5-1-1983

The Sex Typed Activities Test: A Supplementary Measure of Masculine and Feminine Dimensions of Personality

Steven Harold Weaver

The Sex Typed Activities Test: A supplementary measure of masculine and feminine dimensions of personality

by
Steven Harold Weaver

B.A. University of Kansas 1972
M.A. University of North Dakota 1979

A Dissertation

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Grand Forks, North Dakota

May
1983

This dissertation submitted by Steven Harold Weaver in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from the University of North Dakota is hereby approved by the Faculty Advisory Committee under whom the work has been done.

James A. Clark
(Chairperson)

Lola P. Tabor

John

Burdick Hedahl

Owen R. Larson

This dissertation meets the standards for appearance and conforms to the style and format requirements of the Graduate School of the University of North Dakota, and is hereby approved.

G. William Johnson
Dean of the Graduate School

Title The Sex Typed Activities Test: A supplementary measure of masculine and feminine dimensions of personality

Department Psychology

Degree Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in his absence, by the Chairman of the Department or the Dean of the Graduate School. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Signature _____

Date _____

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ACKNOWLEDGMENTS

In spite of the long hours of solitary effort involved in a project of this sort, no dissertation and no research effort is the product of a single individual. I would like to express my gratitude to a number of individuals for their help and support. Very special thanks go to the Chairperson of my Dissertation Committee, Dr. James Clark, for his interest and judicious supervision. My appreciation is also extended to the other members of that committee for their comments and assistance: Dr. Lila Tabor, Dr. Beulah Hedahl, Dr. Omer Larson, and Dr. John Tyler.

I owe a tangible debt to three people who helped to collect the data on which this paper is based. These are Len Burns, of the University of Connecticut; and Cris Scaglione and Dr. Paul Wright, both of the University of North Dakota.

My appreciation goes to the many students who participated in this study by completing questionnaires. In addition, I am grateful to two individuals whom I know only through their collaborative work on sex role measurement: Janet T. Spence and Robert L. Helmreich. Their careful empirical studies and theoretical essays provided the necessary foundation for the work which is presented in this paper.

Because opportunities of this sort are so rare, I would like to take this one to also express my profound gratitude to two teachers whose influence propelled me in this direction, Mrs. Eldora Farley and Dr. Ann Ruth Willner. The first taught me how to read, and the second taught me to take myself seriously enough to realize my educational ambitions.

Finally, I would like to thank my father, Mr. Harold L. Weaver, for his unwavering support and his apparently limitless patience.

S.H.W.

Dedicated to

Gatha Stinette Weaver

and

Douglas Lee Weaver

ABSTRACT

This paper details the development of a personality measure which differentiates among individuals on the basis of their self-reported comfort with a variety of day to day activities. These activities were systematically pre-selected to be either masculine or feminine in nature. The history of masculinity and femininity measurement is examined, and the trend toward increasingly specific measurement devices traced. In line with an emerging multi-dimensional model of masculinity and femininity, it is suggested that a paper and pencil measure of two separate masculine and feminine dimensions composed of items that are behavioral in terms of content would serve to supplement current M and F measures employing trait items alone.

The examination of male and female stereotypes about the overall pool of potential activity items revealed general agreement between the sexes as to the sex-stereotyping of the items, although some systematic differences between the sexes were observed.

The selection of a group of 31 masculine items and 27 feminine items based on the stereotype data, led to the standardization of the Sex Typed Activities Test. This test was shown to be highly reliable, and the predicted symmetri-

cal pattern for the M and F scores of the two sexes was observed. Correlations to other masculinity and femininity measures were appropriate in sign but mild to moderate in magnitude. Within each sex, distributions of scores were skewed for the sex-congruent scales, the M and F scores were found to be positively correlated, and the factor analysis failed to show the clear-cut two factor structure predicted. As a consequence of these discrepancies with the predicted results, a theoretical model connecting item selection procedures, theoretical definitions, and structural characteristics of dualistic masculinity and femininity measures was developed and is described in terms of a general case.

INTRODUCTION

"Real men," we are informed in a recent paperback, "don't eat quiche." In his "guide to all that is truly masculine," Feirstein (1982) lets us know by example what it means to be a Real Man in an increasingly complex world. This comic approach to sex roles in modern America introduces an important point in a humorous way. Masculinity is certainly thought of in many ways, but one of the most common ways is in terms of action or behavior--what it is that people, especially males in this case, do and what they don't (or won't) do. Femininity may be considered in a similar way, the relative degree evaluated by the things a person does or doesn't do.

When ordinary individuals make this kind of judgment about who is more masculine and who is less masculine they are really making a kind of behavioral evaluation. Whether behavior actually reflects the psychological masculinity and femininity of the person being judged may be an arguable point; it is certainly an empirical question. However, it is indisputable that one's behavior is an important aspect of what most people regard as masculinity or femininity. That behavior is considered an important aspect of masculinity and feminity and one well worth examining is readily

supported by an inspection of the literature concerned with sex-roles and their measurement.

The description of the research which follows does two things. First, it analyzes the attempts at systematic measurement of sex role orientation and reviews the problems which have been encountered in this fifty year old effort. Second, it describes the creation of a behaviorally based paper-and-pencil measure of masculinity and femininity. This measure attempts to incorporate contemporary notions about sex role theory and methodology while employing a more behavioral type of item than the trait-based measures now in widespread use.

In a broader sense this paper concerns the process of measuring masculinity and femininity, which are construed as separate sex-valued psychological dimensions. The aim of masculinity and femininity measurement seems fairly clear: it is to discriminate in a systematic and quantitative manner between individuals who are high in those characteristics from those who are low. This would at first glance seem to be a straightforward task. If we could set forth theoretically precise criteria for a description of the masculine and the "unmasculine" individual we should then be able to create items which measure each individual's self-perception in this regard. For comparison's sake, when self-esteem is measured, observations are made about what qualities go along with liking one's self and what qualities

go along with not liking one's self and a questionnaire is thus developed which tallies these up for a particular individual.

Already, and without recourse to any empirical referents whatsoever we can see that several difficulties emerge when a similar approach is attempted with the measurement of normative sex-linked qualities. Individuals certainly have their own notions about what, and who, is masculine and feminine, and these subjective notions are undoubtedly varied and even idiosyncratic (Myers & Gonda 1982b). In spite of these variations, it is clear that collectively we share a certain core of attitudes which help to define the norms of sex-role behaviors and sex-role characteristics within our particular culture. Unfortunately, the translation of these norms into dimensions that can be measured with psychological tests has proven to be problematical.

First is the problem of separating the "unmasculine" individual (an odd turn of phrase itself) from the feminine individual. Is the man who does eat quiche <u>less</u> masculine or more feminine, or are they the same thing? Recent evidence provided by Storms (1979) suggests that most individuals consider masculinity and femininity globally, as opposite ends of a spectrum. Anecdotal evidence suggests that most individuals may even have difficulty conceptualizing them as separate entities. When the layperson thinks of masculinity and femininity in other individuals it is usual-

ly to contrast one with the other, not to notice just the absence of one. It would probably be fair to say that the real concern, outside of the social psychology literature, is not whether one is "very," "moderately," or "not very" masculine, but rather whether in broad terms one conforms to a particular sex-role that is socially prescribed on the basis of gender.

The naive notion of masculinity/femininity also tends to encompass a variety of aspects or "domains" which are blurred when considered globally. Consequently, a second problem for the investigator of sex roles is that of deciding which aspect of masculinity or femininity is being discussed. This can be confusing because the idea of gender so thoroughly permeates our ways of thinking about the world. In many of the world's languages even inanimate objects are assigned gender. As a result, there are a plethora of possible aspects which can be inspected from a variety of viewpoints. Even when the discussion is limited to psychological or behavioral phenomena, the idea of gender and gender-related qualities extends into a large number of domains. To be more concrete, when femininity is discussed, does the discussion concern mainly what one is, what one does, what one likes, what one wears, or how one acts--which is it? Speaking again of the layperson, for most purposes it probably does not matter. But for the psychologist, the answer will determine the direction taken in the development

of appropriate measures. Given this global approach as a starting point, the task of the behavioral scientist who wants to measure these constructs is made infinitely more difficult, as the complex history of measurement in this area attests.

Early in his discussion of psychometric theory, Nunnally (1978) deplores what he calls the "conglomerate" measure--a measure of a disorganized or haphazard group of individual attributes mixed together on a single test. He suggests that psychological measures should measure only one thing or some "isolatable unitary attribute" (p. 4). In M/F research, the first step in this direction was the breaking down of M and F into two distinct dimensions, measured separately. However, if it is true that masculinity and femininity are not only separate dimensions but indeed consist of multiple dimensions within several different domains of behavior and self-concept, then a process must occur of breaking out the identifiable and theoretically significant components for separate psychometric analysis. In fact, this is exactly what has happened in the area of gender-based psychological attributes.

The general movement of masculinity and femininity research has been in the direction of creating instruments which were more specific and less heterogeneous in item content. Granted, the measures which have resulted are not without their critics. It is a matter of opinion whether

progress in this area should be measured in miles or in inches. Nevertheless, the history of measurement in this area is a fascinating example of the way in which scientific approaches to personality develop. Witness the ceremonial launchings, fits and starts, blind alleys, and generous applause followed by blunt criticism. But in practice, the general trend toward more limited, more precise measures is a theoretically defensible one in spite of shortcomings which particular scales might have. All in all, there does appear to be progress.

It is in this spirit of progress that the current research and the development of the Sex-Typed Activities Test (STAT) was undertaken. Simply stated, the problem was to develop a measure of masculine and feminine dimensions which evaluated individuals with reference to behaviors rather than traits. By attempting to extend the dualistic model of separate masculinity and femininity tests into the behavioral realm, it was hoped that some improved knowledge of psychological masculinity and femininity could be gained. It was also hoped that, if all went well, the capacity of researchers to systematically differentiate between individuals on the basis of their sex-role orientation would be improved to a demonstrable degree.

