Sex Roles, Vol. 14, Nos. 7/8, 1986

Masculinity, Femininity, and Androgyny: A Methodological and Theoretical Critique[1]

Herbert W. Marsh[2] and Margaret Myers
The University of Sydney, Australia

The three primary purposes of this investigation are: (1) to apply confirmatory factor analysis (CFA) to responses from three instruments [Bem Sex Role Inventory (BSRI), Comrey Personality Scale (CPS), and Australian Sex Role Scale (ASRS) developed by Antill and his colleagues] designed to measure masculinity (M) and femininity (F); (2) to determine the correlation between M and F on each instrument and to relate the results to the design of the instrument; and (3) to describe a new theoretical model that posits global M and global F to be multifaceted, higher order constructs. When CFA was used to define one M and one F factor from responses to each instrument, the disattenuated correlations between M and F were +0.58 (BSRI), −.50 (ASRS), and −1.0 (CPS). Thus, responses from two instruments supported the separation of M and F, but differed in the direction of the correlation for the two traits, while the third provided remarkably strong support for a bipolar MF continuum. Despite the apparent inconsistency, the observed correlations were explicable in terms of the design of each instrument. While the two-factor model provided a reasonable fit for the BSRI, more complicated models were better able to fit responses to the ASRS and CPS. Results from this investigation and other research were used to formulate a new theoretical model.

Virtually all researchers prior to 1973 and many personality inventories still in use today assume masculinity (M) and femininity (F) to be the end points

[1]The authors wish to thank Douglas Farnill, Ian Ball, and Andres Comrey for providing us with data from previous research used in this study. The authors would also like to thank Samuel Ball, Jenifer Barnes, and Raymond Debus for their helpful comments on drafts of this manuscript.

[2]Requests for reprints should be sent to Herbert W. Marsh, Reader, Department of Educational Psychology, University of Sydney, Sydney, NSW2006, Australia.

0360-0025/86/0400-0397$05.00/0

of a single bipolar dimension. Inherent in this position is the assumption that M and F are correlated close to −1.0; more masculinity (femininity) implies less femininity (masculinity). This widely held belief was challenged in Constantinople's (1973) classic review of Masculinity–Femininity (MF) research. She found convincing evidence that MF is multidimensional, and suggested that the apparent bipolarity in the construct may be a function of the selection and/or construction of items. The social zeitgeist of the women's movement and Constantinople's challenge of the bipolarity assumption in MF research combined to spawn the construct of androgyny, and led to a tremendous resurgence of MF research during the past decade. While debate continues among androgyny researchers about the definition and measurement of the androgyny construct, there appears to be a consensus among these researchers that M and F are distinguishable traits (Cook, 1985), and there appears to have been little protest from traditional personality theorists who previously viewed MF as a bipolar construct. Nevertheless, Constantinople's criticism of MF research was more complicated than has been incorporated into the theory and measurement of androgyny. In particular, she warned that the factor structure underlying M and F may be complex, and that artifacts in the selection/construction of items to measure MF may seriously affect the empirical findings.

Masculinity/Femininity: How Many Constructs?

A Single Bipolar Construct. Traditionally, personality researchers have hypothesized M and F to be the end points of a bipolar dimension, and this is how the construct is represented in many personality inventories such as the California Psychological Inventory (CPI; Gough, 1969) and Comrey Personality Scales (CPS; Comrey, 1970) discussed below. As recently as 1973, Constantinople indicated that "no measure of M-F has been devised that does not incorporate bipolarity from the start" (1973, p. 392).

Megargee (1972) in a summary of the development of the MF scale on the CPI suggested that the scale was originally designed to "differentiate men from women and sexual deviates from normals" (p. 90). Further evolvement of the interpretations based on the scale emphasized a *psychological* continuum rather than one based on gender, and the focus on sexual psychopathology was deemphasized. Megargee concluded that the scale "reflects psychological femininity and not simply sex differences" (p. 93). However, the scale was derived by taking a large item pool and selecting the items most highly correlated with gender, and responses to the scale correlate .64 to .78 with gender. The items in this scale are not designated to be specifically M or specifically F, but the item selection procedure is unlikely to result in items that reflect primarily the characteristics of one gender or the other,

and so a bipolar construct consistent with the assumption of the instrument may be reasonable (but see Cunningham and Antill, 1980). However, Constantinople (1973) suggested that the variety of item clusters included in this scale mean that it is probably multidimensional, though no factor analyses had been performed. Research and development of the MF scale on the CPI is typical of many personality inventories (see Constantinople, 1973, for further discussion).

In an alternative approach, Comrey (1970) developed distinct item clusters that reflect different components of MF on a logical/theoretical basis. He used factor-analytic techniques to revise the scale, and subsequently to demonstrate that scores representing each cluster contributed to a more general MF factor. Each item cluster is labeled to represent the masculine end of the continuum (e.g., no romantic love, tolerance of vulgarity) and each contains two items that define the M end of the bipolar continuum and two that define the F end. However, in the actual factor analyses, responses to the F items and the M items in each cluster are summed to form a single score that represents the cluster. Consequently, such analyses are incapable of identifying separate M and F scales, though it is possible to test the bipolarity assumption in an analysis of responses to individual items rather than item-cluster scores. Nevertheless, the construction of items is such that M and F items are logically opposed (e.g., "It would be hard to make me cry" is a masculine item, while "I am easily moved to tears" is a feminine item) so that a high negative correlation consistent with the assumption of a bipolar scale is likely. Comrey (1970) reported that the MF scale on the CPS correlated about .60 with gender.

In both the CPI and the CPS, as well as many other personality inventories, the MF scale is hypothesized to represent a bipolar psychological construct substantially correlated with gender. Whereas the CPI uses males and females as criterion groups to select items, the CPS defines separate item clusters on a logical/theoretical basis and demonstrates that these combine to form a MF scale. The selection and design of items for both instruments is likely to result in bipolar scales consistent with their theoretical basis, and a test of this assumption for the CPS is one purpose of the present investigation.

Two Distinguishable Constructs: Androgyny. More recently, Bem (1974), Constantinople (1973), Heilbrun (1976), Spence and Helmreich (1979a,b), and others questioned the assumption that M and F represent a bipolar continuum. They argue instead that it is logically possible for a person of either gender to be both masculine and feminine, and the existence of both in the same person has been labeled androgyny. The key assumptions of Bem's 1974 theoretical description of androgyny are that M and F are orthogonal dimensions, and that individuals high on both are mentally healthier and socially more effective.

The most widely used instruments to infer androgyny are the Bem Sex Role Inventory (BSRI) and the Personal Attributes Questionnaire (PAQ). The BSRI and PAQ were constructed according to somewhat different rationales, and their authors also make theoretical distinctions such as differences in the generality of the M and F constructs as inferred by the two instruments (see Cook, 1985). Nevertheless, both the BSRI and the PAQ make inferences about M and F on the basis of socially desirable characteristics, both result in distinguishable M and F scales, and PAQ scores are highly correlated with BSRI scores. Lamke (1982) reported BSRI/PAQ correlations of 0.78 and 0.86 for M and F scales, respectively, and in a comparison of the revised versions of each instrument, Lubinski, Tellegen, and Butcher (1983) concluded that "the short BSRI and the EPAQ were found to be empirically interchangeable" (p. 428). Thus, while the empirical bases and theoretical rationales for the BSRI and the PAQ differ somewhat, the two instruments apparently measure similar constructs.

Androgyny researchers disagree on precisely how androgyny should be defined and measured (see Cook, 1985), but they do agree that M and F reflect two distinguishable traits and not a bipolar construct. Hence, the most frequently tested postulate of androgyny theory is that the correlation between M and F scales must differ significantly from -1.00 in a practical as well as a statistical sense. Bem (1974) argued that the two components are *uncorrelated,* and research with both the BSRI and PAQ has shown the M and F scales to be somewhat *positively* correlated (e.g., Cunningham & Antill, 1980; Lee & Scheurer, 1983, Lubinski et al., 1983; Nicholson & Antill, 1981; also see Spence, Helmreich, & Holahan, 1979).

A second assumption in androgyny theory is that "the combination of masculine and feminine characteristics is deemed to have desirable implications for an individual's behavior regardless of sex" (Cook, 1985, p. 21). Hence, it must be demonstrated that both M and F (or their interaction) contribute uniquely to the prediction of appropriate criterion measures. The most frequently studied criterion for the second assumption has been the predicted positive relationship between androgyny and measures of self-esteem or social well-being. While measures of androgyny that reflect both high F and high M scores are positively correlated with esteem-related measures, most of the predictable variance can be accounted for by the M score alone (e.g., Antill & Cunningham, 1979; 1980; Lamke, 1982; Silvern & Ryan, 1979; Taylor & Hall, 1982; Whitely, 1983) so that the androgynous sex role status is more advantageous to females than males (Heilbrun, 1984). Other research has shown that F scores may contribute positively and uniquely to some other criteria that are nurturant, affiliative, or empathetic in nature (e.g., Bem, 1975; 1977; Cook, 1985; Lee & Scheurer, 1983; Taylor & Hall, 1982). Nevertheless,

support for the unique positive contribution of F to the prediction of esteem that plays a central role in androgyny theory is weak.[2]

Positive and Negative Attributes of M and F. The BSRI and the PAQ primarily consider only socially desirable attributes, and this may constitute an important weakness. For example, the correlation between M and F may be masked by a method effect in responses to the socially desirable items (Baumrind, 1983; Kelly, Caudill, Hathorn, & O'Brien, 1977; Kelly & Worrell, 1977; Pedhauzer & Tetenbaum, 1979). According to such a method-effect hypothesis, responses to two sets of socially desirable items will be positively correlated in a way that is independent of the "true" MF correlation. The operation of such a method effect is also likely to affect correlations between M and F scores and self-esteem measures, since self-esteem is typically inferred by the endorsement of positively valued items and the nonendorsement of negatively valued items. Spence, Helmreich, and Holahan (1979), basing their arguments on intuitive and theoretical perspectives, also contended that many M and F characteristics are socially undesirable, but may still have important consequences. Similarly, on the basis of their review of empirical and theoretical research, Kelly and Worrell argued that "these findings support our position that negative attributes may be a functional part of some or all sex role orientations" (1977, p. 1107).

