Dimensionality, Reliability, and Validity of the Gender Role Conflict Scale – Short Form (GRCS-SF)

APA-Style Citation:

Hammer, J. H., & McDermott, R. C., Levant, R. F., & McKelvey, D. K. (2018). Dimensionality, reliability, and validity of the Gender Role Conflict Scale – Short Form (GRCS-SF). *Psychology of Men & Masculinity, 19*, 570-583. doi: 10.1037/men0000131g

Joseph H. Hammer, Ph.D.
243 Dickey Hall
University of Kentucky
Lexington, KY 40506
(847) 809-8785
jhhammer@gmail.com

Ryon C. McDermott, Ph.D.
UCOM 3813
University of South Alabama
Mobile, AL 36688
(251) 380-2644
rmcdermott@southalabama.edu

Ronald. F. Levant, Ed.D.
CAS 350
University of Akron
Akron, OH 44325
(330) 972-5496
levant@uakron.edu

Daniel K. McKelvey, M.A.
CAS 350
University of South Alabama
Mobile, AL 36688
(251) 460-6371
dkm1421@jagmail.southalabama.edu

Abstract

The Gender Role Conflict Scale – Short Form (GRCS-SF) was derived from the widely-used Gender Role Conflict Scale and is thought to measure men's personal and relational distress from rigid adherence to restrictive masculinities. Extant research suggests the GRCS-SF may be best modeled as a second-order structure, consisting of one second-order gender role conflict (GRC) factor explaining the associations between four lower-order GRC domain factors. The present paper revisited the factor structure of the GRCS-SF using confirmatory factor analysis in a large sample of college men (N = 1117). Contrary to previous research, a bifactor structure and a common factor structure evidenced better fit. Ancillary bifactor measures provided support for conceptualizing the theoretical construct of GRC and the dimensionality of the GRCS-SF instrument as being defined by four independent-yet-related first-order GRC factors. Furthermore, the Restrictive Emotionality and Conflicts Between Work and Family Relations factors more consistently predicted unique variance in the eight criterion variables (i.e., depression, anxiety, social anxiety, substance use, hostility, family distress, academic distress, and self-stigma of seeking psychological help) embedded in the nomological network of gender role conflict than did the Success, Power, and Competition and Restrictive Affectionate Behavior Between Men factors. Conclusions include: modeling the GRCS-SF using a correlated factors solution is recommended, use of the four raw GRC subscales scores may be permissible, and calculation of a raw GRC total score using all 16 items is contraindicated.

Keywords: gender role conflict; bifactor analysis; reliability; validity; scale development

Dimensionality, Reliability, and Validity of the Gender-Role Conflict Scale – Short Form (GRCS-SF)

For more than 30 years, investigations grounded in the gender role strain paradigm (Pleck 1981, 1995) have had a profound impact on psychological practice and research with boys and men (see Levant & Richmond, 2016, for a review). A central tenet of the masculine gender role strain paradigm is that some men experience personal and relational distress when they adhere to rigid, dysfunctional masculine role norms (i.e., dysfunction strain; Pleck, 1995). Numerous studies provide support for this theoretical assertion, as evidenced by generally consistent associations between men's adherence to traditional, sexist masculine norms and mental health problems (Wong, Ho, Wang, & Miller, 2016), physical health problems (Chambers et al., 2016; Christy, Mosher, Rawl, & Haggstrom, in press; Sloan, Conner, & Gough, 2015), and relational health problems (Burn & Ward, 2005; O'Neil, 2008).

Given the relevant threats of masculine gender role dysfunction strain to different aspects of men's wellbeing, several researchers have operationalized the construct using self-report questionnaires (c.f., Levant & Richmond, 2016). The Gender Role Conflict Scale (GRCS; O'Neil, Helms, Gable, David, & Wrightsman, 1986) has emerged as a popular and widely used measure of men's dysfunction strain in the past three decades, in part, because it captures the personal relational distress created by rigid adherence to dysfunctional male roles (O'Neil, 2008). Indeed, the GRCS has been used in at least 280 studies between 1980 and 2014, making it one of the most-used instruments within the gender role strain paradigm (O'Neil, 2015). Across the majority of these 280 investigations, published and unpublished results support positive relationships between GRCS scores and a variety of personal and relational problems in men (see O'Neil, 2008, 2015; O'Neil, Good, & Holmes, 1995 for reviews).

Although the GRCS is widely used, the instrument is 37 items long, and some studies have provided mixed support for its psychometric properties (Good et al., 1995; Rogers, Abbey-Hines, & Rando, 1997). Addressing these concerns, Wester, Vogel, O'Neil, and Danforth (2012) developed a shortened version, the Gender Role Conflict Scale Short Form (GRCS-SF). The GRCS-SF holds promise as being a shorter, more refined measure of GRC (O'Neil, 2015); however, comparatively few studies have examined its psychometric properties or internal structure. Moreover, because the GRCS-SF, like its parent instrument, has traditionally been used to calculate both a total GRC score and individual GRC subdomain subscale scores, it is important to investigate the dimensionality of the instrument to determine whether it provides reliable and valid measurements of general GRC and specific patterns of GRC, respectively. To address these concerns, the present investigation examined the dimensionality of the GRCS-SF, the model-based reliability of the GRCS-SF total and subscale scores, and incremental convergent evidence of validity for the general and specific GRCS-SF latent factors.

Measurement of Gender Role Conflict

GRC, as measured by the GRCS, pertains to the personal and relational psychological consequences of adhering to traditional male roles that severely limit a man's ability to respond adaptively to situational demands (O'Neil, 2015). The original GRCS was developed in a large sample of college men (N = 586) through content analysis and exploratory factor analysis (EFA), yielding four related but theoretically distinct factors of GRC (O'Neil et al., 1986): success power and competition (SPC), restrictive emotionality (RE), restrictive affectionate behavior between men (RABBM), and conflicts between work and family relations (CBWFR). Three of the four GRCS subscales (RE, RABBM, and CBWFR) tap the personal and relational consequences of rigid adherence to masculine roles emphasizing stoicism, heterosexism, and

primacy of work, respectively. By contrast, the SPC subdomain captures men's maladaptive drive for success through competition and pursuit of power (O'Neil, 2015).

Although the correlated factors (also referred to as a four-factor oblique, common factors, or first-order factors model) model of the GRC is widely accepted in the literature (O'Neil, 2015), confirmatory factor analysis (CFA) has provided mixed support (e.g., Good et al., 1995; Moradi, Tokar, Schaub, Jome, & Serna, 2000; Rogers et al., 1997). Indeed, some investigators have argued that certain items should be removed or refined (Good et al., 1995), and at least one study identified that the CFA indices of fit for the instrument did not meet conventional standards (Rogers et al., 1997). To address these concerns, as well as to significantly shorten the instrument, Wester and colleagues (2012) used EFA to select GRCS items with the most evident simple structure (i.e., high factor loadings and low cross loadings) in a diverse sample of 399 adult men and adolescents. From the 37 original items, only 16 were retained to create a short form of the instrument, with factor loadings ranging from .66 to .78. Wester et al. then confirmed that a correlated factors solution provided a better fit to the data than a unidimensional (i.e., one-factor) solution for the GRCS-SF in a sample of 1031 men and adolescents who had taken the original GRCS. Thus, like the original GRCS, the GRCS-SF has been used to calculate subscale and total scores. Subsequent researchers have found associations between these GRC scores and men's self-reports of risky health behaviors (Levant, Parent, McCurdy, & Bradstreet, 2015; Levant & Wimer, 2014), traditional masculinity ideology (Levant, Hall, Weigold, & McCurdy, 2015), mental health stigma (Vogel, Wester, Hammer, & Downing-Matibag, 2014), and violent attitudes toward women (McDermott, Naylor, McKelvey, & Kantra, 2017).

Dimensionality of Gender Role Conflict

Despite wide use of the GRC total and subscale scores, comparatively little research has examined the dimensionality of GRC with either the GRCS-SF or the original GRCS. Theoreticians have argued that specific GRC factors may share common variance with an overall GRC factor driven by fear of femininity and internalized personal and institutional sexism (O'Neil, 2008, 2015). Neither the GRCS-SF nor the GRCS, however, contain items explicitly measuring sexism or fear of femininity. Nevertheless, investigators have noted that GRCS subscale scores correlate moderately (.35) to strongly (.68) with each other (Moradi et al., 2000), which has been interpreted as supporting the possible existence of a general GRC factor. GRCS totals scores have also been robustly related to men's sexism and negative attitudes toward women (e.g., O'Neil, 2008), further supporting the possibility of an overall GRC factor that is congruent with GRC theory. However, the only study (to the best of our knowledge) to examine a general GRC construct with the original GRCS found that a correlated factors model evidenced similar fit compared to a second-order model (i.e., where a superordinate GRC factor indirectly influences item-level variance through the four lower-level patterns of GRC; Norwalk, Vandiver, White, & Englar-Carlson, 2011). Raising additional questions about the structure of GRC, neither the correlated factors model nor the second-order model provided acceptable fit to the data (Norwalk et al., 2011), although a correlated factor model yielded acceptable fit when individual items were artificially parceled together (e.g., Moradi et al., 2000).