PREVIOUS RESEARCH IN MASCULINITY - FEMININITY MEASUREMENT

The psychological approaches to the measurement of masculinity and femininity can easily be grouped into two generations. For the most part, the approaches of the first generation operationalized a hypothetical M-F construct which put highly masculine and highly feminine individuals at opposite ends of a single dimension. These "traditional" M-F approaches are consequently described as bipolar and unidimensional. As will be seen, due to the ways in which these tests were created and standardized, they were omnibus measures and their item content was widely varied.

The "contemporary" or second generation approaches, beginning in the 1970's, treated masculinity and femininity as separate hypothetical constructs. They differ as well in terms of item selection, item content, and scope.

Traditional M-F measures

To illustrate the qualities of the first generation of M-F measures it will be useful to briefly compare three selected examples in terms of item selection and content. The three which have been chosen are the Terman-Miles M-F test, the MMPI Mf scale, and the Gough Femininity Scale.

Terman-Miles M-F Test.

Terman and Miles published the report of their extensive research using the Terman-Miles MF Test in 1936, the same year as the publication of the original Strong MF Scale and the Guilford (GAMIN) Masculinity Scale. The Terman-Miles test was a prototype for M-F measurement for many years. The underlying rationale of the test was reflected in the authors' assertion that sex differences in behavior were "so deep seated and pervasive as to lend distinctive character to the entire personality" (Terman & Miles 1936, p. 1). Consequently, the test was developed to measure mental masculinity and femininity and was based upon actual differences in test responses given by male and female groups. There were two alternate forms of the test each of which contained seven exercises. These were: Word association, Ink-blot association, Information, Emotional and Ethical Response, Interests, Personalities and Opinions, and Introvertive Response. The authors were particularly concerned about the breadth of the test and included some exercises with relatively poor reliabilities in order to retain a wide variety of item content areas. Each item was scored in either a feminine direction (-1), or a masculine direction (+1). In spite of differences within and between the subtests, each individual was assigned a single M-F score on the basis of his or her responses to the entire test battery.

Gough Femininity Scale.

The Femininity Scale in its initial form was a 58-item test with self-descriptive true-false items. High scores indicated psychological femininity. Items were selected on the basis of whether or not they differentiated males from females in two samples--one of high school students in Wisconsin, and one of college students in California. One of the objectives in the creation of this scale was to devise a measure that was as non-obvious as possible. The content of the items according to Gough (1952) reflects a number of clusters relating to femininity: work interests; an interest in feminine roles and rejection of masculine roles; feelings of sensitivity; social timidity and lack of confidence; a sense of compassion and sympathy; and, an apolitical attitude. It even includes a pair of items dealing with physiology (e.g., "I am hardly ever bothered by a skin condition, such as athlete's foot, rash, etc." which is scored for femininity when answered 'True') (Gough 1952).

MMPI Mf Scale.

In the context of the MMPI, a clinical instrument, the central purpose of the Mf test is actually to distinguish homosexuals or "sexual inverts" from heterosexuals. The 60 items were initially drawn from a pool which discriminated between male and female respondents. These items were then administered to a small group of homosexual males, and the

final item selection was based on the ability of items to discriminate between this group and heterosexual males. Actually, as a test of homosexuality, it has never been very effective in distinguishing lesbian females from other females, although it is more effective with males. This is a true-false scale with self-descriptive statements, many of which (23) derived from the work of Terman and Miles. The direction of scores depends on the sex of the individual respondent with high scores representing deviance from the norm of male masculinity and female femininity. According to Dahlstrom and Welsh (1960, p. 64), the content areas of the Mf scale included five separate dimensions: ego sensitivity, sexual identification, altruism, endorsement of culturally feminine occupations, and denial of culturally masculine occupations. Obviously, the construct validation of the MMPI Mf scale is somewhat different from that of the other MF scales since its underlying rationale incorporates the notion of homosexuality in its definition (cf. Constantinople 1973).

One of the striking features of these examples is the emphasis on breadth. Each of these tests samples widely from different content realms, with a limited amount of overlap between the different measures. The common rationale for the item-selection procedures was that M-F scales should differentiate between males and females--the assumption be-

ing that however that was done, one would thereby be choosing items that differentiated between masculine and feminine individuals. The haphazard way in which items were selected and their inevitable variability across dissimilar content areas immediately suggests that they should suffer from relatively poor psychometric properties. These might include weak reliabilities, complex or multidimensional factor structure, and poor inter-item and interscale correlations. Although the nature of these problems varied from scale to scale, there were in fact multiple difficulties with these scales (Constantinople 1973; Lunneborg 1972; Lunneborg & Lunneborg 1970).

Criticisms of the M-F model

A decade ago many of the fundamental assumptions of M-F research were widely re-examined. The review of the M-F literature by Constantinople in 1973 was pivotal in the transformation of psychometric approaches from the first to the second generation. The thoughtful analysis of this article left little room for the defense of the traditional M-F approach. Basically, it posed two questions which can be paraphrased: "Are we going about measuring masculinity and femininity in the right way?" and "Should the attempt to measure masculinity and femininity be made at all?" The answer to this second question remains to be seen; however, the answer to the first question was an emphatic "no". The

M-F approach was seen to suffer from multiple difficulties: a lack of theoretical definition, a poor ability to systematically predict other variables, a demonstrated absence of presumed unidimensionality, and a lack of support for the assumptions on which it was based.

The lack of definition was apparent as Constantinople attempted to put together a definition which could describe Masculinity-Femininity:

> The most generalized definitions of the terms as they are used by those developing tests of M-F would seem to be that they are relatively enduring traits which are more or less rooted in anatomy, physiology, and early experience, and which generally serve to distinguish males from females in appearance, attitudes, and behavior (Constantinople 1973, p. 390).

Drawing a comparison with intelligence, she notes:

> In both cases, we are dealing with an abstract concept that seems to summarize some dimension of reality important for many people, but we are hard pressed as scientists to come up with any clear definition of the concept or indeed any unexceptionable criteria for its measurement. (Constantinople 1973, p. 390)

The consequent use of sex differences as a basis for item inclusion on M-F scales leads to varied notions of what M-F really is depending on the scale used. Constantinople alleged that the heterogeneity of content within, as well as across, M-F measures was a direct result of the lack of an adequate definition of the construct. Even though many items were included which could indeed be intuitively related to a personality construct as such, many other items were

included which had no particular connection to a theory of sex roles in personality.

The single most compelling of Constantinople's observations about M-F and M-F measures was undoubtedly her assertion that it was a major error to consider M-F as a single bipolar construct. The scoring patterns of the M-F tests implied a logical reversal such that what was not masculine was automatically labelled feminine. Built into the rationale of the traditional M-F measures was the notion that masculinity and femininity were opposites by definition. They should therefore be measured in such a way that highly feminine scores would fall on one end of the scale and highly masculine scores would fall on the other.

This was, as Constantinople pointed out, an untested hypothesis treated as an assumption. Moreover, it required that a continuous scale be formed of items selected by their ability to discriminate between the sexes, in spite of the fact that gender is clearly not a continuous variable. Carlson (1971) was also critical of one-dimensional approaches, stating that they failed to reflect adequately the nature of psychosexuality; she both advocated and employed a qualitative, typological approach instead, based on Bakan's (1966) agency/communion dichotomy. A third critic of the bipolar-unidimensional approach, Bem (1974) pointed out that where M-F was represented on a single dimension, it was not conceptually clear what a mid-range score implied in terms of the sex-role of the particular individual.

Finally, the traditional measures displayed a striking absence of the unidimensionality implied by the theoretical M-F construct they were designed to represent. The fundamental assumption involved in creating a scale out of a group of items is that the items, to a greater or lesser degree, measure the same thing. The M-F tests demonstrated poor, that is to say complex, dimensionality as a rule when factor analyzed (Constantinople 1973; Lunneborg 1972). This appears in retrospect to be due, not only to the attempt of the developers to sample a large number of content domains, but also to the tradition of combining masculine and feminine items into a single scale. Constantinople found the practice of using a single score to be inadequate and unjustified. She advanced the idea that M-F might be construed as two separate dimensions, or even as a multi-dimensional phenomenon. If M-F were multidimensional, she reasoned, a profile of scores based on subtraits would be preferable to a single score which fails to account for individual variation. To summarize, in Constantinople's view, the M-F measures lacked a theoretical definition, assumed that M and F were always opposites, used a multiplicity of types of items drawn from varied domains, and displayed poor psychometric properties (perhaps as a consequence of these other problems). Is it any surprise then that no consistent pattern of relationships between M-F and other variables could be said to have emerged? In any case, observed rela-

tionships to other variables must now be viewed in retrospect with extreme caution. For, as Lunneborg (1972) has indicated, it is not exactly clear what it was that the M-F measures were in fact measuring.

As useful as masculinity and femininity may be in everyday life as explanatory devices, the adaptation of them for psychological purposes proved elusive. The need arose for a radical re-examination of the methodologies employed to measure these constructs. If the constructs were to be of use, the approach to measurement had to be refined. Consequently, the second generation of M and F measures were created partly to correct a number of the shortcomings of the first generation approaches, and partly to allow researchers to examine a new construct which they called androgyny.

Second-generation approaches to M & F measurement

The masculinity and femininity measures of the contemporary generation differ from the traditional M-F measures as described above in at least two important ways. First, the newer measures tend to treat masculinity and femininity as separate constructs. Second, they tend to restrict item content to a single theoretical domain. In order to clarify and provide a reference point for the discussion that follows, a brief review of these measures is in order.