In response to this potential weakness, Spence et al. (1979) expanded PAQ (EPAQ) to include comparable M and F scales defined by socially *undesirable* characteristics, and Antill, Cunningham, Russell and Thompson (1981) developed the Australian Sex Role Scale (ASRS) to specifically measure M and F with positively valued characteristics (MP and FP) and with negatively valued characteristics (MN and FN). Consistent with the method-effect proposal, both groups found that the correlation between M and F was most positive for MP and FP scales, and negative when based upon correlations of MP and FN, and MN and FP scales. Spence et al. (1979) also demonstrated that the correlations between the EPAQ scales and self-esteem, though reasonably consistent across sexes, varied dramatically with the scale; correlations were high-positive, low-positive, near-zero, and low-negative for

[2]In subsequent research (Marsh, in press-a) relating responses to the ASRS with multidimensional facets of self-concept it was shown that (a) both M and F contributed uniquely to the prediction of self-concept facets, (b) the relative contribution of M as opposed to F varied substantially with the specific facet of self-concept, and (c) the relative contribution of F tended to be more positive than M for those areas of self-concept in which females have higher self-concepts than do males. In this subsequent study, boys scored significantly higher than girls on MP and MN, while girls scored higher than boys on FP and FN, but none of the correlations was greater than .3.

MP, FP, MN, and FN scales, respectively. A similar pattern was also observed with the ASRS scales (Russell & Antill, 1984). This pattern of correlations suggests that the endorsement of positive items, and the nonendorsement of negative items, on MF scales contributes to the prediction of self-esteem, independent of whether an item represents M or F. This is also consistent with the method-effect proposal.

The Multidimensional Factor Structure of Androgyny Instruments

Researchers tend to treat M and F scales as if these scales measure either one unidimensional bipolar scale, or two distinguishable unidimensional scales representing M and F. The recent extension of the MF scales designed to infer androgyny to include socially undesirable characteristics further complicates this issue. Furthermore, Spence and Helmreich (1979a,b; 1981; Spence, 1983, 1984) argued that none of the scales in PAQ, EPAQ, or BSRI measure global self-images of M or F. Instead, they argue that the F scales measure primarily expressive and communal traits, while the M scales measure instrumental traits. Similarly, the original version of the BSRI contained the items "masculine" and "feminine," and researchers (e.g., Feather, 1978; Gaudreau, 1977; Pedhauzer & Tetenbaum, 1979) have found that these two items form a separate scale that is clearly bipolar and distinguishable from characteristics measured by the other items. Exploratory factor analyses of responses to various androgyny instruments typically result in a complicated pattern of content-specific factors, and only some of these can be unambiguously identified as masculine or feminine (Antill & Russell, 1980; Feather, 1978; Hong, Kavanagh, & Tippett, 1983; Myers, 1982; Pedhauzer & Tetenbaum, 1979; also see Cook, 1985; Myers & Gonda, 1982). These findings offer further support for the Spence–Helmreich contention that global F and global M are each multidimensional constructs that cannot be adequately described as single, unidimensional factors, and they echo the earlier conclusion expressed by Constantinople (1973).

Different approaches have been employed to deal with the apparent multidimensionality of M and F. Bem (1979) claimed that "culture has arbitrarily clustered together heterogeneous collections of attributes into two categories prescribed as more desirable for one sex or the other" (p. 1049), and that the purpose of the BSRI is to determine how individuals self-endorse these clusters. Her position is thus consistent with the multidimensionality of global M and global F, but she preferred to use a conglomerate of items to reflect this multidimensionality rather than to hypothesize and to measure separate components of the global constructs. Other M and F measures have also used atheoretical, empirical procedures for differentiating between M

and F that may be consistent with the Bem perspective (e.g., the CPI and ASRS). Spence intentionally limited consideration to specific components of M and F, and argued that M and F as measured by PAQ are not global measures. Comrey (1970) has taken yet another perspective in defining five specific traits that define M and F on a logical/theoretical basis, and in demonstrating that these combine to form a more general MF scale. In a summary of research on the dimensionality of MF, Spence (1984, p. 25) concluded that empirical findings "disconfirm not only the unifactorial model on which conventional masculinity/femininity tests were predicated but also the more recently proposed two-factor models" but that "alternative conceptualizations of any breadth, based on the insight that masculine and feminine phenomena are multifactorial, have yet to be devised and empirically tested."

The Use of Confirmatory Factor Analysis

Factor-analytic studies in MF research have typically used exploratory rather than confirmatory factor analysis (CFA). In exploratory factor analysis the researcher is unable to define a particular factor structure beyond determining the number of factors to be rotated, and perhaps the degree of obliqueness in the rotated factors. Since the exploratory factor-analysis model is not unique, alternative solutions may be mathematically equivalent in terms of their ability to explain the data but may lead to quite different substantive interpretations. When the observed factor solution does not closely resemble the hypothesized structure, there is no way of determining the extent to which the hypothesized structure would fit the data. In CFA the researcher defines the specific factor structure to be tested, and is able to test its ability to fit the data in an absolute, statistical sense and also in comparison with alternative models (see Joreskog, 1980; Joreskog & Sorbom, 1981; Long, 1983; Marsh, 1985, in press-b; Marsh & Hocevar, 1983, 1984, 1985). Consequently, CFA is a much stronger analytic tool for examining the factor structure underlying a set of measured variables, and the advantages of this procedure are particularly important to the examination of issues in MF research.

A primary purpose of this investigation is to employ CFA in the examination of factor structures designed to explain responses to three instruments that employ different approaches to the measurement of M and F; the BSRI, the ASRS, and the CPS. Alternative models describing a single MF dimension, and separate M and F factors, are examined for each of the three instruments. Models positing separate factors based on positively and negatively valued characteristics (for the ASRS), and distinguishable facets of M and F (for the CPS) are also examined. In addition, the correlation of

scales based on the best solution with criterion measures from two of the studies is explored.

STUDY 1: M AND F WITH ASRS

Method

Sample. Study 1 is a reanalysis of data described by Farnill and Ball (1982a,b; also see Farnill & Ball, 1985), and a more detailed description is presented by those authors (we are indebted to Douglas Farnill and Ian Ball for providing us with the data). Subjects were 158 undergraduates (79% female) enrolled in a teacher education program in Australia who ranged in age between 17 and 35.

Instruments. As part of the study, all students completed Form A of the ASRS (Antill et al., 1981) and the Janis–Field self-esteem instrument (see Crandall, 1973, for a description and a review of this instrument). The ASRS consists of 50 personalitylike characteristics (e.g., logical, anxious, loves children) and subjects respond to each on a 1 (*never or almost never true*) to 7 *(always or almost always true)* scale. The items are classified as M (20 items), F (20 items), or neutral (10 items) with half the items within each group being positively valued (ie., socially desirable) and half being negatively valued. The Janis–Field scale was originally designed to measure feelings of inadequacy and contains 20 items related to social self-esteem (Crandall, 1973), half of which are negatively worded. Crandall also reports reliability estimates in the 0.80s, and moderate convergence with other esteem measures.

Statistical Analysis. All analyses presented here are based on a 52 × 52 correlation matrix representing the 50 ASRS items, gender (male = 1, female = 2), and self-esteem. In the first set of analyses, CFAs were performed with the LISREL V program (Joreskog & Sorbom, 1981) on responses to the 40 ASRS items representing M and F. With LISREL V the researcher is able to define alternative factor models designed to test different hypotheses, and to compare the ability of competing models to fit the original data. The LISREL V program, after testing for identification, attempts to minimize a maximum likelihood function based on differences between the original and the reproduced correlation matrices, and provides an overall chi-square goodness-of-fit test.

The evaluation of how well a hypothesized structure is able to fit observed data represents an important unresolved issue in the application of CFA. In contrast to traditional significance testing, the researcher often seeks a nonsignificant chi-square that indicates that the hypothesized model fits the

data. Since this is like trying to prove a null hypothesis of no differences between predicted and obtained values, the observed chi-square is typically statistically significant and alternative indications of goodness of fit are normally employed. The most commonly used is the ratio of the chi-square to the degrees of freedom. However, perhaps as a consequence of this indicator's dependence on sample size, researchers have disagreed as to an acceptable ratio, with some arguing for ratios as low as 2, and others for ratios as high as 5 as indicative of a good fit. Other indices that may be less related to sample size have also been developed. Joreskog and Sorbom (1981) described two such measures: the root mean square residual (RMSR) that is based on differences between the original and reproduced correlation matrices, and the goodness-of-fit index (GFI) that is "a measure of the relative amount of variances and covariances jointly accounted for by the model" (p. I.41). The Tucker–Lewis index (TLI; see Bentler & Bonett, 1980) scales the observed chi-square along a 0-to-1 scale where 0 represents a null fit—normally one where the reproduced correlation matrix is diagonal—and 1.0 represents an ideal fit (actually the TLI can be slightly greater than 1.0). Marsh 1985, in press-b; Marsh & Hocevar, 1985; also see Fronell, 1983) also argued for the examination of parameter estimates in the hypothesized structure and for the comparison of the goodness-of-fit indicators for the hypothesized model with those from a variety of alternative models. Each of these alternative indications of fit is employed in examining the alternative models.

In the first analysis, a four-factor solution consistent with the design of the model was hypothesized, consisting of MP, MN, FP, and FN factors. In subsequent analyses, goodness-of-fit indicators for various three-factor, two-factor, and one-factor solutions were compared with the four-factor solution. In the second stage of the analyses, six ASRS scale scores were determined by summing responses to the six groups of items, including the neutral/positive (NP) and neutral/negative (NN) items. Correlations among the scales and coefficient alpha estimates of the reliability of each scale (Hull & Nie, 1981) were determined, and the scale scores were correlated with the Janis–Field total score and with gender.