To date, only three studies have examined the factor structure of the GRCS-SF. As mentioned previously, Wester et al (2012) first identified that a one-factor solution was a poor fit to the data over the correlated factors model in the original GRCS-SF validation sample. However, the authors pulled the GRCS-SF items from a sample that had completed the original GRCS, and thus it is possible that participants' responses to the retained items may have been

influenced by their responses to the omitted items (Weinberger, Darkes, Del Boca, Greenbaum, & Goldman, 2006). Next, Zhang and colleagues (2015) tested differences between a correlated factors, unidimensional, and second-order model in a sample of 256 Chinese men. They found that only the correlated factors and second-order models provided acceptable fit, and concluded that the second-order model did not evidence a better fit compared to the common factors model.

Most recently, Levant, Hall, Weigold, and McCurdy (2015) tested differences between a second-order model and a bifactor model of the GRCS-SF, thereby comparing two alternative approaches to modeling an overall GRC construct. Unlike a second order model, which implies the effects of a higher-order factor are mediated through a set of lower-order factors (Kline, 2016), a bifactor model indicates that item variation is explained by the additive contribution of a general factor (often measured by a total score) and a specific factor (measured by a subscale score; Reise, 2012). Said another way, the superordinate GRC construct can be conceptualized as a general factor that exists independently of its specific forms (i.e., bifactor) or as a higher-order factor that is defined by the common element that runs through all four specific forms (i.e., second-order). A bifactor model, therefore, allows researchers to determine the degree of multidimensionality of an instrument, because it may be the case that specific factors no longer explain the observed relationships among items over and above the explanation that is provided by the general factor or that the general factor accounts for insignificant variance and thus only the specific factors matter (Reise, 2012). In addition, because the general and specific factors are all specified as orthogonal to each other (i.e., not allowed to correlate), it becomes possible to calculate ancillary bifactor measures that help researchers answer questions such as “Is the GRCS-SF unidimensional enough to justify the modeling of a general GRC factor?” and “Are the raw GRCS-SF subscale scores reliable enough indicators of their specific factors, or must

SEM be used to model these factors?" Second-order models cannot facilitate answers to these questions (Chen, West, & Sousa, 2006).

Using a sample of 444 community and college men, Levant, Hall, Weigold, and McCurdy (2015) found that the GRCS-SF yielded acceptable fit to the data for both the second-order and the bifactor model. However, the results were somewhat equivocal as to whether the second-order model was a better fit than the less parsimonious bifactor model in which it was nested. Specifically, all indices of fit except one, the Bayseian Information Criterion (BIC), favored the bifactor model, but the chi-square difference between the two models was non-significant. The authors concluded that the more parsimonious second-order model should be retained based on the chi-square difference test. Of note, although Levant, Hall, et al., (2015) decided to retain a second-order model over the bifactor model, several researchers have pointed out that a bifactor model can still be used to more formally test the dimensionality of an instrument as it allows researchers to partition the variance between general and specific factors (Chen et al., 2006; Reise, 2012; Rodriguez, Reise, & Haviland, 2016b).

In summary, the literature provides varying degrees of empirical support for three competing models of the GRCS-SF: the correlated factors, second-order, and bifactor models. Thus, there is a lack of consensus regarding the "true" dimensionality of the GRCS-SF. This lack of consensus also engenders ambiguity about the appropriateness of using raw total and subscale scores (see Brunner, Nagy, & Wilhelm, 2012) to represent the general GRC construct and the specific patterns of GRC, respectively. The present study sought to reconcile these conflicting findings through the application of a more robust set of empirical and theoretical criteria.

The Present Study

This study had three goals. First, the present study sought to verify the dimensionality of the GRCS-SF via global CFA model fit comparisons and ancillary bifactor measures (e.g., explained common variance [ECV] and individual explained common variance [IECV]; see Rodriguez et al. 2016b for a detailed review). Second, once the dimensionality of the GRCS-SF was identified, traditional and bifactor model-based reliability estimates were consulted to determine the appropriateness of using raw total and subscale GRC scores as measures of their intended constructs. Third, both multiple linear regression and structural equation modeling (SEM) were used to examine incremental convergent evidence of validity for the GRC scores. Specifically, the present study drew from comprehensive reviews of GRC, as well as GRC theory, to identify eight relevant criterion variables that reside within the nomological network of GRC but have not been examined fully in relation to the GRCS-SF: depression, anxiety, social anxiety, substance use, hostility, family distress, academic distress, and self-stigma of seeking psychological help. Although, given the ambiguity surrounding the structure of GRC, no hypotheses were advanced related to the reliability or dimensionality of the GRCS-SF, several specific hypotheses (H1-H8) were advanced regarding each of the eight variables imbedded in the GRC nomological network.

H1: Depression. Because the psychological aspects of GRC are believed to be marked by "cognitive, affective, unconscious, or behavioral problems caused by socialized gender roles in sexist and patriarchal societies" (O'Neil, 2008, p. 365), GRC should logically be associated with men's psychological distress in a variety of domains. Most notably, numerous studies have connected GRCS total and subscale scores with greater levels of depression in men (e.g., O'Neil, 2008, 2015), suggesting depression is a key aspect of a GRC nomological network. Given that

the GRCS-SF correlated highly ($r > .90$) with the original GRCS, we hypothesized that GRCS-SF subscale and total scores would be positively associated with depression symptoms.

H2 & H3: Anxiety and social anxiety. Several researchers have identified positive connections between GRC and anxiety symptoms. In theory, GRC restricts men's ability to express emotions, thus increasing anxiety and other internal psychological distress (O'Neil, 2015). Several investigators have noted that men with higher levels of GRC tend to be more socially isolated (e.g., O'Neil, 2008) and GRC demonstrated a positive relationship with social sensitivity, a driver of social anxiety (Blashilll & Vander Wal, 2009). Thus, we hypothesized that GRCS-SF subscale and total scores would be positively associated with general anxiety (H2) and social anxiety symptoms (H3).

H4 & H5: Substance use and hostility. A critical assumption of GRC theory is that men who are restricted personally and relationally may turn to harmful externalizing behaviors, such as substance use or violence, to manage distress (Cochran & Rabinowitz, 2000). Indeed, several researchers have connected higher levels of GRC to substance abuse and aggression (e.g., O'Neil, 2008; 2015). Thus, we hypothesized that GRCS-SF subscale and total scores would be positively related to self-reported substance abuse (H4) and hostility (H5).

H6 & H7: Family distress and academic distress. Although potentially a more distal correlate of GRC, there is some evidence that men with higher levels of GRC also have problems in their primary family unit, possibly because of the familial transmission of GRC (e.g., DeFranc & Mahalik, 2002). Among college students in particular, family support is important for academic and social wellbeing (Fruiht, 2015). With respect to academic distress, there is some evidence that rigid masculine role norms may restrict college men's academic functioning

(Wimer & Levant, 2011). Accordingly, we hypothesized that GRCS-SF total and subscale scores would be positively associated with family distress (H6) and academic distress (H7).

H8: Self-stigma of seeking psychological help. Researchers have long noted connections between men's gender role strain—and GRC in particular—and negative attitudes toward seeking psychological help (see Vogel & Heath, 2016 for a review). Previous studies suggest that self-stigma of seeking psychological help is a key variable within the GRC nomological net (Pederson & Vogel, 2007). Thus, we hypothesized that GRCS-SF subscale and total scores would be positively associated with stigma of seeking help.

Method

Participants, Measures, and Procedure

Participants were 1177 male college students from a large Midwestern university. After institutional review board (IRB) approval, participants were recruited as part of larger, campus mental health needs assessment distributed by e-mail to a random, representative sample of 20,000 students. Although a true response rate could not be determined due to an inability to know if a student actually received the recruitment e-mail, approximately 18% of the target sample participated in the original needs assessment. The present sample was selected from the original sample based on gender and whether participants completed the variables of interest. Three published studies (McDermott, Cheng, Wong, Booth, Jones, & Sevig, 2017; McDermott, Cheng, Wright, Browning, Upton, & Sevig, 2015; M McDermott, Smith, Borgogna, Booth, Granato, & Sevig, 2017*)* have used participants from this original sample; the present study is the first to utilize these participants' GRCS-SF data. All participants were directed to an online survey that began with an informed consent page, followed by the instrument battery and demographic items, and ended with a debriefing page where they could enter to win one of

twenty-four $25 gift cards to the local campus bookstore. Participants ranged in age from 17 to 90 years old (M = 23.09, SD = 4.40). See Table 1 for additional demographic information. See Table 2 for internal consistency estimates (α), means, and standard deviations for all instrument scores in both subsamples.

Gender Role Conflict Scale-Short Form (GRCS-SF; Wester, et al., 2012). The GRCS-SF is shortened version of the original GRCS (O'Neil, et al., 1986). Both instruments and their four subscales (i.e., SPC, RE, RABBM, CBWFR) were described in the introduction. Items are scored on a six-point Likert scale, ranging from 1 (*strongly disagree*) to 6 (*strongly agree*). A higher subscale score indicates more gender role conflict on that dimension. The GRCS-SF scores have been found to correlate highly with the original GRCS scores (.90 < r's < .96, p's < .001; Wester et al., 2012). The GRCS-SF scores have demonstrated acceptable internal consistency estimates (.77 < α's < .80; Levant, Hall et al., 2015). At the request of university stakeholders, we substituted the term "same sex" in two RABBM items (e.g., "Hugging other men is difficult for me" was changed to "Hugging someone of the same sex is difficult for me").