The Bem Sex Role Inventory (BSRI)

In 1974, Sandra Bem described the creation of the Bem Sex Role Inventory (BSRI) which treated M and F as separate dimensions. It asked respondents to rate themselves with regard to a series of adjectival descriptors using a Likert scale ranging from 1 (Never or almost never true) to 7 (Always or almost always true). Examples of the items on this scale, as well as the other scales discussed in this section are given in Table 1. Mean scores for each subject were computed for the ratings on 20 Masculine items, 20 Feminine items, and 20 Neutral items. The items of the M and F scales were designed to be positive traits, so that femininity would not be simply the absence of masculinity, or the opposite of masculinity, but a positive dimension in its own right. Items on the Neutral scale were both positive and negative but were added chiefly to obscure the purpose of the test.

To select the items for the test, groups of males and females rated a large pool of potential items. Each item was rated in terms of its perceived desirability in American society for a male or for a female. Item selection was contingent upon agreement about the stereotype by raters of both sexes in several independent samples. If a characteristic was more desirable for a male than for a female it was eligible for inclusion on the M scale. If it was more desirable for a female than a male it was eligible for inclu-

sion on the Femininity scale. As Bem pointed out, the use of sex-typed social desirability of items as a basis for item selection differentiates this scale from the earlier scales which used differential endorsement by males and females as the criterion for inclusion (Bem 1974). This was a major innovation in itself inasmuch as it involved the use of general stereotypes about males and females for the purpose of measuring masculinity and femininity.

Personal Attributes Questionnaire (PAQ)

At about the same time that Bem developed the BSRI, Spence, Helmreich and Stapp (1975) described the creation of a similar scale called the Personal Attributes Questionnaire (PAQ). The PAQ produces three scores for an individual: Masculine (M), Feminine (F), and Masculine-Feminine (M-F). This test combines separate masculinity and femininity scales with a third bipolar M-F scale. The PAQ, like the BSRI, is limited to traits with a positive valence, although they are expressed in a a bipolar manner, i.e., "Very gentle" is counterposed to "Very rough", and scored on the F scale for degree of gentleness. The PAQ uses a 5-point Likert scale, and the scores are sums of self-ratings on the items.

Item selection on the PAQ is somewhat more difficult to describe in a few words. Again, the items themselves are bipolar in nature with adjectives describing the two end-

TABLE 1

Examples of Items on Four Tests

Bem Sex Role Inventory (Bem 1974)

Masculine	Feminine	Neutral
Self-reliant	Cheerful	Adaptable (+)
Defends own beliefs	Yielding	Jealous (-)
Competitive	Shy	Sincere (+)

Personal Attributes Questionnaire (Spence, Helmreich, & Stapp 1975)

Masculine
Not at all independent a...b...c...d...e Very independent
Feminine
Not at all emotional a...b...c...d...e Very emotional
Masculine-Feminine
Not at all aggressive a...b...c...d...e Very aggressive

PRF ANDRO Scale (Berzins, Welling & Wetter 1978)

MASCUL:
I try to control others rather than permit them to control me. (scored 1 if True)

FEMIN:
I like to be with people who assume a protective attitude toward me. (Scored 1 if True)

Adjective Check list M and F subscales (Heilbrun 1976)

Masculinity	Femininity
aggressive	appreciative
arrogant	excitable
hard-headed	frivolous
outspoken	praising

---

points. (See Table 1 for examples.) In this form, ratings can be made on the items for any particular target (e.g., the typical male, the ideal female, or the self). The 55 original PAQ items were drawn from a larger pool on the basis of raters' judgments about the typical male and female. In this case, the same raters made a separate set of ratings for the typical male and the typical female. Difference scores were computed for each item and average difference scores were tested against the null hypothesis. Where the null hypothesis was rejected for both male and female raters, a significant difference in sex stereotype was demonstrated and the item was retained for the PAQ. In other words, those items were selected which showed significant differences between ratings for the typical male and the typical female. This is not the same of course as actual differences between males and females but again represents different ratings for male and female stereotypes.

These 55 items were assigned to subscales by virtue of a second set of judges' ratings. This time, a separate group of raters were given the same items and asked to make ratings for either the ideal male or the ideal female. Since these were positive characteristics, the ratings for ideal males and ideal females generally fell on the same side of the middle of the scale. To use our example, an item was used for which one endpoint was labelled "Gentle" and the other labelled "Rough". One group of raters rated

the "typical male" and the "typical female" on this item, and a difference in sex stereotype was found. The typical female was judged significantly more gentle than the typical male by both males and females. Another set of raters made ratings for the ideal male and the ideal female. In this case--even though the two sexes were seen as typically different on this trait-- the ideal individual of either sex was seen as more gentle than rough. The ratings for the ideal male and the ideal female fell on the same side of the rating-scale's midpoint. Consequently, this item was placed on the F scale, as a trait desirable for either sex but more typical of females than of males. If a trait were desirable for either sex, but more typical of males, then it was assigned to the M scale instead.

In some cases however average ratings given to the ideal male and the ideal female fell on opposite sides of the midpoint or toward opposing characteristics. For example, "very home oriented" versus "very worldly" was such an item. These items were assigned to the bipolar M-F or "sex-specific" subscale. Subsequently, the 55 item version was reduced to a 24 item version with eight items on each subscale, making this one of the briefest and easiest to administer of all of these inventories.

Personality Research Form ANDRO Scales

The PRF ANDRO scale is a pair of scales composed of items taken from the Personality Research Form (Jackson 1967). It employs a true-false format and contains 29 items on the M scale and 27 items on the F scale. This pair of subscales was not developed independently but was consciously modeled on the BSRI scales. The subscales were designed to provide independent measures of M and F, and to this end the item selection procedure for the BSRI was imitated in two ways: items were chosen which had positive content; and items were selected on the basis of sex-typed desirability. The authors desired that the scales would correlate with the appropriate BSRI subscales, and in fact they did achieve modest correlations. The PRF ANDRO scales, like the next two scales to be described are derivative measures from larger inventories of broader scope. They are less widely used or validated than the two popular measures described above, the PAQ and BSRI.

Adjective Check List M and F subscales

Consistent with the general movement toward assessing masculinity and femininity separately, Heilbrun (1976) chose to revise his traditional style M-F scale taken from the Adjective Check List (Cosentino & Heilbrun 1964). This resulted in the creation of a 28 item M scale and a 26 item F scale.

Item selection proceeded from data which had been previously collected. The original M-F scale was built by choosing those adjectives that discriminated between male college students who were identified with masculine fathers and female college students who were identified with feminine mothers. The resulting list was easily broken down into masculine and feminine adjectives which were then treated as separate M and F subscales. Scores on these newly created scales tended to be somewhat less orthogonal than the PAQ or the BSRI subscales. More importantly, the rationale for item selection is very different, and raises doubts about the comparability of these subscales with other measures of M and F such as the BSRI.

CPI Msc and Fmn subscales

Baucom (1976) described the development of separate masculinity (Msc) and femininity (Fmn) subscales on the California Psychological Inventory (CPI). CPI items are, again, true-false items that are self-descriptive. For an item to be included on the Msc scale it was required that it be endorsed in a given direction by 70% of a male sample, and by at least 10% fewer females. The Fmn subscale was developed in the analogous way, reversing the sexes. Clearly this is a departure from the use of stereotypes for item selection, back to the use of observed sex differences. The implications for comparability with the PAQ or the BSRI consequently are again unclear.

Androgyny Theory

Having described this group of dualistic or second generation measures of masculinity and femininity, a more thorough examination of the theoretical and methodological aspects of this approach can be presented. The last three measures which were just discussed (respectively, the PRF ANDRO; ACL M AND F; and the CPI Msc and Fmn scales) are considerably less prominent in the literature than the PAQ and the BSRI. Also, they are related in terms of their assumptions, methods, and definitions to the more widely-cited measures. Consequently, throughout the remainder of this discussion, the central focus will be the two major inventories, the Personal Attributes Questionnaire (PAQ) and the Bem Sex Role Inventory (BSRI).

Turning to the matter of the theory behind the creation of these two tests, it will be noted that one of the central criticisms leveled at the traditional M-F measures was that they lacked a theoretical definition of the underlying construct they supposedly measured. This seemed to imply two things.

1. Item content was a conglomerate of diverse areas assembled without a rationale other than the fact that items differentiated males from females to a significant degree.
2. It was difficult to validate the scales in any systematic fashion.

The authors of the BSRI and the PAQ asserted that in contradistinction to the traditional M-F scales, the new instruments were in fact based on a theoretical viewpoint. This is a point of some contention, as we shall see.

In relating the contents of these two alternative M and F inventories to theoretical notions about sex-linked personality variables, both Bem and Spence and Helmreich appeal to earlier theoretical explanations of masculine and feminine dimensions. Bakan (1966) described two modalities: agency and communion, the first expressing self-interest and self-assertion, and the second expressing concern for others and altruism. These are respectively identified with male principles and female principles. Parsons and Bales (1955) drew a distinction between the instrumental and expressive roles played by men and women with regard to the family. Instrumental orientation concerns the achievement of goals, the male role, while expressive orientation concerns keeping the family together and harmonious. The content of the M scales are therefore said to represent either agentic or instrumental characteristics of personality, while the F scales are said to reflect expressive or communal concerns or characteristics.

The BSRI and PAQ are often referred to as "androgyny scales" which leads to the supposition that they measure some "thing" called androgyny. In fact they don't; they measure, more or less adequately, and more or less accurate-

ly, constructs called masculinity and femininity through the medium of positive adjectival self-descriptors. Androgyny is a hypothetical construct derived from knowledge of an individual's relative levels of masculinity and femininity.

