050
	Metric	1329.911*	370	.948	.940	.047	.044 .050	.054
	Partial-metric	1312.601*	367	.948	.941	.047	.044 .050	.053
	Scalar	1414.965*	388	.944	.939	.048	.045 .050	.058
	Partial Scalar	---	---	---	---	---	---	---
	Residuals	---	---	---	---	---	---	---
	Partial Residual	---	---	---	---	---	---	---

*p < .001

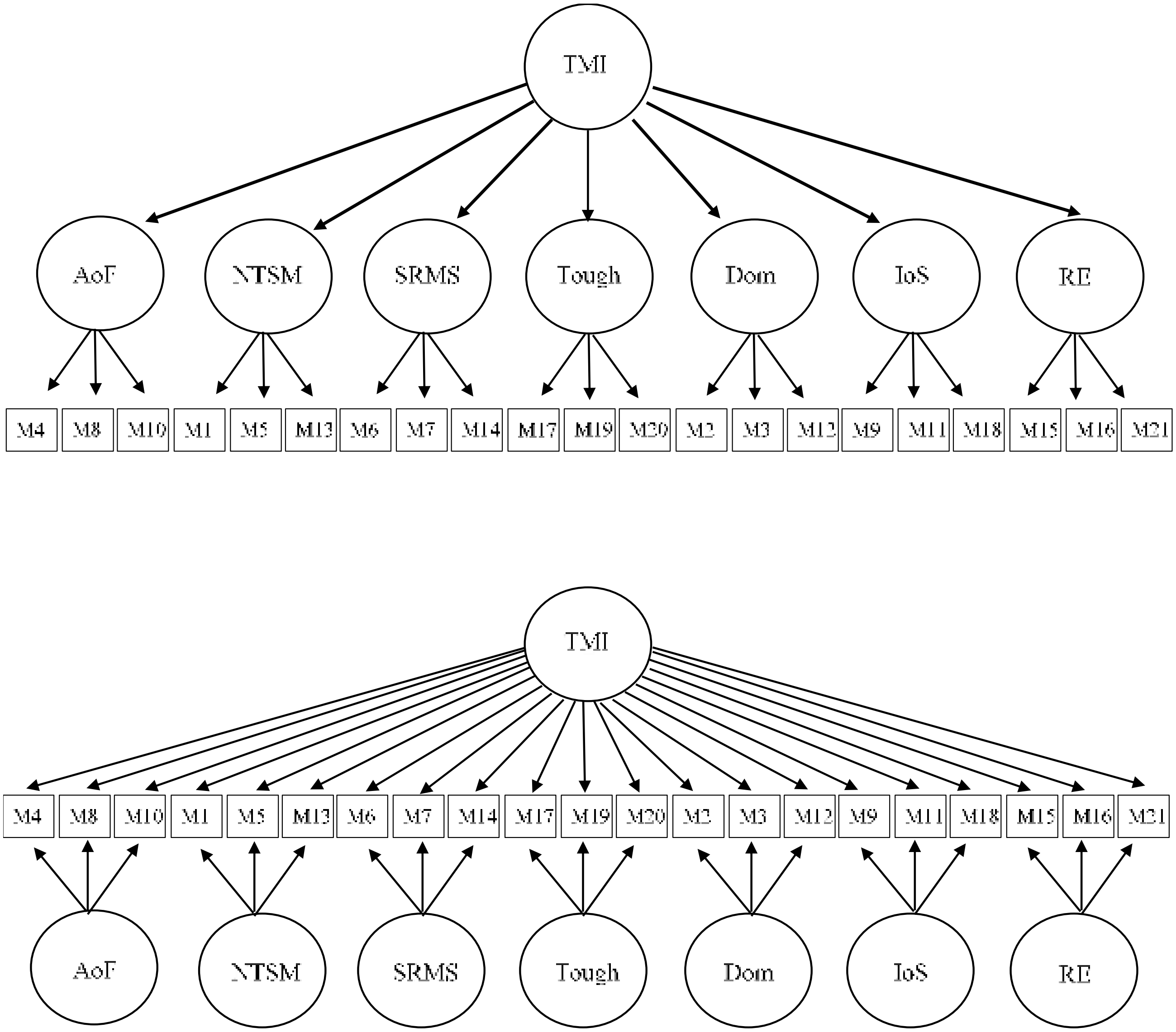

Figure 1a & 1b

Second-order model pictured on top (1a). Bifactor model pictured on the bottom (1b). For each model, the first item of each factor was constrained to 1 to scale the metric. For both models, covariances and error terms and disturbance terms are not pictured for readability. AoF = Avoidance of Femininity, NTSM = Negativity towards Sexual Minorities, SRMS = Self-Reliance through Mechanical Skills, Tough = Toughness, Dom = Dominance, IoS = Importance of Sex, RE = Restrictive Emotionality, TMI = Male Role Norms Inventory total score (i.e., TMI general factor).