Self-Stigma of Seeking Help scale (SSOSH; Vogel, Wade, & Haake, 2006) The SSOSH is a 10-item scale assessing the degree to which individuals believe that their self-image would be negatively impacted by seeking counseling. Responses are rated on a five-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). A sample item is, "I would feel inadequate if I went to a therapist for psychological help." Items are averaged, and higher scores indicate greater self-stigma of seeking help. Vogel et al. (2006) provided initial evidence for the reliability and validity of the SSOSH in a large college student sample and confirmed the unidimensional factor structure. Subsequent research has connected SSOSH scores with more negative attitudes toward counseling in men (e.g., Vogel, Heimerdinger-Edwards, Hammer, &

Hubbard, 2011). Internal consistency estimates for SSOSH scores in college student populations have ranged from .72 to .90 (Vogel et al., 2006).

College Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62; Locke et al. 2011). The CCAPS-62 consists of 62 symptoms reflecting psychological and social concerns among college students. Respondents are instructed to indicate how well each item describes them during the past two weeks on a Likert-type scale ranging from 0 (*Not at all*) to 4 (*Extremely well*). Items are summed and averaged for each subscale: depression (13 items; e.g., "I feel isolated and alone"), substance use (6 items; "I drink more than I should"), generalized anxiety (9 items; "I have spells of terror or panic"), hostility (7 items; "I have difficulty controlling my temper"), social anxiety (7 items; "I am shy around others"), family distress (6 items; "my family is basically a happy one"[reverse scored]), and academic distress (5 items; "I am unable to keep up with my schoolwork"). Each subscale evidenced reliability coefficients greater than .80 and demonstrated significant correlations in the expected directions with widely used epidemiological measures of depression, anxiety, and alcohol use (Locke et al., 2011; McAleavey et al., 2012). Test-retest reliability coefficients over a two-week period were also found to be adequate, ranging between .76 and .92 (Locke et al., 2011).

Results

Data Cleaning

The initial dataset contained 1519 individuals. Cases with 20% or more missing data on any given subscale (n = 342) were deleted, which resulted in a sample of N = 1177. Univariate and multivariate outliers comprised less than 2% of the sample in all instances, thus the outliers were retained (Meyers, Gamst, & Guarino, 2013). No variables exceeded the cutoffs of 3 and 10 for high univariate skewness and kurtosis values, respectively (Weston & Gore, 2006). In

addition, we used the MLR estimator to estimate the Model $\chi 2$ and associated fit indices that use it to protect against deviations from multivariate normality. The total sample (N = 1177) was then randomly split into Initial (n = 588) and Validation (n = 589) subsamples to facilitate examination of the stability of the results.

Dimensionality

The internal structure of the GRCS-SF was tested via a series of CFAs with M*plus* version 6.11 (Muthén & Muthén, 1998-2012). Specifically, four competing measurement models (i.e., unidimensional, correlated factors, second-order, bifactor) were examined (see Figure 1a, 1b, 1c, and 1d). M*plus*' MLR option for maximum likelihood estimation was used, which calculates the scaled chi-square test statistic (scaled χ^2). Model fit was evaluated using the scaled χ^2 statistic, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Standard Root Mean Square Residual (SRMR). The following fit criteria were used: RMSEA ≤ .06, CFI ≥ .95, TLI ≥ .95, SRMR < .08 for good fit and RMSEA ≤ .10, CFI ≥ .90, TLI ≥ .90, SRMR < .10 for acceptable fit (Weston & Gore, 2006).

The unidimensional model and second-order model were nested within both the correlated factors and bifactor models. The correlated factors model was not nested within the bifactor model because the model contained more than three latent variables. Thus, scaled chi-square difference tests ($\Delta\chi 2$), Akaike's information criterion (AIC), and Bayesian information criterion (BIC) were used to compare model fit with one exception: only the AIC and BIC could be used to compare the fit of the correlated factors model to the fit of the bifactor model. Only models that achieved at least adequate model fit were compared via these indices. Burnham and Anderson (2002) state that an AIC value difference exceeding 6 and especially 10 provides evidence of model fit difference (as cited in Symonds & Moussalis, 2011, p. 17). A BIC value

difference exceeding 10 provides strong evidence of model fit difference (Kass & Raftery, 1995). The model with the lower AIC and BIC value is considered to have superior model fit. All analyses used a 5% significance level.

Results indicated that the correlated factors, second-order, and bifactor solutions provided adequate to good model fit in both the Initial and Validation samples (see Table 3 for goodness of fit statistics for all tested models). Relative model fit comparisons revealed that the bifactor solution[1] fit significantly better than the correlated factors solution per the AIC (Initial ΔAIC = 38.17, Validation ΔAIC = 40.54) but not the BIC (Initial ΔBIC = -5.6, Validation ΔBIC = -3.24). The bifactor solution fit significantly better than the second-order solution per the AIC (Initial ΔAIC = 39.42, Validation ΔAIC = 40.52) and the chi-square (Initial Δ scaled χ^2 (12) = 50.03, p < .001; Validation Δ scaled χ^2 (12) = 33.87, p < .001), but the BIC suggested the opposite (Initial ΔBIC = -13.11, Validation ΔBIC = -12.02). In turn, the correlated factors solution did not provide a better fit to the data than did the second-order solution per the chi-square (Initial Δ scaled χ^2 (2) = 5.62, p = .06; Validation Δ scaled χ^2 (2) = 4.29, p = .12) and the AIC and BIC (Initial ΔAIC = 1.25, Validation ΔAIC = -.02; Initial ΔBIC = -7.51, Validation ΔBIC = -8.77), suggesting that the more parsimonious second-order model is preferred over the correlated factors model.

In summary, global fit indices did not provide conclusive evidence for the superiority of one model over all others. Some scholars suggest that BIC should be given more weight in bifactor modeling than AIC and chi-square because BIC imposes greater parsimony penalties, which is an issue in bifactor models, particularly in large sample studies such as this one (Murray & Johnson, 2013). When BIC is given the most weight, the second-order model tended to exhibit the strongest performance compared to the other models. However, given the lack of

robust evidence in favor of the second-order model, we proceeded to calculate ancillary bifactor measures to more accurately determine the dimensionality of an instrument (see Hammer & Toland, 2016, for a video overview). Table 4 summarizes the estimates for all ancillary bifactor measures.

First, we calculated the percent of Explained Common Variance (ECV; Reise, Moore, & Haviland, 2010), an index of unidimensionality, attributable to the general factor and each of the four specific factors. Rodriguez, Reise, and Haviland (2016a) state that "when [general] ECV is > .70 and [Percentage of Uncontaminated Correlations (PUC) is] > .70, relative bias will be slight, and the common variance can be regarded as essentially unidimensional" (p. 232). Whereas the PUC (.80) met the minimum threshold, the general ECV (Initial = .23, Validation = .16) did not, suggesting that the GRCS-SF is primarily multidimensional. The general ECVs indicated that only 16% to 23% of the common variance in the GRCS-SF items was due to the general GRC factor.

Second, we examined the standardized item factor loadings for the unidimensional and bifactor solutions, as well as the Individual Explained Common Variance (IECV) estimates for each item (see Table 5). Unlike the ECV, the IECV provided information about the “extent to which an item’s responses are accounted for by variation on the latent general factor alone...thus acts as an assessment of unidimensionality at the individual item level.” (Stucky, Thissen, & Edelen, 2013, p. 51). Consistent with the low ECVs, we noted that 88% to 100% of the 16 items loaded more strongly on their assigned specific factor than on the general GRC factor, and 88% to 94% had IECVs below .50. Both findings suggest that most GRCS-SF items are better measures of their assigned specific factors than of the general GRC factor, which further suggests the GRCS-SF is primarily multidimensional.

Third, the average difference between items' loading on the general factor in the unidimensional solution and on the general factor in the bifactor solution was .14 to .16, with loading differences ranging from a low of .01 to a high of .49 (see Table 5). Thus, the average relative measurement parameter bias (Rodriguez et al., 2016b) across items was 59% to 72%, which is well above the 10-15% upper limit discussed by Muthen, Kaplan, and Hollis (1987). In other words, forcing the GRCS-SF into a unidimensional solution would lead to a significant amount of bias in the measurement of the GRCS-SF's item factor loadings. This further reinforces that the GRCS-SF is primarily multidimensional.

In conclusion, the ancillary bifactor measures support conceptualizing the GRCS-SF as multidimensional. The weak general GRC dimension highlighted by the low general ECV contraindicates the conceptualization and modeling of an overall GRC dimension, whether it is treated as a higher-level GRC factor driving the four lower-order GRC factors (i.e., second-order model) or a general GRC factor and a series of additive, unique GRC domain factors (i.e., bifactor model). Given that the correlated factors model (a) avoids the problems associated with asserting the existence of an overall GRC dimension and (b) provided an adequate to good global fit to the data, we suggest that the theoretical construct of GRC and the dimensionality of the GRCS-SF instrument are best conceptualized as being defined by four independent-yet-related first-order GRC factors. However, given that a bifactor solution can help refine the measurement of multidimensional constructs by partitioning variance (Rodriguez et al, 2016b), we examined the model-based reliability of the bifactor solution to facilitate greater understanding of the internal structure of the GRCS-SF. Such estimates have not yet been provided in published literature on the GRCS-SF and are useful for determining whether the use of raw total and

subscale scores will yield reliable estimates of a GRC general factor and specific domains of GRC in a bifactor model context, respectively.