The term androgyny was coined by Bem (1974) who, as a researcher and feminist was primarily interested in the differences between sex-typed (masculine males and feminine females) and androgynous individuals. From Bem's point of view, masculinity and femininity are different groups of positive attributes which both males and females have the capacity to internalize. Androgynous individuals are those who possess a more or less equal blending of masculine and feminine characteristics. In developing the BSRI she claimed to be attempting to construct an instrument which could be used by researchers to distinguish between sex-typed and androgynous individuals. Incidentally, Bem has repeatedly stated that her chief interest is in androgyny, and her research program has never concentrated on M and F as such (c.f. Bem 1974; 1979).

As characterized by Bem, the traditional viewpoint, both within and outside of psychology, held that to be a masculine male or a feminine female implied optimal adjustment. She argued for a radical re-examination of the idea that being sex-typed was preferable for individual adjustment. Androgyny theory puts forth an alternative hypothesis: that it is preferable in terms of adjustment to have

some balance between masculinity and femininity regardless of sex. The reasoning behind this is that the androgynous person lacks the conflicts about sex role that the sex-typed person has, and consequently can be more flexible over the variety of situations that are encountered in a day. Kelly and Worell (1977) dubbed this the 'response repertoire model', that is to say, that the androgynous person as compared to the sex typed person, can draw on a greater variety of responses, employing masculine assertiveness as necessary, but also being equally capable of feminine nurturance or empathy when those qualities are called for. In sum, the androgynous person combines both masculine and feminine characteristics, is expected to show a greater degree of behavioral flexibility, and is consequently expected to be better adjusted. This theory was presented in the context of an unabashedly pro-feminist perspective which sometimes gives Bem's writing a polemical air.

The theoretical rationale of the PAQ is somewhat less tied to a political point of view. Instead, it derives from the application of previous research in sex role stereotypes to the problems of personality measurement. Essentially, Spence and Helmreich (1978) borrowed from the work of a group of researchers who were interested in looking at sex role stereotypes as such and used their collection of stereotype descriptions as a basis for comparing persons. The original pool of items from which the PAQ items were drawn

was based upon characteristics listed on the Sex-Role Questionnaire developed by Rosenkrantz, the Brovermans, and their colleagues (see Broverman, Vogel, Broverman, Clarkson, & Rosenkrantz 1972). The work of these researchers was aimed at examining a large group of items and thereby examining in an objective fashion the prevalence and pervasiveness of sex role stereotypes. The results of these studies justify a brief digression. First, the initial study indicated the widespread existence of stereotyped notions about males and females, as well as the strong agreement between the sexes about these stereotypes (Rosenkrantz, Vogel, Bee, Broverman & Broverman 1968). It was also found that the characteristics conventionally associated with males were more highly valued than those associated with females. Finally, the self-ratings of males and females actually paralleled the stereotype ratings, suggesting their veridicality.

The PAQ, which asks respondents to rate themselves on items which have been determined to have stereotypic characteristics, is based on a logical extrapolation from stereotype ratings to personality measurement. It might be said that the PAQ represents a conscious attempt to use stereotypic notions about sex roles as a vehicle for assessing the self-concepts of individuals, insofar as masculinity and femininity are concerned.

The second generation measures, then, are purported to stand on firmer theoretical ground. To reinforce this

claim, the advocates of these approaches pointed to several advances over the traditional M-F approaches. The first, and most salient advance is the abandonment of the assumption of bipolarity. These scales both incorporate the idea that masculinity and femininity should be be approached separately. The second contribution made by the authors of these scales in theoretical terms is the restriction of item content to a single domain, in this case, self-descriptive adjective phrases. Although this was not seen to have preeminent theoretical significance, it marks a refinement of measurement technique that is not adequately appreciated. Third, instead of using sex differences for item selection, these scales employ collective stereotypes about sex roles as an instrument for item selection. This employment of stereotypes divorces the inquiry into sex role orientation from the study of sex differences: i.e., it separates sex and sex role into distinct categories. Finally, some attempt is made to rationalize the scales in terms of theoretical dimensions which are not mutually exclusive or diametrically opposed.

Androgyny: Measurement and Methodology

The separation of masculine and feminine dimensions was a step in the right direction but it raised new questions about the relationship between masculinity and femininity. The cultural assumption that M and F are diametrically opposed continued to have an impact upon the thinking of researchers. This is revealed in the ways in which they set up their inquiries as well as in the ways they analyzed and interpreted their results.

For example, Bem's (1974) initial approach to androgyny was to divide respondents into masculine, feminine, and androgynous subjects based on how close their scores on the M and F scales were. When the M and F scores for a particular individual are combined in a subtractive fashion, he or she is assigned a single score which may range from high masculine to balanced to high feminine. In essence, this converts the dualistic measure into a unidimensional, bipolar measure by distributing individuals along a dimension from very masculine to very feminine. Even though this approach was subsequently renounced by Bem (1977), other researchers have continued to repeat the methodology.

On the PAQ, there are three scales. Spence and Helmreich found it necessary to incorporate a third bipolar scale because on some of their bipolar items the desirability of the two poles differed for the two genders. Inspection of these items reveals that the use of bipolar adjec-

tives incorporates the bipolar notion of M and F into the items themselves. The criticism of this is the same as the criticism of bipolar tests. Given a five point scale counterposing "Very home oriented" to "Very worldly", where does the person who considers herself both worldly and home oriented put her self-rating? In other words, it continues the tradition of bipolar measurement where middle scores are indeterminate in their interpretation (cf. Storms 1979).

In defense of the third subscale of the PAQ, Spence and Helmreich note:

> Since additional analyses convinced us that the M-F scale was not a psychometric accident and since we suspected that it might yield significant information not available from the other scales, we have retained it, despite the conceptual embarrassment of having to embrace simultaneously a dualistic and a bipolar model of masculinity and femininity. (Spence & Helmreich 1978, p. 20)

Since there is in fact no substantive theory of masculinity and femininity measurement at present to describe what the "actual" dimensions of the underlying constructs are, this empirical justification may have to serve.

Spence, Helmreich and Stapp (1975) disagreed with Bem on the meaning of androgyny, which they defined by placing subjects in a two by two typology: Androgynous (High M, High F); Masculine (High M, Low F); Feminine(Low M, High F); and, Undifferentiated (Low M, Low F). The dividing lines were the medians for both sexes weighted for differences in group size. In examining the relationships between sex-role

group and self-esteem they found that the highest self-esteem scores were in the Androgynous group, followed closely by the Masculine individuals, and then by Feminine and Undifferentiated individuals. This analysis, based on a one-way ANOVA model set the pattern for a number of subsequent research efforts. Bem (1977) revised her own scheme, stating that since there appeared to be empirical differences in self-esteem scores between the two "balanced" groups--androgynous and undifferentiated--the four group model seemed to have validity.

One of the problems with the patterns set by these researchers was that it was relatively difficult to ascertain from their reports, and those of researchers who followed their models of data analysis, what the individual effects and the interactions of the M and F scales were. Although the typology described by these authors fit neatly into a 2 X 2 arrangement, Spence and Helmreich used a one-way analysis of variance over four groups. It was not until 1982 that a statement appeared pointing out that the approaches to data analysis that had been used simply overlooked the fact that, whether treated as typological or continuous variables, M and F were separate variables which could be analyzed in a standard 2 x 2 factorial design with an interaction term (Taylor & Hall 1982). This report goes on to point out that as far as "androgyny" was concerned, there were two separate hypotheses. With regard to any dependent

variable, the main effects hypothesis would predict that M and F would affect the dependent variable in an additive fashion. This resembles in conception the arguments that have been advanced by Spence, Helmreich and their co-workers. The interaction hypothesis, in contrast, suggests that the two variables will conjointly have an effect over and above what might be expected strictly from the addition of the separate effects of masculinity and femininity. This seems to reflect Bem's point of view.

The persistent use of seemingly inappropriate models for the analysis of masculinity and femininity data is less remarkable when it is remembered that the traditional underlying model for sex roles is a bipolar one. Despite the psychometric separation of M and F, researchers still have difficulty in thinking about one without immediate, and often confounding, reference to the other. Only gradually have the implications of considering these constructs as completely separate become clear. Simple as these points are, they have tended to clarify--or rather demystify--a number of confusing elements in the androgyny literature.

Limitations of the second generation M & F measures

We have examined in some detail the creation and rationale behind the two most commonly used methods for measuring masculinity and femininity, namely, the Bem Sex Role Inventory and the Personal Attributes Questionnaire. An attempt has been made to show why they were created, what makes the androgyny approach different, and how the approach to data analysis has evolved. A complete review and analysis of the empirical and experimental literature that has been generated around the topic of androgyny is beyond the scope of the current discussion. Instead, this discussion will be limited to those articles which are most directly concerned with the questions of methodology and validity in the measurement of the overall masculinity and femininity constructs. Interested readers will find a substantial recent review in Taylor and Hall (1982) which takes a broader integrative view of the research done so far.

In considering the creation of an alternative measure for M and F, it will be of use to examine with some care the criticisms of the existing measures and their implications for the model of M and F measurement. These criticisms are of two varieties: structural imperfections of the measures, especially the BSRI; and construct validity of the measures.

Structural imperfections

Pedhazur and Tetenbaum (1979) examined in detail the rationale and item selection procedure for the Bem Sex Role Inventory. Note was already made that the claims of enhanced theoretical adequacy for the second generation measures were later disputed. Pedhazur and Tetenbaum were highly critical of what they have characterized as the lack of a theoretical rationale for the BSRI. They set out to examine the structure of the BSRI on the basis of Bem's description of her intentions and her item selection procedures. Even so, they found the BSRI lacking in a variety of ways. They performed a number of analyses on the items of the BSRI with large samples of graduate students in education in the New York City area. Based on the hypothesis initially adopted by Bem that M and F would be two orthogonal dimensions made up of positive attributes, Pedhazur and Tetenbaum predicted that factor analysis of the Bem items should reveal a demonstrable two factor structure primarily reflecting masculine and feminine content. As they point out, the creation of the summative scales or tests presupposes that the items on them are related on a single dimension--i.e., all the items should relate to a common core of meaning. Consequently, the factor structure should be recognizable as Masculinity and Femininity.