Results and Discussion

CFA. In CFA, alternative models are specified by fixing or constraining elements in three matrices conceptually similar to matrices resulting from exploratory factor analysis. In the present investigation these are:

1. LAMBDA Y, a matrix of factor loadings;
2. PSI, a factor correlation matrix that represents the relationships among the factors; and

Table I. CFA of the ASRS: The Four-Factor Solution[a]

Items and scales	Factor loading (LAMBDA) MP	MN	FP	FN	Error/uniqueness (THETA)
MP1	.31[b]	0	0	0	.90[b]
MP2	.70[b]	0	0	0	.51[b]
MP3	.33[b]	0	0	0	.89[b]
MP4	.13	0	0	0	.98[b]
MP5	.39[b]	0	0	0	.85[b]
MP6	.25[b]	0	0	0	.94[b]
MP7	.40[b]	0	0	0	.84[b]
MP8	.41[b]	0	0	0	.83[b]
MP9	.62[b]	0	0	0	.62[b]
MP10	.32[b]	0	0	0	.90[b]
MN1	0	.41[b]	0	0	.83[b]
MN2	0	.44[b]	0	0	.81[b]
MN3	0	.68[b]	0	0	.54[b]
MN4	0	.55[b]	0	0	.70[b]
MN5	0	.37[b]	0	0	.86[b]
MN6	0	.40[b]	0	0	.84[b]
MN7	0	.57[b]	0	0	.67[b]
MN8	0	.64[b]	0	0	.59[b]
MN9	0	.41[b]	0	0	.83[b]
MN10	0	.63[b]	0	0	.61[b]
FP1	0	0	.47[b]	0	.78[b]
FP2	0	0	.33[b]	0	.89[b]
FP3	0	0	.71[b]	0	.50[b]
FP4	0	0	.65[b]	0	.58[b]
FP5	0	0	.72[b]	0	.49[b]
FP6	0	0	.47[b]	0	.89[b]
FP7	0	0	.33[b]	0	.89[b]
FT8	0	0	.49[b]	0	.76[b]
FP9	0	0	.41[b]	0	.84[b]
FP10	0	0	.60[b]	0	.64[b]
FN1	0	0	0	.33[b]	.89[b]
FN2	0	0	0	.19[b]	.96[b]
FN3	0	0	0	.69[b]	.53[b]
FN4	0	0	0	.78[b]	.39[b]
FN5	0	0	0	.28[b]	.92[b]
FN6	0	0	0	.54[b]	.71[b]
FN7	0	0	0	.70[b]	.51[b]
FN8	0	0	0	.80[b]	.37[b]
FN9	0	0	0	.56[b]	.69[b]
FN10	0	0	0	.58[b]	.67[b]

	Factor correlations (PSI)			
Scales	MP	MN	FP	FN
MP	1			
MN	.87[b]	1		
FP	.04	−.31[b]	1	
FN	−.77[b]	−.35[b]	.06	1

[a]N = 158. Parameters with values of 0 and 1 were predetermined (i.e., fixed) so that no tests of statistical significance are possible. See Table II (Model 1) for goodness-of-fit indicators.
[b]$p < .05$.

3. THETA EPSILON, a diagonal matrix of error/uniqueness terms conceptually similar to one minus the communality estimates in exploratory factor analysis.

The results of the four-factor model (see Table I) illustrate the pattern of parameters to be estimated in these three matrices. All coefficients with a value of "0" or "1" are fixed (i.e., predetermined) and not estimated as part of the analysis, while other parameters are free and are estimated by the LISREL program. For this model 40 measured variables—the ASRS items—are used to define four factors corresponding to MP, MN, FP, and FN factors. The free parameters consist of 40 factor loadings in LAMBDA Y, the 6 correlations among the four factors in PSI, and the 40 error/uniquenesses in THETA. This pattern is very restrictive in that it allows each variable to load on one and only one factor, and represents an ideal of simple structure.

The parameter estimates (Table I) for the four-factor solution, Model 1, indicate that the four factors are well defined in that the items designed to define each scale all load in the same direction and all but one of the loadings is statistically significant. The goodness-of-fit indices (see Model 1 in Table II) indicate that the model provides a reasonable description of the data; the chi-square/*df* ratio is less than 2 and the other indicators are also reasonable. Inspection of the factor correlations in PSI (Table I) is particularly important for this study. The MP factor is highly correlated with MN (0.87) and FN (−0.77), while the FP scale is less correlated with the other three factors. This suggests the possibility of a total M scale that incorporates MP and MN, or even a bipolar MF scale that incorporates the MP, MN, and FN factors. These hypotheses are tested with alternative models.

Model 2 proposes an a posteriori, three-factor solution in which the factors are M (comprised of MP and MN items), FN, and FP. While Model 2 does a reasonable job of explaining the data, its fit to the data is significantly poorer than that of Model 1. The difference in chi-square values (58) relative to the difference in *df* (3) is large whether judged in terms of statistical significance or subjective indicators of goodness of fit, and so this model is rejected.

A variety of different two-factor solutions are tested in Models 3–6. The a priori models 3 and 4 hypothesize global M and global F factors (Model 3), or positive and negative item factors (Model 4). In Model 3 the correlation between the M and F factors is substantial and negative (−0.50, with a standard error of 0.07), but is sufficiently different from −1.0 so that these components cannot be justifiably collapsed into a single bipolar scale. In Model 4 the positive and negative item factors are so highly correlated (.97) that the two factors could be collapsed and this model does no better than

Table II. Summaries of Alternative Models' Fit to the ASRS Data[a]

	χ^2	*df*	χ^2/df ratio	TLI	GFI	RMSR
Four-factor solution						
Model 1: MP, MN, FP, and FN (see Table I)	1450	734	1.98	.67	.61	.10
Three-factor solution						
Model 2: MP/MN, FP, and FN	1508	737	2.05	.66	.57	.11
Two-factor solutions						
Model 3: MP/MN and FP/FN	1775	739	2.40	.60	.55	.12
Model 4: MP/FP and MN/FN	1971	739	2.67	.55	.52	.13
Model 5: MP/FN and MN/FP	1734	739	2.35	.61	.55	.12
Model 6: MP/MN/FN and FP	1706	739	2.31	.61	.56	.12
One-factor solution						
Model 7: MP/MN/FP/FN	1972	740	2.66	.55	.52	.13
Null solutions						
Model 8: 40 uncorrelated factors	4661	780	5.98	.00	.40	.19

[a]TLI, Tucker-Lewis indicator; GFI, goodness-of-fit indicator; RMSR, root mean square residual. Factors defined in the various solutions are comprised of combinations of Masculine (M), Feminine (F), Positive (P), and Negative (N) items. Thus, in the two-factor solution "MP/MN and FP/FN", the first factor is defined by Masculine/Positive and Masculine/Negative items, while the second is defined by Feminine/Positive and Feminine/Negative items.

the one-factor model (Model 7) discussed below. Additional a posteriori, two-factor models, Models 5 and 6, were prompted by inspection of the correlations among the factors in Model 1. In the two-factor model that fits best (Model 6), a bipolar MF factor is defined by the MP, MN, and FN items, while the FP items define a separate factor. Nevertheless, the goodness of fit for each of these two-factor models is substantially poorer than that of Model 1 or even of Model 2, and so each of them is also rejected. Model 7 proposes a single MF factor, but it also does substantially poorer than Model 1 and is also rejected. (Model 8, the null model, proposes 40 uncorrelated factors corresponding to each of the measured variables, and such a model is used to define the lower bound (i.e., the zero value) for the TLI.)

In summary, these analyses indicate that the four-factor solution consistent with the design of the ASRS best describes responses to the M and F items. The inability of models positing bipolar traits to fit the data provides support for the androgyny construct. However, consistent with a bipolar hypothesis, at least the *direction* of the correlation between M and F factors is negative (Model 3). Consistent with the rationale for the ASRS instrument,

these findings also demonstrate that responses to positive and negative masculine items, and particularly to positive and negative feminine items, cannot be subsummed to form total M and F scales.

Relationship to Self-Esteem and Gender. Correlations among the six ASRS scores, including the scales comprising responses to MP, MF, FP, FN Neutral/Positive (NP), and Neutral/Negative (NN) items, self-esteem, and gender appear in Table III. Since the coefficient alphas for the ASRS scales vary substantially, particularly for the neutral scales based on only half as many items, correlations corrected for attenuation appear above the main diagonal (the disattenuated correlations are also conceptually more similar to correlations in PSI in Table I). Self-esteem, based on attenuated or disattenuated correlations, is substantially correlated with the MP (positively) and FN (negatively) scales. The overall pattern of correlations suggests that self-endorsing masculine and positive items is positively correlated with self-esteem, while self-endorsing negative items, and perhaps feminine items, is negatively correlated with self-esteem. Consistent with this suggestion, a stepwise multiple regression indicated that the MP and FN each contributed significantly to the prediction of self-esteem, but none of the other scales did so. These findings are also generally consistent with those obtained by Spence et al. (1979) with the EPAQ, and by Russell and Antill (1984) in another study based on the ASRS (but see footnote 2).

The correlations between gender and the six ASRS scores are surprisingly small. MN is significantly correlated with gender, and the direction of this correlation is in the expected direction. However, NN is the only other scale significantly correlated with gender (but see footnote 2). This general lack of correlation between the ASRS scales and gender strongly supports the Spence contention that the psychological constructs that researchers label as M and F must be clearly distinguished from gender, and perhaps that different labels should be used to describe the psychological constructs. However, the relatively small number of males included in this study and the unknown representativeness of the sample dictate that these correlations with gender be interpreted cautiously.