Bifactor Model-Based Reliability

Following the recommendations for more intensive analysis of a bifactor solution (Rodriguez et al., 2016b), we calculated different model-based reliability estimates which provide important information about the reliability of multidimensional measures not available through a Cronbach's alpha. First, Coefficient Omega (ω) measures the proportion of raw total score variance that can be attributed to all common factors (i.e., true score variance, which excludes error variance). It can also be adapted to measure the proportion of raw subscale score variance that can be attributed to all common factors (i.e., Coefficient Omega Subscale). Second, Coefficient Omega Hierarchical (ωH) measures the proportion of raw total score variance that can be attributed to the general GRC factor when simultaneously accounting for the four specific GRC factors. Third, Coefficient Omega Hierarchical Subscale (ωHS) measures the proportion of explained subscale score variance uniquely accounted for by the corresponding specific factor, after partialling out the variance accounted for by the general factor. Fourth, the Percentage of Reliable Variance (PRV) in the raw total score that is due to the general GRC factor is the ratio of ωH to ω. The PRV in each raw subscale score that is due to its intended specific factor is the ratio of ωHS to ω. The PRV is a more useful indicator of bifactor model-based reliability than the omega coefficients because it takes overall explained variance into account.

Regarding ωH and ωHS, Reise, Bonifay, and Haviland (2013) stated that "tentatively, we can propose that a minimum would be greater than .50, and values closer to .75 would be much preferred" (p.137). Li, Toland, and Usher (2016) used a minimum cutoff of .75 for the PRV. We

used these criteria to determine whether the raw total score (thought to operationalize the general GRC factor) and each of the raw subscale scores (thought to operationalize their respective specific GRC factors) appeared to be sufficiently reliable measures of their intended factors (see Table 4).

Results indicated that the general GRC factor only accounted for 35 to 43% of the explained variance (Initial ωH = .43; Validation ωH = .35) and 40 to 49% of the reliable variance (Initial PRV = .49; Initial PRV = .40) in the raw GRC total score. In other words, the raw GRC total score (i.e., score created by averaging or summing the values of the 16 GRC items) primarily measures unintended constructs, rather than the intended construct of general GRC. This constitutes a lack of support for the calculation and interpretation of the raw GRC total score as a measure of the general GRC construct. This finding is not surprising, given the aforementioned lack of dimensionality evidence in favor of modeling a general GRC factor.

Given the strong multidimensionality of the GRCS-SF described above, it is also not surprising that the specific factors demonstrated stronger bifactor model-based reliability. Specifically, the ωHS's for the SPC, RABBM, and CBWFR specific factors were always above the .50 minimum threshold and approached the .75 preferred threshold. Their PRV values always exceeded the .75 threshold. However, the RE specific factor failed to achieve the minimum thresholds in the Initial sample. This is understandable, given that the general GRC factor poached a significant amount of the RE items' variance, leaving the RE specific factor with less variance to account for (see Table 5). In summary, when the GRCS-SF is forced into a bifactor solution, the model-based reliability results suggest that raw scores for the SPC, RABBM, and CBWFR subscales can be safely interpreted as representing meaningful information about their respective specific factors, but that the raw RE subscale score may be an

insufficiently-reliable indicator of the RE specific factor. However, when the GRCS-SF is modeled using the recommended correlated factors model, the raw RE subscale score's reliability will not be compromised by the presence of a general GRC factor. In fact, the raw RE subscale's internal consistency, as appropriately measured by Cronbach's alpha in the context of a correlated factors model, was $\alpha = .82$ in the Initial sample and $\alpha = .82$ in the Validation sample.

Incremental Convergent Evidence of Validity

In line with extant research, we initially specified eight (i.e., depression, anxiety, social anxiety, substance use, hostility, family distress, academic distress, and self-stigma of seeking psychological help) separate SEM structural models in which the (a) GRCS-SF items were set to load in accordance with the aforementioned bifactor model, (b) the criterion variable's items were set to load on the latent criterion variable factor, and (c) the GRCS-SF general and specific factors were simultaneously regressed onto the latent criterion variable factor. Cohen's (1988) guidelines were used to interpret small ($\beta = .20$), medium ($\beta = .50$), and large ($\beta = .80$) effect sizes. Review of these models would help determine whether the weak general GRC factor was successfully measuring a construct that covaried in theoretically-expected ways with criterion variables in the GRC nomological network.

However, pilot testing of this approach revealed that several of the structural models failed to converge, likely due to the weak model-based reliability of the general factor (see above). Therefore, we adjusted the structural models such that, while the general and specific factors were all modeled in the measurement portion of the model, only the GRCS-SF specific factors were simultaneously regressed onto the latent criterion variable factor (i.e., only the specific factors had the chance to account for variance in the criterion variables) in the structural portion of the model. These adjusted models (see Figure 2a) allowed us to remove, from the

subdomain facets, the variance due to the general GRC factor, and examine the unique relationship of each specific pattern of GRC with theoretically-relevant constructs.

Second, given our recommendation that the GRCS-SF be modeled using a correlated factors solution, we conducted a parallel set of eight SEM structural models in which the (a) GRCS-SF items were set to load in accordance with the correlated factors model, (b) the criterion variable items were set to load on the latent criterion variable factor, and (c) the GRCS-SF correlated factors were simultaneously regressed onto the latent criterion variable factor (see Figure 2b). Comparing the results across the two latent methods allowed us to determine the degree to which the general GRC factor variance (when left un-modeled in the context of the correlated factors model) biases the strength of the relationship between the four specific factors and each of the criterion factors.

Third, we conducted a parallel set of eight multiple linear regressions in SPSS (Version 23) in which the four raw GRCS-SF subscale scores were regressed onto the raw score for each criterion variable. Comparing these results with the SEM results allowed us to determine the degree to which using raw scores (i.e., failing to account for measurement error via SEM) biases the strength of the relationship between the four GRC variables and each of the criterion variables. This is important to examine because extant studies using the GRCS-SF often utilized raw subscale scores.

Table 6 summarizes the standardized beta results from these three parallel sets of analyses (i.e., bifactor structural model, correlated factors structural model, multiple linear regression with raw scores). There are three patterns worth noting. First, the RE and CBWFR factors more consistently predicted unique variance in the criterion variables than did the SPC and RABBM factors. Second, the corresponding structural path standardized beta values for the

bifactor and correlated factors models were generally consistent with each other. This indicates that there is negligible structural parameter bias introduced by using the more parsimonious correlated factors model (which was recommended by the global fit indices and ancillary bifactor measures reported above) that purposely avoids specifying the existence of a general GRC factor. This provides further support for our recommendation to conceptualize the construct of GRC, and the GRCS-SF instrument, as being composed of four correlated first-order factors.

Third, the standardized beta values for the multiple linear regression was generally slightly lower (i.e., approximately .04 lower) than the corresponding values from the bifactor and correlated factors models. This indicates that using the raw GRC subscale scores instead of latent GRC factor scores engenders a small degree of downward bias in the estimation of the relationship between the four GRC variables and the criterion variables. In other words, as one might expect, the use or raw GRC subscale scores slightly underestimates the true relationship between the GRC factors and other theoretically-relevant factors, but this underestimation is not alarmingly large, thanks to the sufficient reliability of the four raw subscale scores.

Discussion

The present study investigated (a) the dimensionality of the GRCS-SF, (b) the bifactor model-based reliability of the GRCS-SF total and subscale scores, and (c) incremental convergent evidence of validity for the GRCS-SF scores using two subsamples of college students.

Dimensionality

Global fit indices did not provide conclusive evidence for the superiority of one model over all others, though it is clear a unidimensional model provides a poor fit to the GRCS-SF data. However, the ancillary bifactor measures supported conceptualizing gender role conflict,

as operationalized by the GRCS-SF, as a multidimensional construct. Specifically, evidence of a weak general GRC dimension indicated that a bifactor or second-order model are less appropriate than a correlated factors model. We therefore argue that, if future research on other populations continues to reveal this same pattern of results found in both the Initial and Validation samples, the theoretical construct of GRC and the dimensionality of the GRCS-SF instrument are best conceptualized as being defined by four independent-yet-related first-order GRC factors. Our findings are thus consistent with recent theoretical formulations of GRC (e.g., O'Neil, 2015), but they depart from the theory-based assertion evident in most diagrams and discussion of GRC as being driven by a higher-order GRC factor. Such a distinction became more apparent when examining the bifactor model-based reliability estimates.

Bifactor Model-Based Reliability

Even though a bifactor solution was found to be sub-optimal for the GRCS-SF, a bifactor model can be useful for deeper investigations into the multidimensionality of an instrument as an exercise in variance partitioning (Rodriguez et al. 2016). Thus, we chose to report model-based reliability estimates to investigate the impact of the possible general GRC factor on the model-based reliability of the raw GRCS-SF scores. Results indicated that the raw GRC total score primarily measured the four specific factors rather than the intended construct of general GRC. This constitutes a lack of support for the calculation and interpretation of the raw GRC total score as a measure of a general GRC construct, which was foreshadowed by the aforementioned weak general GRC dimension.

In the context of a bifactor solution, the model-based reliability results suggested that raw scores for the SPC, RABBM, and CBWFR subscales can be safely interpreted as representing meaningful information about their respective specific factors, and provided mixed support for

the use of the raw RE subscale score. However, given our argument that a correlated factors solution is the most appropriate way to model the GRCS-SF, traditional internal consistency estimates take precedence over bifactor model-based reliability estimates in the present case. Internal consistency estimates indicated that all four raw subscale scores are sufficiently reliable measures (α's > .78), which aligns with extant literature (e.g., O'Neil, 2008, 2015).