Pedhazur and Tetenbaum collected two groups of ratings. The first set were stereotype ratings on the 60 M, F, and

Neutral BSRI items, and the second were self-ratings on the 40 BSRI M and F items only. They analyzed these ratings in each of three different ways: an item analysis which focused on descriptive statistics; a discriminant analysis; and a factor analysis.

The discriminant analyses served chiefly to show that in terms of predicting sex of target for stereotype raters, and in terms of predicting sex of self-raters, the two items Masculine and Feminine contributed the largest share of the variance. This means that the sex-typed social desirability of these two items far outstrips that of the other 38 items.

The item analyses revealed that some of the items which Bem had found to differ in terms of their sex-typed desirability ratings had been found by other researchers to be neutral in terms of sex, and that some of the neutral items from the BSRI had been considered by others to be sex-typed. In addition, some of the "positive" feminine traits (e.g. Gullible, Childlike) had lower desirability ratings than some of the "negative" neutral traits!

Stereotype ratings. The factor analyses of stereotype ratings were conducted separately for male and female raters, but the results were sufficiently similar to justify pooling the data for the two sexes. Subjects were asked to rate the desirability of each trait for a man, for a woman, or for an adult in American society. Consequently, there were three separate factor analyses for the three different

stereotypes. Taken together, these analyses seemed to indicate a different factor structure than the hypothesized two factor structure. The first factor, appearing in slightly different forms for the three targets was labelled "Interpersonal Sensitivity". It included items from both the Femininity scale and the Neutral scale. A second factor containing almost all of the M items was labelled "Assertiveness" (significantly, the item Masculine failed to load on on this factor). The third factor included a few Feminine scale traits such as Gullible, Childlike, Shy (for a woman only), and Flatterable (for a man only), as well as a number of the negative Neutral traits. This factor was termed "Immaturity".

Two different positions may be taken on the inclusion of neutral items within this factor analysis. As Bem points out, the neutral items are really used only as filler and are not used in the computation of scores. As a result, the type of common variance that they have with items from the two main scales may tend to create a factor structure which is irrelevant to the application of the scales themselves. On the other hand, the neutral scale is adminstered routinely and the factor structure of the entire test does not reflect the M/F/Neutral division, suggesting a re-examination of the division of these items into subscales. Actually the observed factor structure does suggest a division between M and F factors giving some support to the BSRI as a measure

of Assertive traits (Masculinity?) versus traits reflecting Interpersonal sensitivity (Femininity?). The fact that all of the femininity items did not load on the Interpersonal Sensitivity dimension does not detract from the fact that many did, and that this particular factor is separated from the factor on which masculine items loaded. It would clearly be desirable to alter the test and delete certain items, but the results are are not sufficiently damaging to justify wholesale rejection of the test.

Self-ratings. It is unfortunate, but when the self-ratings were made for this study, the neutral items were omitted. The subjects in this second study were 171 male and 400 female graduate students of education. Comparisons of the mean self-ratings for each of the 40 items disclosed that with the exception of three items the mean differences between male and female respondents were relatively small. In terms of total scores, about half of the actual differences between male and female mean scores on the overall BSRI M and F scales were due solely to the items Masculine and Feminine.

The factor analyses for the self-ratings were carried out separately for males and females. In both the male and female cases, a four factor solution resulted. The factor structures were somewhat dissimilar which invites the question of whether it is justifiable to use the same scales with both males and females and to treat the scores as

equivalent. In both cases, the fourth factor was fairly similar: a bipolar factor with masculinity loading positively and femininity loading negatively. This again is evidence that these items differ significantly from the other items on these subscales.

Female self-ratings produced one main factor on which almost all the M items loaded, but the male self-ratings produced two M factors. For females, the first factor included meaningful loadings for 17 of the 20 M traits (two more traits did not make the .40 cut-off, but still loaded above .30). In comparison, M traits loaded on two separate factors for male respondents. One represented "Independence", and the other represented "Self-sufficiency". It would appear that males make more refined distinctions between types of masculine traits than do females.

In contrast, the male respondents seemed to view all of the feminine items as related, while females made more distinctions among them. Eleven of the F traits loaded above .40 on the first of the males' factors (12 greater than .30). The females' second factor also had 12 F traits loaded on it. The third factor for females, however, seemed to be bipolar. It contrasted the negatively signed F traits like Childlike and Gullible, with the M traits Independent, Self-sufficient, and Self-reliant. This seems to suggest that females discriminate between those F traits which deal with "Interpersonal sensitivity" from these other F traits

which we have already seen to be poor items on the basis of their lack of positive social desirability.

Pedhazur and Tetenbaum summarize these analyses in the following terms: "the factor analyses of self ratings for males and females do not reflect the dimensions of masculinity and femininity proposed by Bem." (Pedhazur & Tetenbaum 1979, p. 1012) This conclusion is based on the pattern of the observed factors which differed between males and females. But once again, as with the stereotype ratings, there is a general separation of masculine and feminine traits on separate factors. It must be granted that the two hypothesized major factors did not appear in an unqualified fashion. But again, these data provide room for the interpretation that although there is a need for revision, there is also a basis for the validity of this approach. The data themselves are sufficiently equivocal, in my own view, to allow for alternative interpretations.

These factor analytic data do show at least two important things. First, the items Masculinity and Femininity, whatever their worth in measuring the overall constructs, don't belong on these scales. They are rather too tightly bound to sex of respondent with the responses of individuals reflecting their biological sex more than their psychological sex role. Second, several items on the femininity scale are not very desirable in absolute terms although they may be <u>less undesirable</u> for females than for males. Clearly a

revision of the BSRI is indicated based on these factor analytic data. Such a revision is forthcoming according to Bem (1979).

These data tend to indicate that the self-ratings are not as cohesive within scales as they might be. In practical terms however, they may be sufficiently cohesive to justify their inclusion on a single scale if revised.

Other criticisms of methodology. In addition to the Pedhazur and Tetenbaum study, several other attempts have been made to re-examine the items of the BSRI in terms of their sex-typed social desirability. Bem's premise was that any group of judges could serve as informants for the universal masculine and feminine stereotypes. However, Edwards and Ashworth (1977) found very little replication of Bem's original findings regarding the stereotypy ratings of the individual BSRI items. As Bem (1979) has pointed out, however, this failure was not in the replication of results but in the replication of instructions to the raters, making the two studies non-equivalent. Subsequently, another study at the same university (Walkup & Abbott 1978) found that all but three of the original BSRI items were in fact judged to be differentially desirable for males and females by both male and female judges. The other three were judged significantly different but only by female raters, indicating a trend, but an incomplete one in terms of Bem's original item-selection criteria, which required agreement by male and female judges.

Myers and Gonda (1982a) have reported the results of two studies on a preselected subset of BSRI items. Subjects in the first study were visitors to a participatory museum in Toronto, mostly non-students, and about half were from the U.S. The implications of this study for the validity of the BSRI are difficult to discern, although the tone of the article is very critical--taking exception not only to the particular items of the scales but to the entire trait approach and even the use of Likert scales in research.

In the first study, subjects were given a small (11 item) subset of the BSRI and asked to make no less than eight separate sets of ratings and one open-ended response for each of those items. The central concern of the study was whether the same kinds of sex-role stereotypes would be reflected across all of these various rating conditions. However, with regard to the crucial concern of sex-typing or stereotyping, the stability of the ratings was extremely variable across all conditions. This led the authors to assert that basing a self-rating scale on sex-stereotyping may be dangerous because the way in which the stereotypes are procured or elicited may influence what is found. The second study asked subjects to give open-ended descriptions of themselves in different specified situations, to fill out the BSRI, to define "aggressive" (which is a BSRI item), and to give examples of situations in which they might be aggressive.

Although this report must be mentioned for reasons of completeness, it is not clear, to be frank, what the implications of the research are for the measurement of M and F. As is immediately apparent, the analyses of these multiple requests of subjects were complex and the focus of the study diffuse. The complicated methodology of the first study especially, encourages one to wonder at what point repeated ratings of the same stimuli begin to take on a more random, less systematic, aspect. The more germane question of whether BSRI scores accurately discriminate between masculine and non-masculine or feminine and non-feminine individuals within a given sex does not appear to be at issue here. That, it seems, is really the important issue. The focus here seems to be on the validity of the items as representative of stereotypes. It seems that the authors feel that an absence of unequivocal evidence about the stereotypical features of the BSRI items under certain instruction conditions is an adequate criticism of the scale itself. But the absence of evidence is not the evidence of absence. Simply because sex stereotyping does not influence every possible rating pattern does not negate the fact that the items do contain an element of sex stereotyping. Because of their logical and methodological complexity and their equivocal implications for M and F measurement, these two studies will not be discussed further.

In contrast to the BSRI, the factor structure for the PAQ as reported by Helmreich, Spence and Wilhelm (1981) appears to be largely in line with the expectation that M and F subscale items will load on separate factors and that a two factor solution will be the most adequate explanation for a large portion of the observed variance. In this regard, the PAQ may be a more uniform and cohesive set of scales than the BSRI.

Summary. It is clear then that a certain amount of criticism was due to the BSRI due to the inadequate attention which had been paid to the initial scale construction. Factor analysis, if included in the original construction could have avoided some of the difficulties and objections which were outlined by Pedhazur and Tetenbaum (1979). At the same time, I have argued that the evidence provided by Pedhazur and Tetenbaum as well as the evidence presented by Helmreich, Spence and Wilhelm (1981) gives support to the construction of dualistic measures of M and F based on stereotyped traits.