STUDY 2: M AND F USING A MODIFIED BSRI

Methods

The Sample and the Data. Data for Study 2 come from a study designed to explore the relationship between androgyny and occupational choices for adolescent girls (see Myers, 1982, for more detail). Subjects were Year-8 (n = 146) and Year-10 (n = 123) adolescent girls from two single-sex high

Table III. Correlations Between ASRS Scales, Self-Concept, and Gender[a]

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1 Masculine/Positive	(67)	82	−04	−71	64	−02	78	−12
2 Masculine/Negative	59[b]	(78)	36	−28	15	55	36	−36
3 Feminine/Positive	−03	−28[b]	(77)	19	47	−01	−03	−03
4 Feminine/Negative	−52[b]	−22[b]	15[b]	(81)	44	60	−79	−10
5 Neutral/Positive	42[b]	11	33[b]	−32[b]	(65)	−29	55	03
6 Neutral/Negative	−01	37[b]	−01	41[b]	−18[b]	(58)	−34	−33
7 Self-esteem	52[b]	28[b]	−02	−64[b]	36[b]	−20[b]	--	06
8 Gender (1 = male, 2 = female)	−08	−28[b]	−02	−08	−02	−19[b]	06	—

[a]Coefficients above the diagonal are corrected for attenuation. Coefficients are presented without decimal points. Values in parentheses are coefficient alpha estimates of reliability. The alphas for the two neutral scales for the ASRS are lower because they are based on only five items, whereas the other four scales are based on ten items. The average of interitem correlations for the neutral scales (0.24) is approximately the same as for the other four scales (0.25). All correlations were corrected for unreliability in the ASRS scales (but not the other variables since alpha estimates were not available) and these disattenuated correlations appear above the main diagonal.

[b]$p < .05$.

schools in a predominantly middle-class region of metropolitan Sydney, Australia. Materials were administered near the end of the Australian school year, and this is particularly relevant for Year-10 students, since this has been the traditional "school-leaving" age for students in this state. Most students complete schooling through year 10, at which time a school certificate is awarded. Until recently, throughout the state less than one-third of Year-10 students returned to complete Year 11 and 12, and many of these tended to be more academically oriented and to have aspirations for higher education. The finding that 64% of the Year-10 students in this study intended to return to school for Year 11 probably reflects the recent trend of high youth unemployment.

BSRI. The original BSRI was modified for this study because certain items in the original version were found to be beyond the vocabulary range for this age group. It was also modified to fulfill demands of the NSW State Department of Education, which had to approve the materials before they were administered. The original BSRI consisted of 20 MP items, 20 FP items, 10 NP items, and 10 NN items, but these were represented by 14, 14, 8, and 6 items, respectively, in the present study (the item numbers in Table IV identify the original BSRI items that were used). The 7-point response scale used on the BSRI was also shortened to five categories (*never or almost never true, rarely true, sometimes true, often true, always or almost always true*).

Job Aspiration/Expectation Scales. Subjects were presented with a list of jobs and were told that this covered a wide range of jobs "that might be available to you at the end of Year 10." Subjects were asked to indicate jobs that were the "type of job they preferred" (job aspiration) and the "type of job they would settle for" (job expectation). After completing all the materials, the girls then classified each job as traditional "those that girls usually went in for or the kind parents expect girls to take") or unusual (jobs that "not many people would expect girls to consider"). Jobs judged to be traditional by at least 50% of the subjects were classified as "traditional." The proportion of nontraditional jobs that each girl preferred and the proportion that each girl would settle for were taken to be measures of the nontraditionality of their job aspirations and expectations. Preferred jobs tended to be somewhat less traditional than jobs subjects would settle for (means = 28 vs 22%), though responses on the two variables were moderately correlated ($r = .54$).

Statistical Analysis. In the first set of analyses, CFA models were defined to explain responses to the 14 MP items and the 14 FP items from the BSRI (no MN nor FN items appear on the BSRI). A two-factor model, consisting of M and F factors, was tested across all subjects. Though not reported here, CFA models in which responses by Year-8 and by Year-10 students were

Table IV. CFA of Bem Scale: Two-Factor Solution[a]

Items and scales	Factor loadings (LAMBDA) MP	FP	Error/uniqueness (THETA)
MP1 (4)	.50[b]	0	.75[b]
MP2 (7)	.43[b]	0	.81[b]
MP3 (10)	.14[b]	0	.98[b]
MP4 (16)	.60[b]	0	.64[b]
MP5 (25)	.58[b]	0	.67[b]
MP6 (43)	.47[b]	0	.78[b]
MP7 (31)	.33[b]	0	.89[b]
MP8 (34)	.46[b]	0	.79[b]
MP9 (28)	.33[b]	0	.89[b]
MP10 (46)	−.02	0	.99[b]
MP11 (49)	.41[b]	0	.83[b]
MP12 (52)	.38[b]	0	.85[b]
MP13 (55)	.26[b]	0	.93[b]
MP14 (58)	.53[b]	0	.72[b]
FP1 (5)	0	.39[b]	.85[b]
FP2 (8)	0	−.06	.99[b]
FP3 (17)	0	.48[b]	.77[b]
FP4 (23)	0	.53[b]	.72[b]
FP5 (26)	0	.56[b]	.68[b]
FP6 (29)	0	.57[b]	.67[b]
FP7 (35)	0	.60[b]	.65[b]
FP8 (38)	0	.23[b]	.95[b]
FP9 (41)	0	.71[b]	.50[b]
FP10 (44)	0	.58[b]	.67[b]
FP11 (50)	0	.09	.99[b]
FP12 (53)	0	.15	.98[b]
FP13 (56)	0	.29[b]	.91[b]
FP14 (59)	0	.59[b]	.65[b]

Scales	Factor correlations (PSI) MP	FP
MP	1	
FP	.58[b]	1

[a] $N = 269$ high-school students. Parameters with values of 0 and 1 were predetermined (i.e., fixed) and so tests of statistical significance were not possible. The numbers in parentheses refer to the item numbers in Bem (1974).

[b] $p < .05$.

analyzed separately demonstrated that the factor solution was reasonably invariant across the two age groups. In subsequent analyses, four BSRI scales were computed by summing responses across items in each of the four item groups (i.e., MP, FP, NP, NN), and these were correlated with the job aspiration scores.

Results and Discussion

CFA Models. For Model 9 (see Table IV) the 14 MP items and the 14 FP items are hypothesized to define an M and a F factor. These scales are well defined in that 24 of the 28 items load significantly and positively, and none of the remaining 4 items load significantly in a negative direction. The goodness-of-fit indicators (Table V) suggest that the fit is reasonable. Also, the two-factor solution represents a significant and substantial improvement over the one-factor solution (Model 10). The correlation between the M and F factors ($r = .58$; see PSI in Table IV) also suggests that while the two factors are correlated, they cannot be subsummed into a single factor, and certainly not into a bipolar factor in which the MF correlation would have to be negative. Consequently, the *direction* of this correlation is theoretically important. The direction of this correlation, while consistent with other research with the BSRI and also the PAQ, is exactly the opposite to that hypothesized in bipolar factors and is also, perhaps, inconsistent with the uncorrelated factors proposed by Bem (1974).

Correlation Between BSRI Scales and Job Aspirations. Correlations between the four BSRI scores, including the NP and NN scales, and the job aspiration variables appear in Table VI for the entire sample. The MP and FP scales are less correlated, even after correcting for attenuation, than in Table IV, but the direction is still positive and highly significant. However, the MP and FP scales are each more highly correlated with the NP scale than with each other. This appears to be consistent with the method-effect proposal, and suggests that the positive correlation between the MP and FP scales may be due to all the items in the MP and FP scales being positive (i.e., socially desirable). While the low estimated reliability of the NN scale makes its interperetation dubious, correlations between it and the other BSRI scales are also consistent with this explanation.

Girls who score lower on the MP scale are more likely to prefer and to be willing to settle for traditional jobs, while FP is not significantly cor-

Table V. Summaries of Alternative Models' Fit to the Bem Scales[a]

	χ^2	*df*	χ^2/df ratio	TLI	GFI	RMSR
Two-factor solution						
Model 9: MP and FP	763	349	2.19	.52	.81	.08
One-factor solution						
Model 10: MP/FP	882	350	2.52	.44	.73	.09
Null solution						
Model 11:28 uncorrelated factors	1716	378	4.54	.00	.50	.17

[a]See footnote a in Table II.

Table VI. Correlations Between Bem Scales and Job Aspiration Measures[a]

Variables	(1)	(2)	(3)	(4)	(5)	(6)
1 Masculine/Positive	(71)					
2 Feminine/Positive	28[b]	(71)				
3 Neutral/Positive	44[b]	66[b]	(72)			
4 Neutral/Negative	23[b]	06	−.7	(33)		
5 Preferred Job	22[b]	02	15[b]	02	–	
6 Job Settle For	28[b]	05	07	14[b]	54[b]	–

[a]Coefficients are presented without decimal points. Values in parentheses are coefficient alpha estimates of reliability. The Masculine/Positive and Feminine/Positive scales each contain 14 items, and the average of interitem correlations is 0.16. The Neutral/Positive and Neutral/Negative scales contain 8 and 6 items, respectively, and the average of interitem correlations are 0.25 and 0.08. The job aspiration scales vary between 0 and 1, where 0 indicates traditionally female jobs so that the two scales measure the nontraditionality of jobs the girls prefer and will settle for.
[b]$p < .05$.

related with either of the job aspiration variables. Since Year-10 girls, particularly those who do not plan to continue their schooling, face an imminent entry into the job market, the correlations were examined separately for the two year groups (see Table VII). For Year-8 girls, whether or not they plan to

Table VII. Correlations Between Bem Scales and Job Aspiration Measures for Subgroups[a]

	Year 8		Year 10	
	Leavers (n = 62)	Nonleavers (n = 75)	Leavers (n = 42)	Nonleavers (n = 74)
Correlation between preferred job and				
Masculine/Positive	−.05	−.07	.42[b]	.34[b]
Feminine/Positive	−.13	−.20	.12	.01
Neutral/Positive	−.15	−.07	.16	.10
Neutral/Negative	.03	−.11	.16	.12
Correlations between job settle for and				
Masculine/Positive	.13	−.12	.67[b]	.41[b]
Feminine/Positive	−.07	−.12	.20	.17
Neutral/Positive	−.09	−.18	.19	.16
Neutral/Negative	.18	−.09	.32	.18

[a]Leavers are those students who indicated that they plan to leave school at the end of Year 10, the typical "school-leaving" time in Australian schools. The job aspiration scales measure the nontraditionality of jobs the girls prefer and will settle for.
[b]$p < .05$.

continue schooling beyond Year 10, there is no significant correlation between any of the BSRI scales and either job aspiration variable. For Year-10 girls, particularly those not planning to continue school the following year, both job aspiration variables are substantially and positively correlated with the MP scale, but not with any of the other BSRI scales. For school leavers the traditionality of the jobs they are willing to settle for correlates with MP close to the limits of the reliability of the scale. These findings demonstrate that M scales, though perhaps not F scales, may have relevance for occupational choices when the changing structure of employment makes it critical for girls to consider jobs outside of gender stereotypes.