Incremental Convergent Evidence of Validity

We investigated incremental convergent evidence of validity for the GRC scores using three complementary approaches (i.e., bifactor structural model, correlated factors structural model, multiple linear regression with raw scores). We hypothesized that GRCS-SF total and subscale scores would be positively associated with each of the eight criterion variables (i.e., depression, anxiety, social anxiety, substance use, hostility, family distress, academic distress, and self-stigma of seeking psychological help) based on findings from within the broader GRC nomological network. However, our results only partially supported these hypotheses. In support of our hypotheses and consistent with previous published and unpublished studies of GRC (e.g., O'Neil, 2008), results indicated that the RE and CBWFR factors more consistently predicted unique variance in the eight criterion variables. Our results provided incremental convergent evidence of validity for the GRCS-SF raw and latent subscale score for RE and CBWFR, indicating that men's difficulties expressing vulnerable emotions and developing a healthy work-life balance may be particularly relevant to their personal and relational distress.

Contrary to our hypotheses, SPC and RABBM factors were often weak or inconsistent unique predictors of our criterion variables. Such findings point to the variability within the GRC literature with respect to findings related to SPC and RABBM (e.g., O'Neil, 2008), and thus they do not necessarily reflect poor psychometric properties of the GRCS-SF. To this point, it is

important to emphasize that the present analyses were focused on determining which of the GRCS-SF factors accounted for unique (i.e., incremental) variance. When SPC and RABBM did not have to compete with RE and CBWFR to account for variance in the criterion variables (i.e., when examined with a series of simple bivariate correlations; not shown), SPC and RABBM more consistently demonstrated statistically significant correlations with the criterion variables. However, researchers and clinicians must think critically about the clinical significance of SPC and RABBM, given their frequent difficulty in accounting for unique variance beyond the RE and CBWFR factors. Specifically, in what contexts is it important to assess or intervene with SPC and RABBM, and in what contexts do RE and CBWFR seem to obviate SPC and RABBM's influence on men's functioning? Additional research, particularly meta-analyses, will be needed to determine the overall effects of certain GRC domains on specific outcomes. In the meantime, it should be noted that prior research supports RE and CBWFR's greater predictive strength of clinical problems. In a narrative compilation of GRC findings (O'Neil, 2008), RE was reported to be the most consistent predictor of both depression and alexithymia, and RE and CBWFR were most strongly correlated with lack of psychological well-being and shame. Disconnection from one's feelings and work-life balance difficulties may have exert a more potent and chronic form of distress than does concern with success and one's intimate interactions with other men, which may be less salient in daily living for many men.

In addition to providing incremental convergent validity evidence for the GRCS-SF, our results offered further support for modeling the GRCS-SF as four interrelated factors. Specifically, through comparison of the results for the bifactor and correlated factors structural models, we determined that whether a general GRC factor was modeled, this did not substantially influence the strength of the relationship between the four GRC factors and each of

the criterion factors. This provides further support for the use of the more parsimonious correlated factors model. Lastly, through comparison of the results for the correlated factors model and the multiple linear regression, we determined that the use of raw subscale scores to estimate the relationship between the four patterns of GRC and criterion variables introduced a small degree of downward bias. Researchers and clinicians who wish to use raw GRCS-SF subscale scores in lieu of SEM factors (which can account for measurement error) should therefore carefully consider how much downward bias they are willing to tolerate, given their intended use of the scores.

Cautions, Limitations, and Future Directions

As of this writing, scholarship on bifactor modeling is rapidly evolving (Rodriguez et al., 2016). Guidelines and standards for conducting bifactor analysis and interpreting results are in flux, so we offer the present interpretations with humility and the understanding that our conclusions are subject to future clarification. For example, we were cautious about making strong conclusions related to the potential superiority of the bifactor model, given the overfitting bias that the bifactor model enjoys as the least parsimonious model (Bonifay, Lane, & Reese, 2017; Murray et al., 2013). Given the undergraduate student makeup of this sample, it cannot be assumed that these findings will automatically generalize to other populations. Most sample participants were also heterosexual, White, educated, and reported experiencing rare to occasional financial distress. Thus, the generalizability of these findings should be tested directly in future research, rather than assumed. For example, perhaps the GRCS-SF demonstrates a different factor structure with a much stronger general GRC factor among community adults, or within populations scoring very low or very high in gender role conflict. Likewise, future research is necessary to determine whether these findings generalize to the

original 37-item GRCS or to the various adaptation and translations of the GRCS that have been published (e.g., O'Neil, 2015). We also remind researchers that, because GRC is a social construction influenced by situation and context, it is important to discuss the significance of GRCS-SF scores in a manner that does not essentialize masculinity (Mahalik, 2014).

Conclusion and Recommendations

In summary, if the present findings demonstrate stability across populations in future studies, we recommend the following. First, users should not calculate or interpret or raw GRC total score using all 16 items. Second, it is permissible to attempt to model the general GRC factor in the context of a bifactor solution, but the weakness of the general GRC factor may render such a model less useful. Third, we recommend conceptualizing and modeling the GRCS-SF using a correlated factors solution in which four first-order factors represent the four patterns of GRC. Fourth, the use of the four raw GRC subscale scores may be permissible, provided users are aware and comfortable with the small downward bias that may occur when using those raw scores.

Footnote

[1] It is important to consider that a bifactor model forces the general factor to first account for as much shared variance across all items as possible, prior to letting the specific factors account for as much remaining variance as they can in their respective sets of items (Rodriguez et al., 2016). Thus, even when an instrument is truly best represented by a correlated factors solution, a bifactor solution will still typically be able to extract enough shared variance across the items to form a general factor. This general factor shared variance can represent the general construct of interest, a general theoretical construct that was not intended to be measured, or even a general method effect factor. What the general factor truly represents is a matter of logical debate, informed by both empirical analysis and rational-theoretical argument.

References

Blashill, A. J., & Vander Wal, J. S. (2009). Mediation of gender role conflict and eating pathology in gay men. *Psychology of Men and Masculinity, 10*, 204-217.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796-846. doi: 10.1111/j.1467-6494.2011.00749.x

Burn, S. M., & Ward, A. Z. (2005). Men's conformity to traditional masculinity and relationship satisfaction. *Psychology of Men & Masculinity*, *6*(4), 254–263. https://doi.org/10.1037/1524-9220.6.4.254

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference, 2nd ed*. New York, NY: Springer.

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychology Science, 5,* 184-186.

Chambers, S. K., Hyde, M. K., Oliffe, J. L., Zajdlewicz, L., Lowe, A., Wootten, A. C., & Dunn, J. (2016). Measuring masculinity in the context of chronic disease. *Psychology of Men & Masculinity*, *17*(3), 228–242. https://doi.org/10.1037/men0000018

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

Christy, S. M., Mosher, C. E., Rawl, S. M., & Haggstrom, D. A. (in press). Masculinity beliefs and colorectal cancer screening in male veterans. *Psychology of Men & Masculinity*, https://doi.org/10.1037/men0000056

Cochran, S. V., & Rabinowitz, F. E. (2000). *Men and depression: Clinical and empirical perspectives.* San Diego, CA, US: Academic Press.

DeFranc, W., & Mahalik, J. R. (2002). Masculine gender role conflict and stress in relation to parental attachment and separation. *Psychology of Men & Masculinity, 3*(1), 51–60. https://doi.org/10.1037/1524-9220.3.1.51

Fruiht, V. M. (2015). Supportive others in the lives of college students and their relevance to hope. *Journal of College Student Retention: Research, Theory and Practice*, *17*(1), 64–87. https://doi.org/10.1177/1521025115571104

Good, G. E., Robertson, J. M., O'Neil, J. M., Fitzgerald, L. F., Stevens, M., DeBord, K. A., … Braverman, D. G. (1995). Male gender role conflict: Psychometric issues and relations to psychological distress. *Journal of Counseling Psychology*, *42*(1), 3–10. https://doi.org/10.1037/0022-0167.42.1.3

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

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

Kass, R. E., & Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. doi: 10.2307/2291091

Levant, R. F., Hall, R. J., Weigold, I. K., & McCurdy, E. R. (2015). Construct distinctiveness and variance composition of multi-dimensional instruments: Three short-form masculinity measures. *Journal of Counseling Psychology*, *62*(3), 488–502. https://doi.org/10.1037/cou0000092

Levant, R. F., Parent, M. C., McCurdy, E. R., & Bradstreet, T. C. (2015). Moderated mediation of the relationships between masculinity ideology, outcome expectations, and energy drink use. *Health Psychology*, *34*(11), 1100–1106. https://doi.org/10.1037/hea0000214

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

Levant, R. F., & Wimer, D. J. (2014). Masculinity constructs as protective buffers and risk factors for men's health. *American Journal of Men's Health, 8*(2), 110-120. https://doi.org/10.1177/1557988313494408

Li, C. R., Toland, M. D., Usher, E. L. (2016). *Dimensionality, scoring, and interpretation of the Short Grit Scale*. Manuscript in preparation.

Locke, B. D., Buzolitz, J., Lei, P., Boswell, J. F., McAleavey, A. A., Sevig, T. D., & ... Hayes, J. A. (2011). Development of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58*(1), 97-109. doi:10.1037/a0021282

Mahalik, J. R. (2014). Both/and, not either/or: A call for methodological pluralism in research on masculinity. *Psychology of Men and Masculinity, 15*, 365-368.