The Construct Validity Issue

Concerns about construct validity have been the other central focus of criticism addressed toward the current measures of M and F. The first of these concerns relates to the problem of adequate definitions for masculinity and femininity. The second revolves around the desirability of us-

ing stereotypes about males and females for the purpose of creating self-rating scales. The final aspect of construct validity considered here involves differences in interpretation between Bem and Spence and Helmreich regarding the narrowness or breadth of the M and F constructs. Prior to considering these three points, however, a brief discussion of validity in psychological measurement is offered to clarify the use of terminology.

Types of validity. To briefly review, there are a number of different subtypes of validity which are commonly discussed with regard to psychological measures. "Face validity" implies that what is being measured is clear from the content and structure of the test. "Concurrent" or "convergent" validity refers to the fact that a measure correlates well with other measures of the same construct. "Discriminant" validity suggests that a measure does not relate too closely to measures of other constructs, which might imply that it is measuring something other than what the investigator intends (Campbell & Fiske, 1959).

Nunnally (1978) describes three major types of validity. The first of these, "predictive" validity, concerns the ability of a score to accurately predict some quality or some behavior for which some external criterion exists. For example, do SAT scores predict level of success in college? If so, the SAT is said to have predictive validity. At this stage of investigation, if such a criterion exists at all for sex-role orientation it is obscure.

A second type of validity asks whether a measure includes items from all the possible areas which the investigator wishes to survey. "Content" validity is desired whenever there is no criterion but when the instrument itself is the object of the measurement. This is most common in in-class exams where the goal is to adequately sample the knowledge students have gained about the material covered in the class.

The third type of validity, "construct validity", is the most relevant in the present context. It is also more difficult to pin down than the other types. The terms "construct" or "hypothetical construct" refer to abstract notions about psychological variables of interest. Consequently, construct validity for a measure indicates that the measure is indeed tapping the variable it was intended to measure. This seems simple enough, although a trifle circular.

As Nunnally points out, psychological constructs vary in terms of breadth, complexity, and degree of abstraction. The more abstract they are, the more necessary it is to measure them with some validity. Unfortunately, it is also the case that the the more abstract such constructs are, the more difficult it is to be sure that measures of them are accurate and valid. Similarly, constructs also vary in terms of the size of the domain from which items can be selected, and in the relative specificity or generality of the

construct definition. The larger the potential domain and the less specifically the construct is defined, the more difficult it is to decide what kinds of items belong in the domain and which don't. It is probably safe to say that up to this point very few other psychological constructs have wider potential domains and looser definitions than masculinity and femininity. Such a realization suggests that an elaborate groundwork must be laid down in this area before substantive and reliable conclusions can be drawn about these constructs.

Definitions and validity. If the first generation measures could be criticized for the lack of a theoretical definition of the construct they were attempting to measure, the second generation measures are only mildly improved in this regard. In both cases one looks in vain for explicit statements about the nature of the constructs, their appropriate domains, and the expected relationships between variables. Rather, the definitions of the constructs themselves are largely deduced from the assumptions and structures of the measuring devices themselves. This approach has obvious limitations.

Pedhazur and Tetenbaum (1979) have strenuously argued that the BSRI is not based on a discernible theory about masculinity and femininity but is totally reliant on an empirical approach to measurement. The BSRI's methodological shortcomings, highlighted by their factor analytic data,

stem in part from this problem of definition. Nothing, they note, is done to differentiate what area of sex roles is being investigated. Though the BSRI derives from stereotypes, nothing is explicitly stated about what aspects of stereotypes are being studied or how they are being applied. Such oversights, they argue, inevitably lead to questionable validity and ambiguity in the conceptualization of the research. Two of their comments are especially pertinent at this point:

> Instead of defining the domains of masculinity and femininity and attempting to construct measures consistent with the definitions, Bem has chosen a strictly empirical approach. (Pedhazur & Tetenbaum 1979, p. 998)

> The absence of theoretical definitions of the constructs precludes attempts to determine whether or not a given set of traits is representative of a given domain. How can one assess the validity of a measure when the construct it is supposed to be measuring is undefined? (Pedhazur & Tetenbaum 1979, p. 1012)

In her rebuttal (Bem 1979), Bem defends her work by asserting that the Bem Sex Role Inventory is indeed based on a theory--one which distinguishes sex-typed from androgynous individuals. In her discussion it becomes abundantly clear that Bem does <u>not</u> hold a theory of masculinity and femininity as such, but instead views them as a "hodgepodge" of attributes linked together by "historical accident". She goes on to state that the BSRI is thus based on "a theory about both the cognitive processing and the motivational dynamics of sex-typed and androgynous individuals" (Bem 1979, p.

1048); namely, that some individuals pay more attention to sex roles than other individuals, and therefore are more likely to monitor and modify their behavior than are other individuals. The lack of descriptiveness about the masculine or the feminine individual in this definition of sex role orientation is obvious. However, "the purpose of the BSRI is to discriminate between those individuals for whom this hodgepodge does form a unitary cluster and those individuals for whom it does not." (Bem 1979, p. 1049)

In a sharp retort, Pedhazur and Tetenbaum state that they are unable to find any record of such a theory of sex roles. Neither do they accept the description offered as even a "rudimentary theory". They state:

> Asserting, as Bem does, that some individuals are motivated to conform to sex-typed cultural norms and that others are not motivated to do so, or that some individuals are "consistent" and others are "inconsistent", is tautological unless one articulates a theoretical explanation for such phenomena. (Pedhazur & Tetenbaum 1979, p. 1016)

In this heated exchange, Pedhazur and Tetenbaum's point is perhaps well taken. Bem did not lay out a theoretical explanation for what she was trying to measure and if she had, the BSRI might not have displayed some of the methodological shortcomings that were outlined earlier. On the other hand, Bem has clearly stated that her interest in this area is not in masculinity and femininity per se, but in androgyny theory. Logically, this may imply a misapprehension of the purpose of psychological measures and their uses.

The approach to measurement of Spence and Helmreich is more straightforward and defensible. To review their definition of psychological masculinity and femininity:

> (C)lusters of socially desirable attributes stereotypically considered to differentiate males and females and thus to define the psychological core of masculine and feminine personalities. (Spence & Helmreich 1978, p. 4)

It may be argued that this kind of definition is relatively uninformative insofar as it echoes the structure of the PAQ itself. Further, it leaves the reader in the dark as to what masculine and feminine personalities are. It does include several aspects which give clues to the presumed nature of the constructs, however. First, of course, they are treated separately. Second, the definition is limited to attributes. Although attributes in itself is a relatively vague term, "attributes" can safely be distinguished in this case from behaviors, interests, vocations, attitudes, role enactments, and demeanor--all of which have previously been labelled as aspects of masculinity and femininity by the first generation of M-F theorists. Clearly, Spence and Helmreich intend that such a distinction be made, since they draw a precise distinction between personality characteristics such as those measured by the PAQ and other types of role-related phenomena. They tend to see the PAQ as a reflection of the individual's repertoire of sex-role traits, in contrast to Bem's point of view that the BSRI reflects a behavioral repertoire.

Implicit in the definition given by Spence and Helmreich is an unspecified connection between sex role traits which are clustered together in some identifiable fashion and gender itself. This leads one to speculate what the similarities and differences might be in terms of their respective relationships to other phenomena, i.e., how the effects of sex-role are to be distinguished from those of sex as a variable.

In essence, this definition is a description of what the PAQ does, not of what M and F are. From the point of view of Pedhazur and Tetenbaum, which maintains that measures should be created only after the constructs have been meaningfully defined, this approach to definition may still be seen as inadequate. On the other hand, there is an alternative possibility that it is not always possible to create definitions with the precision that would be desirable. The assertion that a theoretical definition can be given a priori without a more substantial basis in empirical work may be in error. It may be premature to state a "definitive definition" for these constructs in the absence of more information about sex roles and personality. These tentative operational definitions and the process through which they have developed and been refined, may in fact be the necessary precursors of an adequate construct definition.

Myers and Gonda (1982b) have also been critical of the lack of definition. In the first place, they argue that the

BSRI and the PAQ have defined constructs as orthogonal without any real theoretical justification for doing so. "Separate", they say in effect, is different from "orthogonal", a term which implies a mathematical relationship--the total lack of correlation between two variables. Some of the second generation measures do display orthogonality for some populations, others do not. They also suggest that the point of view of the subject should be taken into account when we define masculinity and femininity: what we call masculinity and femininity may not be what they do.

If by this they mean that subjects should be enlisted in the process of uncovering the constructs of M and F and helping to validate measures in the manner described by Mischel (1977), this is a defensible and even cogent comment. One would not infer this from their methodology, however. The data they present to buttress their argument were gathered by asking large numbers of respondents to give open-ended definitions of the words masculinity and femininity. When coded into categories, these definitions were found to be largely unrelated to the content of the BSRI. Only between 10 and 14 percent were similar to the BSRI traits. As the authors fail to indicate either what question they hoped to answer by this procedure or inform us what is indicated by the results, we can only speculate that they wished to show that the content of the BSRI varies from the definition of the terms which subjects might have. In fact, all that

they have really shown, once again, is that stereotypes elicited in an open-ended format show significantly less agreement and less overlap than stereotypes elicited in a structured format. Open-ended definitions of intelligence would not be likely to reflect the content of psychological tests of intelligence, either.

The problem of construct validity does not require that operational definitions match those of the public at large, What it does require is that the phenomenon in question be related in a predictable and theoretically reasonable fashion to other phenomena of interest. It is in this sense that the lack of adequate definitions poses a severe difficulty in establishing the construct validity of the second generation measures of M and F.

The use of stereotypes in self ratings. The PAQ and BSRI are both based on stereotypes about males and females. In each case, groups of judges were asked for their evaluations about the appropriateness or desirability of certain characteristics for males or females. The use of the PAQ or BSRI presupposes that the degree of an individual's masculinity (for example) can be represented by the degree to which that individual identifies with the attributes that are generally judged to be more characteristic or desirable in males. The phrase "identifies with" in this context indicates nothing more than the relative endorsement of various items as self-descriptive by the individual.