STUDY 3: M AND F USING CPS

The CPS is designed to measure eight bipolar dimensions of personality, one of which is a bipolar MF scale. Each personality dimension is defined by five item clusters, and each cluster contains two positively and two negatively worded items. Published factor-analytic studies of responses to the CPS have always been based on the 40 scores representing these item clusters rather than on responses to individual items, and Comrey (1970) provided a strong rationale for this approach. However, since each of the five MF item clusters is represented by a sum of M and F items, separate M and F factors are not possible. For purposes of this investigation, CFA was conducted on a correlation matrix representing the 20 individual items that comprise the MF scale (the authors are indebted to Andrew Comrey for providing them with this correlation matrix). In exploratory factor analyses of responses to just these 20 items, Comrey found that a five-factor solution clearly identified the five item clusters designed to define the MF scale, for the total sample considered here and in separate analyses of responses by males and females (A. L. Comrey, personal communication). The purpose of this analysis is to employ CFA to compare the goodness of fit of a five-factor solution based the design of the CPS with results from other models in which separate M and F factors are hypothesized.

Method

Sample and Materials. Data for Study 3 come from the original group used to norm the CPS (see Comrey, 1970, pp. 14–17 for further discussion). Subjects (362 males, 384 females) were either visitors to a university open house day, or university students. On the CPS, the MF scale is defined by five item clusters labeled to reflect the masculine end of the scale: "no fear

of bugs", "no crying", "no romantic love", "tolerance of blood", and "tolerance of vulgarity". Each of the clusters in turn is defined by four items, two scored in the M direction and two in the F direction (the item numbers as they appear in the CPS and the direction of their scoring appear in Table IX).

Statistical Analysis. The goodness of fit for a five-factor solution was compared with that obtained for a one-factor solution and for a two-factor solution in which separate M and F scales are hypothesized. CFA models similar in logic to those employed in studies 1 and 2 were used for this purpose. However, implicit in the design of the CPS and the logic of the five-factor solution is the assumption that these five factors combine to form a higher order factor that reflects a more general MF factor.

Limitations in the application of exploratory factor analysis are even more critical in the analysis of higher order factor models. However, recent advances in CFA in the analysis of higher order factor structures do not have many of these weaknesses (Bentler & Weeks, 1980; Joreskog, 1980; Joreskog & Sorbom, 1981; Marsh, 1985, in press-b; Marsh & Hocevar, 1985; Olson, 1983; Tanaka & Huba, 1984). The technical details of how analyses are performed are beyond the scope of this paper but the procedure used here is similar to that described by Marsh 1985, in press b; Marsh & Hocevar, 1985).[3] The logic of this analysis is a straightforward extension of the analysis of first-order structures described earlier. The solution based on the design of the CPS hypothesizes five first-order factors corresponding to the item clusters, and the correlations among these factors appear in the PSI matrix of factor correlations. Implicit in the design of this model is the assumption that these five factors are all positively correlated with each other and combine to form a higher order MF factor. In order to test this assumption, a sixth higher order factor is defined by each of the five first-order factors, and this factor is hypothesized to account completely for correlations among the first-order factors. Thus, the 10 correlations among the five first-order factors are explained in terms of a single second-order factor. Conceptually, it is as if the

[3]For purposes of the higher order model, the factor loading of one measured variable for each of the five first-order factors was fixed to be 1.0, and it served as a reference indicator. The factor variances in the PSI matrix, including the second-order factor, were then freed and estimated by the LISREL program. Factor loadings for the second-order factor were estimated in the beta matrix described by Joreskog and Sorbom (1981). The formulation of such a model for higher order CFA and its rationale is described by Marsh 1985, in press-b; Marsh & Hocevar, 1985). For the purposes of the present investigation, analyses were conducted on the total population, but Marsh (1985) demonstrated that both the first-order and second-order factor structures examined here are relatively invariant across responses by males and responses by females.

correlations among the first-order factors were the basis of a second factor analysis. Since the higher-order factor is merely trying to explain the correlations among first-order factors in a more parsimonious way (i.e., one that requires fewer estimated parameters), even when the higher order model is able to explain the factor correlations, the goodness of fit for the higher order model will produce a chi-square value no better than the corresponding first-order model. If the goodness-of-fit indicators for the higher order model do not differ substantially from the corresponding first-order solution, then the hierarchical ordering of the factors is supported.

Results and Discussion

First-Order Models. In Model 12, the five-factor solution based on the design of the CPS (see Table VIII), each of the five factors is well defined in that every item loads in the hypothesized direction, the loading is statistically significant, and is substantial. Model 12 is based on five bipolar factors, each of which is designed to measure one component of a more general bipolar MF scale. Inspection of the factor loadings in Table VIII demonstrates that each of these factors is bipolar in that all M items load positively in the M direction, and all F items load negatively in the F direction. The statistically significant, positive correlation between each pair of factors is also consistent with the design of the CPS. The goodness-of-fit indicators (see Table IX) demonstrate that the model does a good job of explaining the data. While the chi-square and chi-square/*df* ratio for Model 10 is somewhat higher than for the best models in Studies 1 and 2, this is due to the substantially larger sample size employed in Study 3; the goodness-of-fit indicators that are less effected by sample sizes (i.e., TLI, GFI, and RMSR) are substantially better than in Studies 1 and 2.

Models 13 and 14 hypothesize a single bipolar MF factor (Model 13) and a two-factor solution in which M and F are separate but correlated traits (Model 14). Inspection of the factor loadings for the single factor in Model 13 clearly demonstrates that it is a bipolar factor, but inspection of the fit indices shows that the fit of this model is substantially poorer than that of Model 12. However, the goodness of fit for Model 14 is little better than Model 13 (see Table IX), and also fails to explain the data nearly as well as Model 12. Furthermore, the estimated correlation between the M and F factors in Model 14 ($r = -1.07$) is slightly larger than -1.0. The fact that the correlation is more negative than -1.0, even if only slightly, means that the solution is improper and may suggest the inadequacy of the model. The size of the correlation in Model 14, whether interpreted as not differing substantially from -1.0 or as indicative of a poor model, coupled with the

Table VIII. CFA of CPS with Five Factors Representing Item Clusters Designed to Measure Masculinity–Femininity[a]

Item number and direction	Factor loadings (LAMBDA) I	II	III	IV	V	Error/uniqueness (THETA)
53 (+)	.72[b]	0	0	0	0	.48[b]
143 (+)	.80[b]	0	0	0	0	.37[b]
8 (−)	−.69[b]	0	0	0	0	.53[b]
98 (−)	−.55[b]	0	0	0	0	.70[b]
71 (+)	0	.71[b]	0	0	0	.50[b]
161 (+)	0	.91[b]	0	0	0	.17[b]
26 (−)	0	−.84[b]	0	0	0	.29[b]
116 (−)	0	−.62[b]	0	0	0	.62[b]
89 (+)	0	0	.75[b]	0	0	.44[b]
179 (+)	0	0	.53[b]	0	0	.72[b]
44 (−)	0	0	−.78[b]	0	0	.40[b]
134 (−)	0	0	−.33[b]	0	0	.89[b]
62 (+)	0	0	0	.89[b]	0	.21[b]
152 (+)	0	0	0	.86[b]	0	.26[b]
17 (−)	0	0	0	−.68[b]	0	.54[b]
107 (−)	0	0	0	−.27[b]	0	.93[b]
80 (+)	0	0	0	0	.68[b]	.54[b]
170 (+)	0	0	0	0	.67[b]	.55[b]
35 (−)	0	0	0	0	−.74[b]	.46[b]
125 (−)	0	0	0	0	−.59[b]	.66[b]

Scales	Factor correlations (PSI) I	II	III	IV	V
I	1				
II	.40[b]	1			
III	.47[b]	.43[b]	1		
IV	.43[b]	.29[b]	.17[b]	1	
V	40[b]	.35[b]	.27[b]	.25[b]	1

[a]Parameters with values of 0 and 1 were predetermined (i.e., fixed) and so tests of statistical significance were not possible. Item numbers are from the CPS. Those marked "+" are scored as masculine and those marked "−" are scored as feminine. The five item clusters are labeled: no fear of bugs, no crying, no romantic love, tolerance of blood, and tolerance of vulgarity.

[b]$p < .05$.

similarity of goodness of fit for Model 13 and 14, provides strong support for the bipolarity of the MF factor consistent with the design of the CPS. Nevertheless, neither of these models do nearly as well as the five-factor solution employed in Model 12, and thus both are rejected.