McAleavey, A. A., Nordberg, S. S., Hayes, J. A., Castonguay, L. G., Locke, B. D., & Lockard, A. J. (2012). Clinical validity of the Counseling Center Assessment of Psychological Symptoms-62 (CCAPS-62): Further evaluation and clinical applications. *Journal of Counseling Psychology*, *59*(4), 575-590. doi:10.1037/a0029855

McDermott, R. C., Cheng, H.-L., Wong, J., Booth, N., Jones, Z., & Sevig, T. (2017). Hope for help-seeking: A positive psychology perspective of psychological help-seeking intentions. *The Counseling Psychologist*, 45, 237-265. doi: https://doi.org/10.1177/0011000017693398

McDermott, R. C., Cheng, H.-L., Wright, C., Browning, B. R., Upton, A. W., & Sevig, T. D. (2015). Adult attachment dimensions and college student distress: The mediating role of hope. The Counseling Psychologist, 43(6), 822–852. doi: https://doi.org/10.1177/0011000015575394

McDermott, R. C., Naylor, P. D., McKelvey, D., & Kantra, L. (2017). College men's and women's masculine gender role strain and dating violence acceptance attitudes: Testing sex as a moderator. *Psychology of Men & Masculinity*, *18*(2), https://doi.org/10.1037/men0000044

McDermott, R. C., Smith, P. N., Borgogna, N., Booth, N. Granato, S. & Sevig, T. D. (2017). College students' conformity to masculine role norms and help-seeking intentions for suicidal thoughts. *Psychology of Men and Masculinity,* Advance online publication. doi: http://dx.doi.org/10.1037/men0000107

Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation*. SAGE Publications.

Moradi, B., Tokar, D. M., Schaub, M., Jome, L. M., & Serna, G. S. (2000). Revisiting the structural validity of the Gender Role Conflict Scale. *Psychology of Men & Masculinity*, *1*(1), 62–69. https://doi.org/10.1037/1524-9220.1.1.62

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. I*ntelligence, 41*, 407–422.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*(3), 431-462. doi:10.1007/BF02294365

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (7th ed)*. Los Angeles, CA: Authors.

Norwalk, K. E., Vandiver, B. J., White, A. M., & Englar-Carlson, M. (2011). Factor structure of the gender role conflict scale in African American and European American men. *Psychology of Men & Masculinity*, *12*(2), 128-143. http://dx.doi.org/10.1037/a0022799

O'Neil, J. M. (2008). Summarizing 25 years of research on men's gender role conflict using the Gender Role Conflict Scale: New research paradigms and clinical implications. *The Counseling Psychologist*, *36*(3), 358–445. https://doi.org/10.1177/0011000008317057

O'Neil, J. M. (2015). *Men's gender role conflict: Psychological costs, consequences, and an agenda for change*. Washington, DC, US: American Psychological Association.

O'Neil, J. M., Good, G. E., & Holmes, S. (1995). Fifteen years of theory and research on men's gender role conflict: New paradigms for empirical research. In R. F. Levant, W. S. Pollack, R. F. Levant (Ed), & W. S. Pollack (Ed) (Eds.), *A new psychology of men.* (pp. 164–206). New York, NY, US: Basic Books.

O'Neil, J. M., Helms, B. J., Gable, R. K., David, L., & Wrightsman, L. S. (1986). Gender-Role Conflict Scale: College men's fear of femininity. *Sex Roles, 14*(5–6), 335–350. https://doi.org/10.1007/BF00287583

Pederson, E. L., & Vogel, D. L. (2007). Male gender role conflict and willingness to seek counseling: Testing a mediation model on college-aged men. *Journal of Counseling Psychology, 54*(4), 373-384. http://dx.doi.org/10.1037/0022-0167.54.4.373

Pleck, J. H. (1981). *The myth of masculinity*. MIT Press (MA).

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

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. http://dx.doi.org/10.1080/00223891.2012.725437

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

Reise, S. P., Moore, T. N., & 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. http://dx.doi.org/10.1080/00223891.2010.496477

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98,* 223-237.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*, 137-150.

Rogers, J. R., Abbey-Hines, J., & Rando, R. A. (1997). Confirmatory factor analysis of the Gender Role Conflict Scale: A cross-validation of Good et al., 1995. *Measurement and Evaluation in Counseling and Development, 30*(3), 137–145.

Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement, 37*(1), 41-57. https://doi.org/10.1177/0146621612462759

Sloan, C., Conner, M., & Gough, B. (2015). How does masculinity impact on health? A quantitative study of masculinity and health behavior in a sample of UK men and women. *Psychology of Men & Masculinity*, *16*(2), 206–217. https://doi.org/10.1037/a0037261

Symonds, M. R. E., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference, and model averaging in behavioral ecology using the Akaike's information criterion. *Behavioral Ecology and Sociobiology, 65*(1), 13-21. doi: 10.1007/s00265-010-1037-6

Vogel, D. L., & Heath, P. J. (2016). Men, masculinities, and help-seeking patterns. In Y. J. Wong, S. R. Wester, Y. J. Wong (Ed), & S. R. Wester (Ed) (Eds.), *APA handbook of men and masculinities.* (pp. 685–707). Washington, DC, US: American Psychological Association.

Vogel, D. L., Heimerdinger-Edwards, S. R., Hammer, J. H., & Hubbard, A. (2011). “Boys don’t cry”: Examination of the links between endorsement of masculine norms, self-stigma, and help-seeking attitudes for men from diverse backgrounds. *Journal of Counseling Psychology, 58*(3), 368–382. https://doi.org/10.1037/a0023688

Vogel, D. L., Wade, N. G., & Haake, S. (2006). Measuring the self-stigma associated with

seeking psychological help. *Journal of Counseling Psychology, 53*(3), 325-337. doi:10.1037/0022-0167.53.3.325

Vogel, D. L., Wester, S. R., Hammer, J. H., & Downing-Matibag, T. M. (2014). Referring men to seek help: The influence of gender role conflict and stigma. *Psychology of Men & Masculinity*, *15*(1), 60–67. https://doi.org/10.1037/a0031761

Weinberger, A. H., Darkes, J., Del Boca, F. K., Greenbaum, P. E., & Goldman, M. S. (2006). Items as context: Effects of item order and ambiguity on factor structure. *Basic and Applied Social Psychology*, *28*(1), 17–26. https://doi.org/10.1207/s15324834basp2801_2

Wester, S. R., Vogel, D. L., O'Neil, J. M., & Danforth, L. (2012). Development and evaluation of the Gender Role Conflict Scale Short Form (GRCS-SF). *Psychology of Men & Masculinity*, *13*(2), 199–210. https://doi.org/10.1037/a0025550

Weston, R., & Gore, P. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist, 34*(5), 719-751. https://doi.org/10.1177/0011000006286345

Wimer, D. J., & Levant, R. F. (2011). The relation of masculinity and help-seeking style with the academic help-seeking behavior of college men. *The Journal of Men's Studies*, *19*(3), 256–274. https://doi.org/10.3149/jms.1903.256

Wong, Y. J., Ho, M.-H. R., Wang, S.-Y., & Miller, I. S. K. (2016). Meta-analyses of the relationship between conformity to masculine norms and mental health-related outcomes. *Journal of Counseling Psychology*. https://doi.org/10.1037/cou0000176

Zhang, C., Blashill, A. J., Wester, S. R., O'Neil, J. M., Vogel, D. L., Wei, J., & Zhang, J. (2015). Factor structure of the Gender Role Conflict Scale-Short Form in Chinese heterosexual and gay samples. *Psychology of Men & Masculinity*, *16*(2), 229–233. https://doi.org/10.1037/a0036154

Table 1

Demographic Characteristics of Initial and Validation Samples

	Initial (***n*** **= 588)**	**Validation** (***n*** **= 589)**
Race		
White or Caucasian	66.3%	68.8%
Black/African American	3.1%	3.1%
Hispanic/Latino(a)	2.9%	1.9%
Asian/Asian-American	22.1%	20.2%
Multiracial	2.9%	3.9%
Missing/Did Not Respond	2.7%	2.2%
Class Standing		
Freshman/1st year	17.9%	14.8%
Sophomore/2nd year	13.4%	11.4%
Junior/3rd year	15.1%	17.3%
Senior/4th year+	12.8%	13.6%
Graduate	36.4%	39.4%
Professional	4.4%	2.9%
Missing/Did Not Respond	0.0%	.7%
Sexual Orientation		
Heterosexual	89.6%	88.1%
Gay	5.1%	7.1%
Bisexual	2.9%	2.4%
Questioning	1.2%	1.4%
Missing/Did Not Respond	1.2%	1.0%
Current Financial Stress		
Always Stressful	8.7%	6.1%
Often Stressful	17.0%	17.1%
Sometimes Stressful	32.0%	33.4%
Rarely Stressful	31.3%	32.3%
Never Stressful	10.9%	11.0%
Missing/Did Not Respond	.2%	0.0%

Table 2

Score Means, Standard Deviations, and Internal Consistency Estimates (α)

	α (95% CI)	*M*	*SD*
Initial Sample			
GRCS-SF total score	.80 (.78, .83)	3.38	.75
RE subscale score	.82 (.80, .85)	3.29	1.26
SPC subscale score	.81 (.78, .83)	4.00	1.14
RABBM subscale score	.82 (.80, .84)	2.44	1.14
CBWFR subscale scores	.86 (.84, .88)	3.77	1.32
Validation Sample			
GRCS-SF total score	.78 (.75, .80)	3.37	.70
RE subscale score	.82 (.79, .84)	3.31	1.21
SPC subscale score	.83 (.81, .85)	4.03	1.14
RABBM subscale score	.80 (.78, .83)	2.32	1.09
CBWFR subscale scores	.85 (.83, .87)	3.83	1.25

Note: GRCS-SF = Gender Role Conflict Scale - Short Form, RE = Restrictive Emotionality, SPC = Success, Power, Competition, RABBM = Restrictive Affectionate Behavior Between Men, CBWFR = Conflict Between Work and Family Relations.