In simplest terms, if I rate myself highly on terms like Independent, Self-reliant, and Dominant, I receive a higher M score than the person who rates himself lower on such terms.

Naturally, this transposition from stereotype ratings to self ratings involves an assumption that the same stereotypes apply to the terms used in rating scales when individuals rate themselves as had applied when the stereotype ratings were made. Is there any reason to doubt that this is indeed the case?

Pedhazur and Tetenbaum (1979) have speculated that since the pattern of factors for stereotype ratings looked different from the pattern of factors for self ratings, the two situations might not be equivalent. Such speculation is not justified by the data since one analysis included all three BSRI subscales (M,F, and Neutral) and the other only included the M and F subscales, a fact that the authors themselves noted.

Approaching the problem from a different direction, Locksley and Colten (1979) present an extensive deductive argument against the assumption that stereotype items are useful for self-ratings. On the basis of this argument, they question the validity of the BSRI and PAQ as measures of masculinity, femininity, and androgyny. The first of their overall concerns is "the feasibility of using inventories developed to tap general perceptions of aggregate sex

differences as measures of individual differences" (Locksley & Colten 1979, p. 1018).

This objection bears upon the validity of conclusions about the sex-role orientation of individuals which are drawn from inventories made up of items reflecting sex-role stereotypes. Locksley and Colten question whether it is any more valid to use the dichotomous stereotypes for males and females as a basis for continuous M and F scales than it was to use actual sex differences. They dispute the premise that sex role orientation within individuals actually covaries with traits popularly thought to be desirable or typical of men versus women.

In what way might the ratings of stereotypes have shortcomings as a source of information for building M and F measures? One way suggested by these authors is that in making stereotype ratings, judges may be confounding the "pure" masculine and feminine dimensions with male and female family and work roles. Consequently, judges may try to express these roles in trait terms; as a result the trait ratings for stereotypes may not reflect the pure dimensions. In response to this suggestion it could be said that we have little or no reason to believe that judges are using these role-conceptions in this manner. The argument rests on the speculation of the authors. It is also possible that family and work roles that are sex-stereotyped are important in the discrimination between masculine and non-masculine individu-

als and consequently will not affect the validity of the scales adversely, but rather may actually enhance it.

For Locksley and Colten, however, the implications of such problems are significant. For one thing, the use of self-rating scales for high school and college students may be problematical since they have not attained maturity and developed into those same work and family roles, which have been surreptitiously built into the M and F scales. Consequently, it may appear that there are many more androgynous individuals than there would be in a different population. This point of view directly contrasts with the view of others that the study of sex roles in young populations may be perilous for precisely the opposite reason: they may be overly conscious of and rigid in their adherence to stereotyped sex roles (e.g. Pleck 1975). However, another problem, according to these authors, may be that general sex stereotypes may be too global for interpreting and guiding behavior at the level of individual self-perception or self-direction. In other words, we may all have global stereotypes but we really don't apply them with regard to ourselves. The authors propose that instead of using global stereotypes it might be preferable to study those concepts most pertinent to self-description and self-perception.

This brings us to a second major concern of Locksley and Colten, regarding measurement, namely "the appropriateness of a traditional individual differences approach to the

phenomena of sex roles, sex differences in personality and behavior, and sex identity." (1979, p. 1018) An excellent point made in this context is that the use of the kind of abstract items which make up the PAQ and the BSRI invite the subject to make automatic adjustments in terms of comparing self to others. That is, the subject who endorses the items is making mental comparisons in order to rate him- or herself on the scale continuum. The mental set which the subject adopts may already be adjusted for comparisons only to members of the same sex. If so, this might raise a question about the comparability of items between males and females. The objective scores may lend a "surface impression of equivalence" that is specious.

If one admits that the use of stereotypes is ill-advised and the individual differences approach inadequate then some other model must be put forward for the study of sex roles in personality. Locksley and Colten advocate the adoption of the cognitive model of prototypes as a basis for M-F research. In this context, a prototype is defined as a "good instance" of a category represented in a particular object surrounded by other objects of decreasing similarity to the prototype and decreasing degrees of membership.

The problem as I see it in suggesting alternatives in the study of sex roles is that any alternative must in some way rely on an ability to distinguish among individuals with regard to different aspects of sex role. To the degree that

this is true, it involves the general process of measurement. If the constructs being measured by the BSRI and the PAQ are either irrelevant or unrelated to what needs to be measured, then some definition of what needs to be measured must be made and some suggestion about how to measure it must be developed. In this regard, the critique of Locksley and Colten suffers equally with the work they are criticizing from the lack of explanation of the meaning of masculinity and femininity.

Whether one discusses the topic in terms of traits (Hogan, DeSoto & Solano 1977; Spence & Helmreich 1978), cognitive schemata (Bem 1979; Markus 1977; Myers & Gonda 1982b), prototypes (Locksley & Colten 1979), sex role salience, sex role transcendance, or cognitive complexity theory (Myers & Gonda 1982b), there still needs to be a means of systematically differentiating between individuals along these significant dimensions of personality. In any case, we would need to know how important the sex-linked constructs are to individuals, or how much they incorporate them into their self-concepts, or how much they identify with them. The excellence of theory in this area is chiefly determined by the ability of researchers to make quantitative discriminations between individuals.

Breadth of the M and F constructs. The third of the principal issues regarding construct validity to be considered revolves around the question of how broadly or narrowly

masculinity and femininity scores should be interpreted. The topic will be introduced here in terms of the theoretical statements which have been made or can be inferred from the literature. In the next section of this chapter, the implications of opposing assumptions on this issue will be considered more fully.

As we have already seen, early researchers working with the M-F construct assumed not only that M and F were opposite characteristics, but that they bridged many different kinds of phenomena. The lists of subtests and dimensions for the various tests included not only traits, but attitudes, vocational interests, recreational interests, behavior, preferences, and so on. The assumption was made that M-F was such a distinctive and broad-based construct that all of these areas would be consistently correlated. The masculine male would be masculine in all or most ways. The femininity of the feminine female would be visible regardless of the particular aspect of of M-F being sampled. Very likely, this assumption was as much at fault for the psychometric deficiencies of these measures as the assumption of bipolarity.

The PAQ and the BSRI by way of contrast defined the proper domain of masculinity and femininity as "attributes" or positive aspects of self-concept, which were operationalized through questionnaires composed of traits. But there are substantial differences of interpretation between Bem

and her co-workers and Spence, Helmreich and their co-workers regarding the breadth of the constructs measured by their respective tests of M and F. Bem's early research, although based on a measure of more limited content has nevertheless continued conceptually to treat M and F as broad based constructs (Pedhazur & Tetenbaum 1979; Helmreich et. al. 1979). Spence and Helmreich on the other hand have been meticulous in divorcing the instrumental and expressive constructs they have dubbed Masculinity and Femininity from the various other aspects of sex role and sex role behavior.

Bem's notion of androgyny implied a strong linkage between self-concept and behavior. The nature of this linkage is not always precisely delineated. Rather, there is a conceptual blurring that sometimes occurs in her writing, such that traits and behaviors are spoken of as virtually the same thing. This seems to stem in part from a disinclination on her part to admit that the BSRI is actually a trait measure, the trait concept having fallen into some disrepute (cf. Hogan, DeSoto & Solano 1977). Bem describes the relationship between the self-concept and behavior in the following terms:

> Thus whereas a narrowly masculine self-concept might inhibit behaviors that are stereotyped as feminine, and a narrowly feminine self-concept might inhibit behaviors that are stereotyped as masculine, a mixed, or androgynous, self-concept might allow an individual to freely engage in both "masculine" and "feminine" behaviors. (Bem 1974, p. 155)

The androgyny theory as proposed by Bem, then, implies a strong connection between the measurement of personality characteristics and the degree of behavioral flexibility. It assumes that self-concept and behavior are are closely bound, so that a reasonable test of the theory of personality traits is the prediction of behavior in situations calling for masculine or feminine behavior. This presumption leads to certain kinds of predictions about the relationship of BSRI M and F scores to masculine and feminine behaviors. In her research program, Bem attempted to demonstrate that androgynous individuals were just as masculine as masculine individuals on such tasks as resisting conformity (said to represent masculine independence, an instrumental trait), and just as feminine as feminine individuals on such tasks as playing with a kitten (Bem 1975), interacting with a human baby, or showing "nurturance" to a lonely transfer student (Bem, Martyna & Watson 1976). These studies had implications both for androgyny theory and for the construct validity of the BSRI. The results were generally supportive of the hypothesis that androgynous subjects were more flexible than the sex-typed subjects although there were exceptions. Feminine subjects did not appear to perform well even on some "feminine" activities. The ability of the BSRI scores to predict in certain defined circumstances the performance of tasks which were chosen to represent masculine/instrumental themes and feminine/expressive themes was also

supportive of the construct validity of the BSRI as a measure of those constructs (Taylor & Hall 1982).

But it is also true that there is an underlying tendency here to confound behavior with personality traits. Spence and Helmreich have consistently differentiated their position from Bem's. Arguing that "the literature suggests the utility of traitlike notions when one's intent is to understand the implications of individual differences for broad areas of real-life functioning," (Spence & Helmreich 1978, p. 15) they promote a view of the PAQ as a measure of personality characteristics which will not necessarily correlate highly with specific masculine and feminine behaviors. As they describe Bem's position:

> This theoretical position rests on the supposition that the empirically diverse indicators of masculinity and femininity are all highly correlated, so that an individual exhibiting one set of attributes or behaviors in the class can reasonably be assumed to exhibit approximately the same degree of all other attributes and behaviors. (Helmreich, et. al. 1979, p. 1633)

Their position in contrast is that a distinction should be drawn between role expectations and behaviors on the one hand, and the internal properties of the actor on the other. These internal properties are of considerable importance insofar as they govern behavior over a large number of situations but they cannot be said to correlate highly with specific behaviors or to other aspects of sex role. Thus the psychological dimensions measured by the PAQ are "only weak-

ly related within each sex to the broad spectrum of sex-role behaviors" (Spence & Helmreich 1978, p. 3).