In the analyses described thus far, one bipolar MF factor explains the data as well as separate M and F factors, but the five-factor solution does better yet. However, Model 12 is also consistent with the bipolarity assumption in that each of the five factors is defined by two M items that load

Table IX. Summaries of Alternative Models' Fit to the CPS Data[a]

	χ^2	*df*	χ^2/df ratio	TLI	GFI	RMSR
Five-factor solution						
Model 12: original item clusters (see Table VIII)	482	160	3.01	.90	.88	.06
Two-factor solution						
Model 13: M and F	2968	169	17.56	.40	.58	.12
One-factor solution						
Model 14: M and F combined	2987	170	17.57	.40	.58	.12
Model 15: original item clusters broken into M and F components	390	125	3.12	.89	.87	.05
Higher-order solution						
Model 16: five first-order factors and one higher-order factor	511	165	3.10	.89	.88	.06
Null solution						
Model 17: 40 uncorrelated factors	5548	190	29.20	.00	—	—

[a]TLI, Tucker–Lewis indicator; GFI, goodness-of-fit indicator; RMSR, root mean square residual. Factors defined in the various solutions comprise combinations of items representing the five item clusters, and each item cluster is made up of half masculine and half feminine items. In the higher order solution the five first-order factors represent the five item clusters while the higher order factor is designed to explain the correlations among these factors with a single bipolar masculinity–femininity factor.

positively and two F items that load negatively. In Model 15, a ten-factor solution is proposed, dividing each of the five factors in Model 12 into separate M and F components. However, for Model 15 none of the five correlations between the M and F components of the same content-factor differed substantially from -1.0 (*r*s of -1.1, $-.99$, $-.97$, -1.1, $-.93$, and -1.0). Also, the goodness-of-fit indicators for Model 15 suggest little improvement over Model 12. Consequently, the comparison of Models 12 and 15, as does the comparison of Models 13 and 14, provides strong support for the bipolarity of the MF construct as measured by responses to the CPS.

Higher Order Factor Solution. Support for Model 12 suggests that the CPS measures five distinguishable factors designed to reflect the MF construct, and that the MF component in each of these factors is bipolar. However, the data is not adequately explained by a single bipolar scale that combines the five factors into a single MF factor (Model 13). The question to be examined here is how well the correlations among the five first-order factors can be explained by a single higher order factor (Model 16 in Figure 1). While a detailed presentation of technical aspects of this analysis is beyond the scope of this investigation, conceptually the analysis is as if the factor correlation matrix for Model 12 (PSI in Table VIII) was factor analyzed and a one-factor

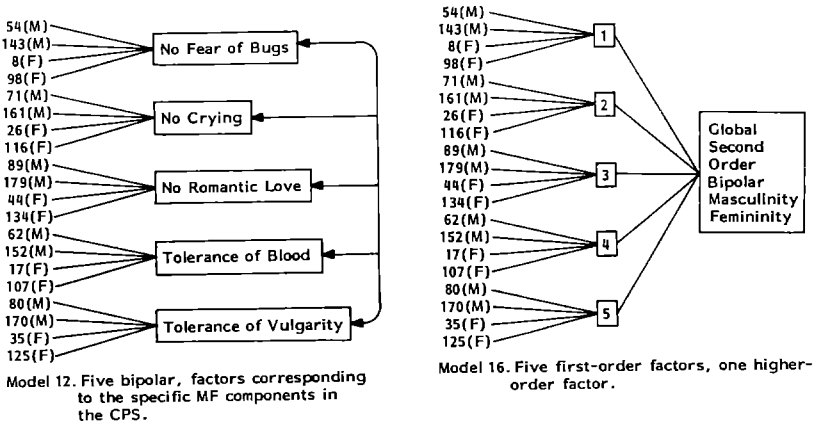


Fig. 1. Higher order factor analysis of CPS data.

solution was tested. According to this model, each of the 20 CPS items reflects one of the five first-order factors, and each of the five first-order factors reflects the higher order factor (see Figure 1).

The parameter estimates for the higher-order CFA indicate that each of the 20 items loads significantly in the appropriate direction in defining the first-order factors, and that each of the first-order factors loads positively and significantly in the definition of the second-order factor. Furthermore, the goodness-of-fit indicators for Model 16 are nearly as good as those for Model 12. This suggests that the relationships among the five first-order factors designed to measure a general MF construct can adequately be explained by a single higher order factor, and this provides further support for the design of the CPS.

While a higher order CFA as illustrated in Fig. 1 has not previously been performed with responses to the CPS, the logic of the analysis is consistent with exploratory factor analyses described by Comrey (1970). In CFA, the first-order factors are defined by responses to individual items, and then the correlations among these first-order factors are used to infer a higher-order factor. Comrey used an unweighted average of individual items that represent our first-order factors, and inferred more general factors on the basis of correlations among them. Nevertheless, the facility to summarize the goodness of fit and to compare the fit with competing models is an important advantage not possible with conventional approaches to exploratory factor analysis. The conclusions described here are also consistent with suggestions by Spence, by Pedhauzer, by Antill, by Constantinople, and perhaps even by Bem, that global F and global M factors cannot be ade-

quately defined by single unidimensional factors. While those researchers may not agree with the particular factors chosen by Comrey to represent M and F, nor with its bipolar representation imposed by the use of logically opposed items, the hierarchical CFA approach is consistent with their arguments.

OVERVIEW, SUMMARY, AND IMPLICATIONS

This investigation offers important methodological, conceptual, and substantive contributions to the growing body of MF research. Methodologically, this is apparently the first reported application of CFA in MF research, and it demonstrates the advantages of CFA over exploratory factor analysis. Also, the extension of CFA to test the higher order model in Study 3 offers an analytic tool to test hypotheses about global M and F suggested by other researchers but not tested previously. Conceptually and substantively, the investigation provides a demonstration of how observed correlations between M and F are substantially influenced by the selection and design of items used to infer the constructs, and argues for the multidimensionality of global M and F constructs.

In 1973, Constantinople criticized MF research in that (a) the implicitly assumed bipolar relation between M and F was not tested, (b) the implicitly assumed unidimensionality of global M and/or global F was not tested, and (c) strategies used to select/construct MF instruments were largely atheoretical and offered a weak basis for developing and refining theory. MF research during the last decade has focused almost exclusively on the first point, and the other two seem to have been largely ignored or not incorporated into the measurement of MF constructs. However, a review of this research and the findings of the present investigation offer further support for the continued relevance of all three criticism (see Cook, 1985; Myers & Gonda, 1982; Spence, 1984 for further discussion).

Size and Direction of MF Correlations

Personality and androgyny researchers argue for the importance of M and F as psychological constructs, though they may disagree about whether M and F should be conceived as a single bipolar construct, two distinguishable constructs, or as global, higher order factor(s) defined by lower level factors that reflect specific components of M and F. Broad personality inventories have typically defined M and F to represent a single, bipolar scale, while androgyny researchers have proposed M and F to be two separate, distinguishable traits. Results of CFAs on responses to three instruments used

to measure M and F have shown the two traits to be somewhat positively correlated (BSRI), somewhat negatively correlated (ASRS), and almost perfectly negatively correlated (CPS). While these findings are remarkably inconsistent with each other, they are each explicable in terms of the selection and construction of items used in each instrument.

BSRI items represent primarily socially desirable characteristics chosen to represent M and F stereotypes. Thus it is likely that a method effect, such as social desirability, will increase the apparent similarity in responses to the M and F scales on the BSRI. The method effect will act to produce an observed correlation more positive than the "true" MF correlation. When the M and F scale is definitely by an unweighted sum of responses to M and F items, and the correlation between M and F is not corrected for unreliability, the findings here (Table V) and elsewhere suggest that the correlation is small and positive. Use of CFA yields a correlation between the two factors that is also positive, but somewhat larger (Table IV). This reflects the correction for attenuation in the CFA analysis, and also perhaps reflects that items that more strongly reflect both social desirability *and* M (or F) are likely to load more highly on the M (or F) factors. Thus, while the size of the positive correlation may be somewhat surprising, the direction of the correlation and its proposed explanation are consistent with other research and the design of the BSRI (and the PAQ).

ASRS items use half socially desirable and half socially undesirable characteristics to represent masculine and feminine stereotypes. Thus, for total M and total F scores, the influence of social desirability as a method effect is likely to be nullified. Nevertheless, the superiority of the four-factor solution—MP, MN, FP, and FN—suggests that the influence may still operate. The correlations between M and F scales and self-esteem also vary drastically, depending on whether the M and F scales are defined by positively or negatively valued items. Consistent with the method-effect hypothesis, the observed MF correlation for ASRS responses (Model 13) is moderately negative, and much more negative than the MF correlations for responses to the BSRI (Model 9). The direction of the MF correlation is similar to findings with the same instrument reported by Antill et al. (1981), and by Hong et al. (1983). Even with the ASRS, the selection of items works to underestimate the MF correlation in that Antill et al. reasoned: "Independence of the resultant scales was also deemed important so that items that correlated highly with a scale to which they had not been allocated were removed. This criterion was applied strictly to M+ or M- items correlating with F+ or F- scales or vice versa" (p. 176).

CPS items were not selected to be either socially desirable or undesirable, and correlations between responses to the items clusters are nearly uncorrelated with the Response Bias scale used to infer social desirability (Comrey, 1970, Table XI). However, CPS items were specifically

selected/constructed to represent logical opposites, and thus it is not surprising that the correlation between M and F scales is more negative than is observed with either of the other instruments. Nevertheless, the size of the negative correlation, approximating -1.0 after correction for attenuation, was surprisingly high.

The observed correlation between M and F scales apparently depends to a considerable extent on the way in which items are selected or constructed, and so it is difficult to say what the "true" correlation is. If, as with the CPS, M and F items are logically opposed (or literal opposites), then the scales are likely to be so negatively correlated that they can be adequately characterized as bipolar. If, as with the ASRS, M items are selected that are least correlated with the F scale, and vice versa, then the correlation between M and F scales is likely to be only modestly negative or to even approach zero. If, as with the BSRI and PAQ, M and F items are substantially alike on other, perhaps irrelevant, characteristics such as social desirability, then the correlation between M and F scales is likely to be close to zero or positive. The substantial negative correlation between self-descriptions on the single items "Masculine" and "Feminine" that appeared on the original BSRI, and the moderately negatively correlation between M and F scales on the ASRS (despite the selection bias in items to counteract this correlation), provides evidence that the direction of the correlation is negative, but it seems unlikely that the size of this correlation is so negative that MF constructs can be adequately explained as opposite ends of a bipolar continuum, unless items are specifically selected/constructed to be logically/literally opposed or to reflect some external criterion that is bipolar (e.g., gender). Nevertheless, the question of whether or not MF is bipolar cannot be answered by empirical findings and must be resolved on the basis of subsequent research into the construct validity of MF measures derived from these different approaches.