Table 3

Goodness of Fit Statistics for All Tested Measurement Models

Model	Scaled $\chi 2$	*df*	RMSEA [90% CI]	CFI	TLI	SRMR	AIC	BIC
Unidimensional (Initial Subsample)	3092.63	104	.221 [.214, .228]	.005	-.148	.165	32927.61	33137.69
Unidimensional (Validation Subsample)	2293.41	104	.189 (.182, .196)	.253	.138	.177	32508.16	32718.32
Correlated Factors (Initial Subsample)	239.49	98	.050 (.042, .058)	.953	.942	.044	30478.69	30715.04
Correlated Factors (Validation Subsample)	233.37	98	.048 (.040, .056)	.954	.943	.041	30041.72	30278.16
Second Order (Initial Subsample)	244.95	100	.050 (.042, .058)	.952	.942	.047	30479.94	30707.53
Second Order (Validation Subsample)	237.76	100	.048 (.040, .056)	.953	.944	.043	30041.71	30269.39
Bifactor (Initial Subsample)	193.00	88	.045 (.036, .054)	.965	.952	.047	30440.53	30720.64
Bifactor (Validation Subsample)	199.66	88	.046 (.038, .055)	.962	.948	.037	30001.18	30281.40

Note: All models were statistically significant at the $p < .001$ level. GRCS-SF = Gender Role Conflict Scale - Short Form. Statistics are based on MLR estimation. Scaled $\chi 2$ = scaled chi-square test statistic, RMSEA = Root Mean Square Error of Approximation, CI = Confidence Interval, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, SRMR = Standard Root Mean Square Residual.

Table 4

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

	Gen	RE	SPC	RABBM	CBWFR
Explained Common Variance	.23 (.16)	.41 (.83)	.91 (.75)	.84 (.83)	.96 (.94)
Omega / Omega Subscale	.89 (.88)	.87 (.83)	.81 (.85)	.82 (.81)	.87 (.85)
Omega Hierarchical / Omega Hierarchical Subscale	.43 (.35)	.31 (.79)	.74 (.66)	.69 (.67)	.83 (.81)
Percentage of Reliable Variance	.49 (.40)	.35 (.84)	.91 (.77)	.84 (.83)	.96 (.95)
H Index	.78 (.64)	.66 (.80)	.81 (.81)	.77 (.75)	.87 (.88)
Factor Determinacy	.84 (.72)	.81 (.88)	.90 (.89)	.87 (.85)	.93 (.94)

Note: GRCS-SF = Gender Role Conflict Scale - Short Form, Gen = General Gender Role Conflict Factor, RE = Restrictive Emotionality Specific Factor, SPC = Success, Power, Competition Specific Factor, RABBM = Restrictive Affectionate Behavior Between Men Specific Factor, CBWFR = Conflict Between Work and Family Relations Specific Factor. Initial sample results are displayed first and validation sample results are displayed second in parentheses.

Table 5

Confirmatory Factor Analysis Standardized Loadings for the GRCS-SF

			Bifactor					
	Uni	RPB	IECV	Gen	RE	SPC	RABBM	CBWFR
Restrictive Emotionality Items								
g1 Talking (about my feelings) during sexual relations is difficult for me.	**.54 (.22)**	10% (20%)	.44 (.07)	**.49 (.18)**	**.55 (.69)**			
g2 I have difficulty expressing my emotional needs to my partner.	.55 **(.23)**	11% (15%)	.31 (.11)	**.50 (.27)**	**.74 (.79)**			
g3 I have difficulty expressing my tender feelings	**.61 (.27)**	10% (28%)	.68 (.20)	**.68 (.38)**	**.46 (.75)**			
g4 I do not like to show my emotions to other people	**.55 (.27)**	28% (25%)	1.00 (.36)	**.76 (.36)**	.04 **(.47)**			
Success, Power, and Competition Items								
g5 Winning is a measure of my value and personal worth	**.40 (.70)**	56% (19%)	.17 (.61)	**.26 (.59)**		**.58 (.47)**		
g6 I strive to be more successful than others	.33 **(.72)**	100% (104%)	.06 (.23)	**.16** (.35)		**.67 (.64)**		
g7 Being smarter or physically stronger than other men is important to me	**.43 (.80)**	67% (159%)	.09 (.12)	**.26 (.31)**		**.82 (.84)**		
g8 I like to feel superior to other people	**.33 (.65)**	74% (270%)	.08 (.06)	**.19 (.18)**		**.67 (.69)**		
Restrictive Affectionate Behavior Between Men Items								
g9 Affection with other men makes me tense.	**.48 (.29)**	42% (8%)	.22 (.14)	**.34 (.27)**			**.64 (.67)**	

g10 Men who touch other men make me uncomfortable	**.45 (.28)**	82% (2%)	.11 (.15)	**.24 (.29)**		**.71 (.68)**	
g11 Hugging someone of the same sex is difficult for me.	**.42 (.25)**	61% (23%)	.13 (.22)	**.26 (.32)**		**.66 (.60)**	
g12 Being very personal with people of the same sex makes me feel uncomfortable	**.49 (.24)**	48% (20%)	.19 (.18)	**.33 (.31)**		**.68 (.65)**	
Conflict Between Work and Family Relations Items							
g13 Finding time to relax is difficult for me	**.37 (.22)**	116% (26%)	.06 (.22)	**.17 (.30)**			**.67 (.57)**
g14 My needs to work or study keep me from my family or leisure more than I would like.	.34 **(.21)**	183% (27%)	.02 (.04)	.12 (.16)			**.84 (.82)**
g15 My work or school often disrupts other parts of my life (home, health, leisure, etc).	.37 **(.20)**	139% (178%)	.03 (.01)	**.15** (.07)			**.85 (.90)**
g16 Overwork and stress, caused by a need to achieve on the job or in school, affects/hurts my life	.39 **(.20)**	122% (23%)	.06 (.05)	**.17 (.16)**			**.71 (.68)**

Note: GRCS-SF = Gender Role Conflict Scale - Short Form. Uni = Unidimensional Model, Gen = General Gender Role Conflict Factor, RE = Restrictive Emotionality Specific Factor, SPC = Success, Power, Competition Specific Factor, RABBM = Restrictive Affectionate Behavior Between Men Specific Factor, CBWFR = Conflict Between Work and Family Relations Specific Factor, IECV = Individual Explained Common Variance, RPB = Relative Parameter Bias. Loadings are standardized and based on MLR estimation. All bolded loadings significant at $p < .05$. Initial sample results are displayed first and validation sample results are displayed second in parentheses.

Table 6

Standardized Betas Quantifying the Relationship between the GRCS-SF Scores and Criterion Variables

	Academic distress	Anxiety	Depression	Family distress	Hostility	Social anxiety	Stigma	Substance use
RE (Bifactor)	**.24 (.23)**	**.24 (.26)**	**.30 (.33)**	**.17 (.11)**	.13 **(.18)**	**.41 (.53)**	**.21 (.26)**	.10 (.07)
RE (Correlated Factors)	**.24 (.19)**	**.21 (.26)**	**.29 (.33)**	**.17 (.11)**	**.11 (.17)**	**.40 (.41)**	**.11 (.25)**	**.13 (.07)**
RE (Raw)	**.18 (.17)**	**.19 (.22)**	**.26 (.29)**	**.14 (.09)**	**.12 (.16)**	**.33 (.35)**	**.14 (.24)**	**.11** (.05)
SPC (Bifactor)	-.06 (.06)	.05 (.03)	.03 (.04)	-.08 (-.06)	.13 **(.18)**	-.03 **(.15)**	**.27 (.28)**	**.13 (.11)**
SPC (Correlated Factors)	-.09 (-.02)	.01 (.00)	-.01 (.01)	-.08 (-.06)	**.12 (.11)**	-.08 (.02)	**.23 (.26)**	**.14 (.12)**
SPC (Raw)	-.06 (-.02)	.02 (.00)	.00 (.01)	-.06 (-.04)	**.13 (.13)**	-.05 (.01)	**.20 (.27)**	**.14 (.11)**
RABBM (Bifactor)	.03 (.09)	.05 (.00)	**.19** (.03)	-.01 (.00)	**.14 (.13)**	**.27 (.26)**	**.38** (.08)	-.07 **(-.12)**
RABBM (Correlated Factors)	-.01 (.02)	-.01 (-.02)	.05 (.02)	-.01 (-.00)	**.12 (.12)**	**.16 (.15)**	**.33** (.05)	-.09 **(.14)**
RABBMD (Raw)	.01 (.02)	-.01 (-.01)	.05 (.04)	-.02 (.01)	.08 (.08)	**.14 (.12)**	**.27** (.04)	-.08 **(-.12)**
CBWFR (Bifactor)	**.33 (.41)**	**.27 (.28)**	**.26 (.62)**	**.14 (.16)**	**.18 (.20)**	.08 **(.15)**	.02 (.05)	.02 (.00)
CBWFR (Correlated Factors)	**.34 (.39)**	**.27 (.28)**	**.25 (.28)**	**.14 (.16)**	**.17 (.20)**	.06 (.10)	-.02 (.04)	.01 (-.01)
CBWFR (Raw)	**.29 (.37)**	**.26 (.31)**	**.25 (.31)**	**.13 (.17)**	**.18 (.21)**	**.08 (.12)**	.00 (.05)	.01 (.00)

Note: GRCS-SF = Gender Role Conflict Scale - Short Form, RE = Restrictive Emotionality specific factor, SPC = Success, Power, Competition specific factor, RABBM = Restrictive Affectionate Behavior Between Men specific factor CBWFR = Conflict Between Work and Family Relations specific factor, Bifactor = Structural equation modeling-based bifactor model, Correlated Factors = Structural equation modeling-based correlated factors model, Raw = Multiple linear regression using raw scores. Standardized betas are based on MLR estimation. All bolded standardized betas were significant at $p < .05$. Initial sample results are displayed first and validation sample results are displayed second in parentheses.