Multifaceted Higher Order M and F Constructs: A New Model

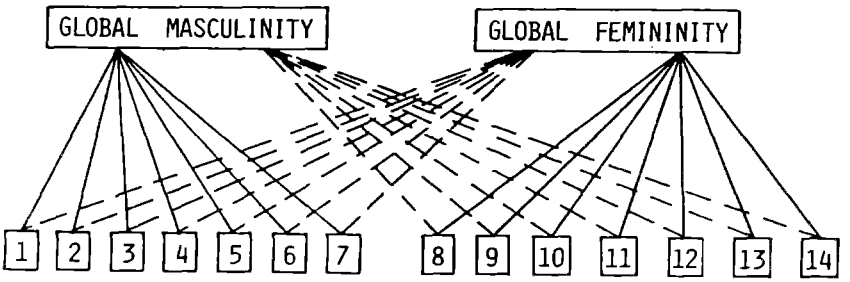
The inability of the two-factor, M and F, solution to explain responses to the ASRS and the CPS adds to the mounting evidence against the unidimensionality of global F and global M. This lack of unidimensionality was clearly proposed in the Constantinople (1973) review, but this has been largely ignored in the construction of subsequent MF measures. Consistent with the theoretical design of the CPS, M and F can more appropriately be viewed as higher-order constructs defined by a variety of specific traits. However, unlike the CPS scale, such a hierarchical model does not imply that M and F must represent a bipolar construct. Instead, separate hierarchies are likely to exist for M and F constructs, and to result in two correlated higher order traits. While the actual value of the correlation between higher

order M and F factors will depend on the construction/selection of items, the content of lower order factors, and perhaps the subject population, the direction will probably be negative.

The multifaceted hierarchical perspective of global M and global F is apparently consistent with the conceptualization of Bem, Spence, and their colleagues, even though it is not reflected in the design of the PAQ, the EPAQ, the BSRI, and other androgyny instruments. However, a conceptual model based on this logic could be used to construct new instruments more firmly based on an explicit theoretical model, and more amenable to empirical tests. Systematic reviews of MF research emphasize the lack of a theoretical basis for most of the measurement instruments employed in this field. As psychological constructs, M and F are hypothetical constructs whose usefulness must be demonstrated by investigations of their construct validity. The determination of whether theoretically consistent and distinct facets of global M and global F exist, and the determination of their content if they do exist, should be prerequisite to the study of how these facets, or the global constructs that they represent, are related to other variables. In recommending such an approach, we, as did Pedhauzer and Tetenbaum (1979), adamantly reject the atheoretical, empirical approach often used to develop M and F scales. Instead, an explicit theoretical model should be the starting point for instrument construction, and empirical results should be used to support, refute, or revise the instrument *and* the theory on which it is based. In applying such an approach, the first step is to review theoretical and empirical research to determine the lower order factors that underlie the global M and F constructs, and this is the step that has apparently been neglected in the construction of most instruments. Once the lower order factors have been explicitly hypothesized, the construction of appropriate items to measure these specific factors will be much easier than when the appropriate constructs have not been adequately defined. Finally, the empirical testing of the proposed higher order structure with CFA will constitute an important part of the demonstration of the construct validity of interpretations based upon responses to the proposed instrument. Implicit in this approach is the edict that theory building and instrument construction are inexorably intertwined, and that each will suffer if the two are separated.

One possible representation of a multifaceted hierarchical model of global M and global F is illustrated in Figure 2.[4] According to this model,

[4]The structure of the multifaceted hierarchical model and the propositions used to further define the theoretical model were stimulated in part by the Shavelson model of self-concept (Shavelson, Hubner & Stanton, 1976) and research based on the model described by Marsh and Shavelson (1985).



Masculine Specific Factors

1 = Rational, quantitative, mathematically oriented
2 = Goal Directed, success/achievement oriented
3 = Aggressive, dominant, need to control
4 = Self-sufficient, autonomous, independent
5 = Competitive, assertive
6 = Tough, vulgar, insensitive
7 = Physical, athletically oriented

Feminine Specific Factors
8 = Emotional, anxious, cries easily
9 = Dependent, submissive, yielding, passive
10 = Nurturant
11 = Traditional, conventional
12 = Empathetic, sensitive to interpersonal needs
13 = Romantic, love oriented
14 = Verbally expressive, verbal/language oriented

Fig. 2. One possible representation of a multifaceted, higher order representation of the masculinity and femininity constructs.

global F and global M are negatively correlated with each other, and each is defined by specific traits that have been suggested in previous research. The dotted lines connecting first-order F factors to the higher order M factor, and first-order M factors to the higher order F factor, represent the possibility that the same first-order factor contributes to both higher order factors. If this occurs, the first-order factor is hypothesized to load positively and substantially on the higher order factor that it is specifically designed to measure, and negatively and less substantially on the other higher order factor.

The number and the content of the first-order factors used to define each global factor in Figure 2, though reasonable, are merely heuristic for the demonstration of a multifaceted hierarchical model of global masculinity and femininity. This theoretical model can be further defined by the set of hypotheses presented below, which, though reasonable, are also designed

to be heuristic. It would be surprising if subsequent research does not offer substantial improvements to this model, and does not lead to the revision or outright rejection of some of the facets and some of the hypotheses. However, such a systematic interplay between theory and research will lead to a better understanding of the MF construct. The hypotheses are as follows:

1. The facets that comprise global M and global F are structured and organized. Individuals summarize much information about their own sex self-concepts with these facets and relate these facets to one another. Nevertheless, consistent with Bem's gender schema theory, individuals may differ in the extent to which they use this structure to process incoming information.
2. The facets and their structure reflect a category system adopted by a particular individual, are influenced by group membership, and are a function of sex stereotypes that exist in a particular culture at a given period in time.
3. A similar structure will exist for both males and females.
4. Gender (1 = male, 2 = female) will be (a) positively correlated with global F and each of the specific facets that comprise it (b) negatively correlated with global M and each of the facets that comprise it.
5. Responses to the adjectives *Masculine* and *Feminine* (and synonyms) will be more negatively correlated than will scores for global M and global F, and will tend to form a distinguishable bipolar scale that will be substantially correlated with both global M and global F. Furthermore, most of the correlation between gender and the two global constructs will be explicable in terms of this two-item, bipolar scale (i.e., the correlations between gender and the two global constructs, after controlling for the bipolar MF scale, will be very small). [Marsh (1986) notes that self-concept researchers alternatively define global self-concept in terms of a hierarchy such as posited here or in terms of responses to a relatively unidimensional scale—sometimes called esteem—that may be like the proposed MF scaoe defined by the adjectives masculine and feminine (and their synonyms).]
6. Across a representative sample, the correlation between global M and global F will be negative, but the size of the correlation will vary systematically for adults in different subpopulations. The correlation will be less negative when educational level and/or SES is high, and in certain occupational categories and other subgroups where sexual stereotypes are weak.
7. The size of the negative correlation between global M and global F will vary systematically with age. The correlation will be most negative during adolescent years when sex typing is a typical development stage. (Lamke, 1982, based on theoretical work by Erikson,

Kohlberg, and others, suggested that the adoption of stereotypic sex stereotypes during adolescence is healthy, whether or not more androgynous stereotypes are desirable at other ages.)

8. M and F will become increasingly multifaceted with age, maturity, and experience in that (a) the number of specific facets may increase, (b) the lower order factors may become more clearly defined, (c) the magnitude of correlations among first order facets may become smaller, or (d) the proportion of variance in lower order factors explained by the global factors may decrease.
9. Specific facets of M and F, and particularly global M and global F, are self-evaluative as well as self-descriptive. The evaluative component will be particularly strong during early adolescent years when individuals are more sensitive to sex role stereotypes and in subpopulations where conformity to sex role stereotypes is more pervasive.
10. Specific facets of M and F will be more highly correlated with specific sex-related criteria and behaviors to which they are most logically and theoretically related—more highly correlated than other first-order factors, and more highly correlated than the global scores. Particularly during adult years when the hierarchical ordering of the facets is weaker, psychological M and F will not be adequately summarized by global M and F or the adjectives "masculine" and "feminine." The logic of this approach to construct validity is related to multitrait-multimethod analyses where validity is inferred when a construct is most highly correlated with other constructs to which it is most logically/theoretically related, and less highly correlated with other constructs.
11. As posited by androgyny theory, M and F facets will each contribute significantly and uniquely to the prediction of sex-related criteria, but the size and direction of these effects will vary with the specific criteria. In particular, the positive contributions of F facets will be larger than M facets for relevant criteria independently determined to be feminine, while the contribution of M will be larger for criteria independently determined to be masculine (see Marsh, in press-a, for a demonstration of how the circularity of this hypothesis can be broken; also see footnote 2).

The Current Status of Androgyny

Twelve years after Bem's first formulation of androgyny theory, few researchers claim that psychological masculinity and femininity represent

bipolar ends of a single continuum, yet debate continues on the definition of androgyny, its measurement, and its relation to external criteria of social effectiveness and competency. We question neither the utility of the concept of androgyny, nor the existence of males and females whose self-images are high in both psychological masculinity and femininity. The social zeitgeist reflected in the women's movement helped stimulate androgyny research, and this research has been productive.

The model posited above is based on the need to look at the unique contributions of both global M and global F not as unidimensional constructs combining to form sex role identities labeled as androgyny, sex-typed, or undifferentiated, but as two higher order factors reflecting a complex of more specific facets. Within this model androgyny does not represent a single construct, but rather a theoretical hypothesis about the relationship between global M and global F, and their relationship to other constructs such as self-concept and social competency. In this model, the degree of sex typing for a particular subpopulation is reflected in the size of the correlation between global M and global F, and perhaps in the strength of the hierarchy connecting specific facets to the higher order global constructs. While it may be possible to collapse information from the multiple facets into one of three or four sex role classifications, we find such a classification as overly simplistic and counter to the richness and diversity of self-images that individuals of both genders actually have. The denial of such richness seems opposed to the aim of androgyny research to demonstrate that existing sex stereotypes are too narrow, too rigidly defined, and too confining. Future research with this model may reflect this richness and diversity.

REFERENCES

Antill, J. K., & Cunningham, J. D. Self-esteem as a function of masculinity in both sexes. *Journal of Consulting Psychology,* 1979, *47,* 783-785.

Antill, J. K., & Cunningham, J. D. The relationship of masculinity, femininity, and androgyny to self-esteem. *Australian Journal of Psychology,* 1980, *32,* 195-207.

Antill, J. K., & Russell, G. A preliminary comparison between two forms of the Bem Sex Role Inventory. *Australian Psychologist,* 1980, *15,* 427-435.

Antill, J. K., Cunningham, J. D., Russell, G., & Thompson, N. L. An Australian Sex-Role Scale. *Australian Journal of Psychology,* 1981, *33,* 169-183.

Baumrind, D. Are androgynous individuals more effective persons and parents? *Child Development,* 1982, *53,* 44-75.

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

Bem, S. L. Sex role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology,* 1975, *31,* 634-643.

Bem, S. L. On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology,* 1977, *45,* 196-205.

Bem, S. L. Theory and measurement of androgyny—A reply to Pedhauzer, Tetenbaum, and Locksley-Colten critiques. *Journal of Personality and Social psychology,* 1979, *37,* 1047-1054.

Bentler, P. M., & Bonett, D. G. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin,* 1980, *88,* 588-606.

Bentler, P. M., & Weeks, D. G. Linear structural equations with latent variables. *Psychometrica,* 1980, *45,* 289-308.

Comrey, A. L. *Manual for Comrey Personality Scales.* San Diego, CA: Educational and Industrial Testing Service, 1970.

Cook, E. P. *Psychological androgyny.* Pergamon Press: New York, 1985.

Constantinople, A. Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin,* 1973, *80,* 389-407.

Crandall, R. The measurement of self-esteem and related constructs. In J. Robinson & P. Shaver (Eds.), *Measurements of social psychological attitudes.* Ann Arbor, MI: Institute for Social Research, 1973.

Cunningham, J. D., & Antill, J. K. A comparison among five masculinity-femininity-androgyny instruments and two methods of scoring androgyny. *Australian Psychologists,* 1980, *15,* 437-448.

Farnill, D., & Ball, I. L. Some psychometric data on the Australian Sex-Role Scale. Paper presented at the 17th Annual Conference of the Australian Psychological Society Melbourne 1982. (a)

Farnill, D., & Ball, I. L. Some psychometric data on the Australian Sex-Role Scale (Summary). *Australian Psychologist,* 1982, *17,* 349. (b)

Farnill, D., & Ball, I. L. Male and female factor structures of the Australian Sex-Role (Form A). *Australian Psychologists,* 1985, *20,* 205-213.

Feather, N. T. Factor structure of the Bem Sex Role Inventory: Implications for the student of masculinity, femininity and androgyny. *Australian Journal of Psychology,* 1978, *30,* 241-254.

Fornell, C. Issues in the application of covariance structure analysis. *Journal of Consumer Research,* 1983, *9,* 443-448.

Gaudreau, P. Factor analysis of the Bem Sex Role Inventory. *Journal of Consulting and Clinical Psychology,* 1977, *45,* 299-302.

Gough, H. G. *Manual for the California Psychological Inventory* (rev. ed.). Palo Alto, CA: Consulting Psychological Press, 1969.

Heilbrun, A. B. Measurement of masculine and feminine sex role identities as independent dimensions. *Journal of Consulting and Clinical Psychology,* 1976, *44,* 183-190.

Heilbrun, A. B. Sex-based models of androgyny: A further cognitive elaboration of competence differences. *Journal of Personality and Social Psychology,* 1984, *46,* 216-229.

Hong, S. U., Kavanagh, K., & Tippett, V. Factor structure of the Australian Sex Role Scales. *Psychological Reports,* 1983, *53,* 499-505.

Hull, C. H., & Nie, N. H. *SPSS update 7-9.* New York: McGraw-Hill, 1981.

Joreskog, K. G. Structural analysis of covariance and correlation matrices. *Psychometrica,* 1980, *43,* 443-477.

Joreskog, K. G., & Sorbom, D. *LISREL V: Analysis of linear structural relations by the method of maximum likelihood.* Chicago: International Educational Services, 1981.

Kelly, J., & Worrell, J. L. New formulations of sex roles and androgyny: A critical review. *Journal of Consulting and Clinical Psychology,* 1977, *45,* 1101-1115.

Kelly, J., Caudill, S. Hathorn, S., & O'Brien, C. Socially undesirable sex-correlated characteristics: Implications for androgyny and adjustment. *Journal of Consulting and Clinical Psychology,* 1977, *45,* 1186-1187.

Lamke, L. K. The impact of sex-role orientation on self-esteem in early adolescence. *Child Development,* 1982, *53,* 1530-1535.

Lee, A. G., & Scheurer, V. Psychological androgyny and aspects of self-image in women and men. *Sex Roles,* 1983, *9,* 289-306.

Long, K. S. *Confirmatory factor analysis.* Beverly Hills, CA: Sage, 1983.

Lubinski, D., Tellegen, A., & Butcher, J. N. Masculinity, femininity, and androgyny viewed and assessed as distinct concepts. *Journal of Personality and Social Psychology,* 1983, *44,* 428-439.

Marsh, H. W. Masculinity, femininity and androgyny: Their relations with multiple dimensions of self-concept. *Multivariate Behavioral Research,* in press. (a)

Marsh, H. W. The hierarchical structure of self-concept: An application of hierarchical confirmatory factor analysis. *Journal of Educational Measurement,* in press. (b)

Marsh, H. W. The structure of masculinity/femininity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. *Multivariate Behavioral Research,* 1985, *21,* 427-449.

Marsh, H. W., & Hocevar, D. Confirmatory factors analysis of multitrait-multimethod matrices. *Journal of Educational Measurement,* 1983, *20,* 231-248.

Marsh, H. W., & Hocevar, D. The factorial invariance of students' evaluations of college teaching. *American Educational Research Journal,* 1984, *21,* 341-366.

Marsh, H. W., & Hocevar, D. The application of confirmatory factor analysis to the study of self-concept: First- and higher-order factor models and their invariance across groups. *Psychological Bulletin,* 1985, *97,* 562-582.

Marsh, H. W., & Shavelson, R. Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist,* 1985, *20,* 107-125.

Megargee, E. I. *The California Psychological Inventory handbook.* San Francisco: Jossey-Bass, 1972.

Myers, M. R. *The relationship of androgyny and attributions for outcome to occupational behavior.* Unpublished thesis, Department of Education, University of Sydney, 1982.

Myers, A. M., & Gonda, G. Utility of the masculinity-femininity construct: Comparison of traditional androgyny approaches. *Journal of Personality and Social Psychology,* 1982, 514-522.

Nicholson, S. I., & Antill, J. K. Personal problems of adolescents and their relationship to peer acceptance and sex-role identity. *Journal of Youth and Adolescence,* 1981, *10,* 309-325.

Olson, G. H. *Covariance structure analysis involving measured variables assumed to have higher-order factor structures.* Paper presented at the annual meeting of the American Educational Research Association, Division D, Montreal, Canada, 1983.

Pedhauzer, E. J., & Tetenbaum, T. J. The Bem Sex Role Inventory: A theoretical and methodological critique. *Journal of Personality and Social Psychology,* 1979, *37,* 996-1016.

Russell, G., & Antill, J. An Australian Sex-Role Scale: Additional psychometric data and correlations with self-esteem. *Australian Psychologist,* 1984, *19,* 13-18.

Shavelson, R., Hubner, J. J., & Stanton, G. C. Self-concept: Validation of construct interpretations. *Review of Educational Research,* 1976, *46,* 407-441.

Silvern, L., & Ryan, V. Self-rated adjustment and sex-typing of the Bem Sex Role Inventory: Is masculinity a primary factor of adjustment? *Sex Roles,* 1979, *5,* 739-763.

Spence, J. T. Comment on Lubinski, Tellegen, and Butcher's "Masculinity, femininity, and androgyny viewed and assessed as distinct concepts." *Journal of Personality and Social Psychology,* 1983, *44,* 440-446.

Spence, J. T. Masculinity, femininity and gender-related traits: A conceptual analysis and critique of recent research. In B. A. Maher & W. B. Maher (Eds.), *Progress in experimental personality research,* (Vol. 13). New York: Academic Press, 1984.

Spence, J. T., & Helmreich, R. L. On assessing androgyny. *Sex Roles,* 1979, 5, 721-738. (a)

Spence, J. T., & Helmreich, R. L. The many faces of androgyny: A reply to Locksley and Colten. *Journal of Personality and Social Psychology,* 1979, *37,* 1032-1042. (b)

Spence, J. T., & Helmreich, R. L. Androgyny versus gender schema: A comment on Bem's gender schema theory. *Psychological Review,* 1983, *88,* 365-368.

Spence, J. T., Helmreich, R. L., & Holahan, C. K. Negative and positive components of masculinity and femininity and relations to self-reports of neurotic and acting out behaviors. *Journal of Personality and Social Psychology,* 1979, *37,* 1673-1682.

Tanaka, J. S., & Huba, G. J. Confirmatory factor analyses of psychological distress measures. *Journal of Personality and Social Psychology,* 1984, *46,* 621-635.

Taylor, M. C., & Hall, J. A. Psychological androgyny: A review and reformulation of theories, methods and conclusions. *Psychological Bulletin,* 1982, *92,* 347-366.

Whitely, B. E. Sex role orientation and self-esteem: A critical meta-analytic review. *Journal of Personality and Social Psychology,* 1983, *44,* 447-455.