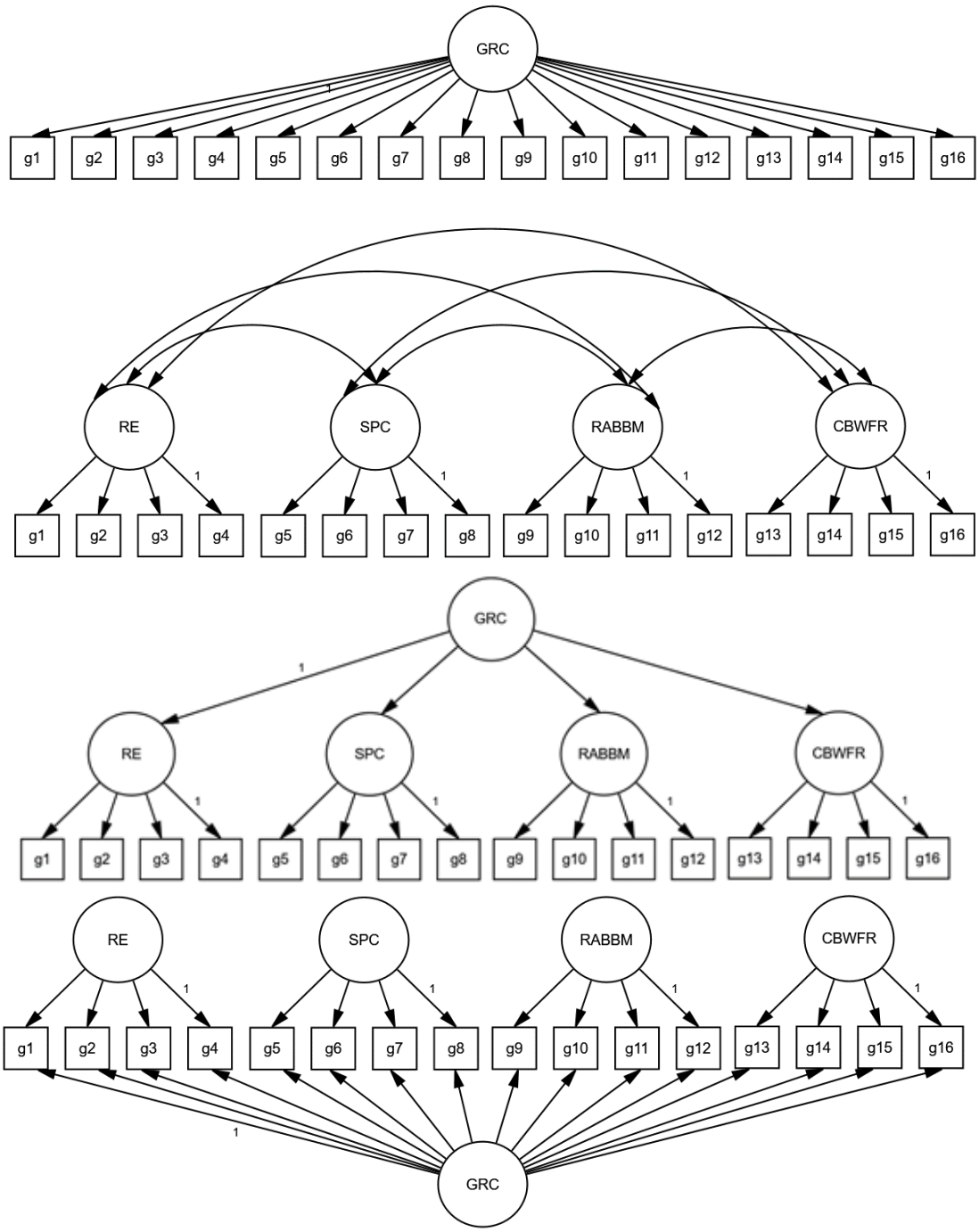

Figure 1a, 1b, 1c, 1d. Different ways of modeling the GRCS-SF. Models pictured from top to bottom: Unidimensional (1a), Correlated Factors (1b), Second-order (1c), and Bifactor (1d). GRC = Gender Role Conflict, RE = Restrictive Emotionality, SPC = Success Power and Competition, RABBM = Restrictive Affectionate Behavior Between Men, and CBWFR =

Conflicts between Work and Family Relations. Item error terms and latent variable disturbance terms are not displayed.

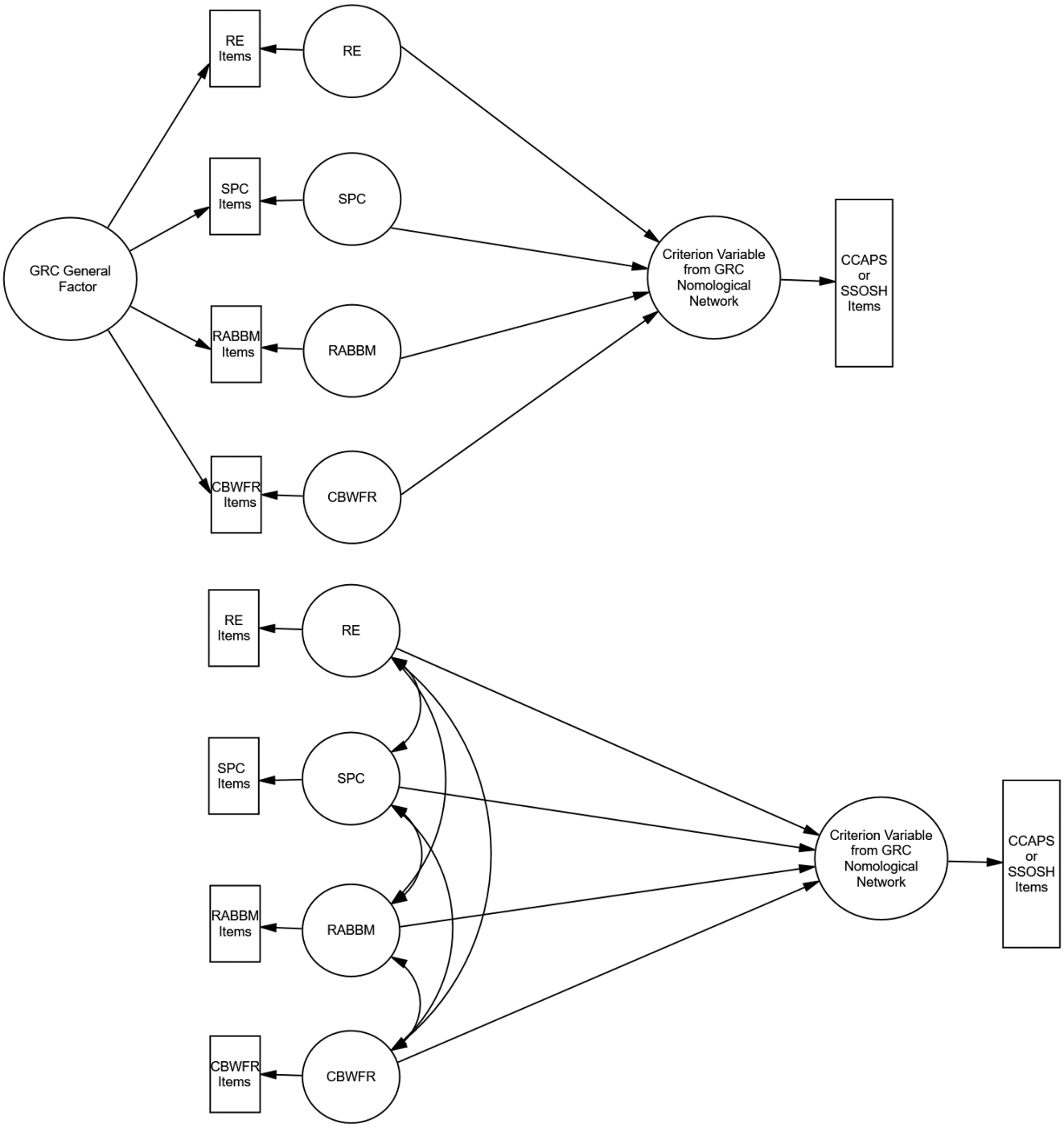

Figure 2a and 2b. Structural equation models used to examine incremental convergent evidence of validity. Models pictured from top to bottom: Adjusted bifactor structural model (2a) and Correlated Factors structural model (2b). GRC = Gender Role Conflict, RE = Restrictive

Emotionality, SPC = Success Power and Competition, RABBM = Restrictive Affectionate Behavior Between Men, CBWFR = Conflicts between Work and Family Relations, CCAPS = College Counseling Center Assessment of Psychological Symptoms, and SSOSH = Self-Stigma of Seeking Help. Individual items and their error terms, and latent variable disturbance terms are not displayed.