The question of the theoretical domains to which the content of the BSRI and PAQ can justifiably be expected to relate, then, is in some dispute. In a mundane sense, this issue boils down to asking whether it is legitimate to call the scales "Masculinity" and "Femininity" which at least in everyday parlance, seem to be so much more encompassing than the domains actually represented by these scales. Indeed, in examining the items of the BSRI, Pedhazur and Tetenbaum (1979) have asked what justification there is for naming the subscales Masculinity and Femininity. After all, in terms of the self ratings factors, the two terms Masculinity and Femininity form an entirely separate bipolar factor distinct from the factors on which the other items load. They have argued that the factors that contain most of the other BSRI items, and consequently the shortened subscales, be referred to by titles more appropriate to their content, such as 'Assertiveness' or 'Instrumentality' for Masculinity or 'Interpersonal sensitivity' for Femininity. This is a direct criticism of the extrapolation from the limited constructs of the BSRI to the over-arching masculinity and femininity constructs. Bem (1979) has conceded that the two items Masculinity and Femininity are in fact the worst items on the subscales in terms of their relationship to other items. She has failed to provide a compelling justification for re-

taining the titles Masculinity and Femininity for the subscales.

Spence and Helmreich (1979) have claimed that the ultimate justification for the appellations Masculinity and Femininity lies in the demonstration that the subtests discriminate between the sexes in their self-report. This is in some ways a weaker argument than could be advanced. A stronger argument would be that the item selection process for these instruments was closely linked to general perceptions about males and females. Nevertheless, Spence and Helmreich have more recently advocated that the expressive and instrumental dimensions measured by the PAQ, although significant in their own right, should be "disentangled" from the overall or total class of masculine and feminine attributes (1980).

By narrowing the definition of what is being measured by these scales from "Masculinity" and "Femininity" to an aspect of the overall masculinity and femininity constructs, Spence and Helmreich, at least, are acknowledging the problems of construct validity which have been discussed here. They are saying in effect that the overall constructs are too encompassing to be adequately measured by one short test, based in one domain.

However, in direct contrast to this point of view there is another important statement on this topic aside from the theoretical and deductive evidence which has been summarized

here. In reviewing the literature on androgyny, Taylor and Hall (1982) argue that there has been considerable confusion between those experiments and effects which are pertinent to the issue of the construct validity of the PAQ and the BSRI, and those effects that directly pertain to the androgyny hypothesis as such. They re-analyzed a number of studies, looking at three kinds of dependent variables--male typed, female typed, and non-sex-typed psychological health variables. Analyses of the sex typed variables seemed to provide significant support of the construct validity of the M and F scales of the PAQ, the BSRI and other similar measures. It would seem that this kind of empirical meta-analysis should carry substantial weight.

In summarizing, it is difficult to make an unequivocal overall statement about the construct validity of these measures, but some judgments can be made. There are important difficulties in the area of definition of the constructs which lead to difficulties in other areas. Although there is no substantive evidence against the use of stereotypes as a source of items for self-rating M and F tests, the validity of such an approach and the implications of such an approach have been questioned. In many ways, this question will depend on further evidence and bears examining in an empirical forum. Finally, related to the adequacy of definition in this area, is the problem of determining the breadth or narrowness of the M and F constructs as measured

by the PAQ and the BSRI. Construct validity asks whether what was supposed to be measured is the same as what is being measured. Regardless of the value of the dimensions measured by the PAQ/BSRI, profound doubts have been raised as to whether they are measuring masculinity and femininity per se. In the next section, we will examine one of the trends in this area which has directly resulted from some of these criticisms, the trend toward examining other domains of masculinity and femininity.

NEW DIRECTIONS FOR PSYCHOLOGICAL SEX ROLES

The future resolution of the problems with the second generation measures would appear to depend in part on a careful re-definition of the masculinity and femininity constructs. If these are in fact extended over a wider range than can be accurately assessed by a single brief instrument, this necessarily implies a broadening of measurement efforts, and the proliferation of unique devices aimed at distinguishing separate or distinct dimensions of personality related to sex role orientation. A look at recent theory and a survey of another set of sex-role measures will precede the explanation of the reasoning which led to the creation of the Sex Typed Activities Test.

"The Many Faces of Androgyny"

The breadth of the operational definitions of masculinity and femininity, historically, has narrowed to the point where it is questionable whether the widely used measures of the constructs are sufficiently encompassing to be called measures of masculinity and femininity as such. In a trio of articles, Spence et. al. acknowledge this fact and suggest new directions for the inquiry into sex roles as they relate to personality.

In the first of these three articles, by Helmreich, Spence and Holahan (1979), a "conceptual replication" was made of an important early study by Bem and Lenney (1976). The overall purpose of the experiment was to demonstrate the greater behavioral flexibility of androgynous individuals, i.e., that androgynous individuals were less likely than sex-typed individuals to avoid cross-sex-typed behaviors. It was thought that cross-sex-typed activity would be motivationally problematic for the sex-typed individual. Briefly, Bem and Lenney set up a ruse which offered subjects the chance to pose for photographs while acting out everyday activities. Prior to posing, subjects were given a chance to rate their preferences on the tasks. Tasks were chosen to be representative of either sex-stereotyped or neutral activities. As an example, nailing two boards together was a masculine task. Avoiding cross-sex tasks cost the subject money since they were to be paid a few cents for each task, but less for sex appropriate tasks. After performing a few tasks (three masculine, three feminine, and three neutral) they were asked to make a series of "comfort" ratings about how they felt for each activity. Males were asked how masculine they felt, females how feminine, and all were asked to make ratings on how attractive, likable, peculiar, and nervous they were performing each task. Interestingly, "comfort" itself was not one of the items.

The experimental design was unfortunate since it failed to treat M and F as separate variables. Instead a subtractive method was used. Subjects were pre-sorted into three groups: extreme masculine, androgynous (balanced), and extreme feminine. As a result, it it difficult to assess from the reported results the independent main effects of M and F. Nevertheless, the results indicated a main effect for sex role group such that sex-typed individuals were significantly more stereotyped in their preferences than either androgynous or sex-reversed subjects. Using post-session comfort ratings, a "negativity" score was compiled from the separate ratings. Again, sex role showed a significant main effect with sex typed subjects reporting greater discomfort with cross-sex activities than either sex-reversed or androgynous individuals. When the experimenter was of the other sex, they were especially uncomfortable demonstrating sex-inappropriate activities.

This study is significant both in its implications for the broad vs. narrow distinction and in terms of the construct validity of the M and F scales of the BSRI. It seems to imply that the M and F constructs are broad, encompassing both behaviors and personal qualities. It also seems to lend a quasi-predictive validity to the BSRI as a general measure of sex-roles.

Helmreich, Spence and Holahan (1979), in a re-appraisal of these data interpreted Bem's position as exemplifying a

'global' view of masculinity and femininity which seems to imply that the constructs are "manifested in a variety of gender-related behaviors, role attitudes, and personal qualities." (1979, p. 1633) Assuming that the PAQ is only a weak predictor of other classes of sex role behavior, they set out to demonstrate a more limited hypothesis. The PAQ, they said, would predict to only a mild degree subjects' comfort with, or preferences for, sex typed activities.

In this study, two groups of ratings were made by subjects in anticipation of getting photographed or videotaped. The ratings were made for anticipated comfort with and preference for a group of everyday behaviors much like the ones used by Bem and Lenney. The results tended to support the authors' contention that the PAQ would be only mildly related to sex role behaviors. The relationships of M and F as measured by the PAQ were positive in sign, but generally quite low in magnitude. The major contribution to the variance between types of tasks was sex, with both sexes preferring to perform and feeling more comfortable with tasks congruent for their sex. The Attitudes Towards Women Scale (Spence & Helmreich 1978), a measure of profeminist attitudes, proved to be a superior predictor of sex typed preference in males and of overall comfort in females than the PAQ. One finding of some interest was that the three types of comfort ratings, M, F and Neutral were all highly correlated (average $\underline{r}$ = .67). Due to this, Helmreich et. al.

combined all three types of tasks in one measure of overall comfort, although the theoretical implications of overall comfort ratings are unclear.

On the basis of their results, Helmreich et. al. challenge the view that instrumental and expressive traits are both "stongly and directly related to other masculine and feminine behaviors and attributes" (1979, p. 1642). The complexity of the data they present leaves some room for dispute on this point, but in general they argue that this study reflects their point of view that relationships between different kinds of indicators of masculinity and femininity are generally weak and complex. Certainly, they did not find any strong relationships between PAQ M and F scores and comfort with or preference for the sex typed tasks that they used. Finally, they conclude that there is little support for the androgyny hypothesis which predicts that androgynous individuals are more flexible than sex-typed individuals.

An even stronger statement of this theoretical perspective was presented in the formal reply (Spence & Helmreich 1979) to the critique of Locksley and Colten (1979). In this paper, entitled "The Many Faces of Androgyny," they significantly extend their definition of masculinity and femininity. In essence, the constructs measured by the PAQ are seen as only one aspect of masculinity and femininity. The choice of this particular method of approaching the con-

structs is defended but again they note their skepticism that one aspect of M and/or F will necessarily relate to all other aspects. They admit that it is unclear how salient the particular aspects measured by the PAQ are to individuals whose own definitions of the constructs are going to be weighted differently. Still, they maintain that it is justifiable to call these constructs masculinity and femininity. But more importantly, there are, they say, many masculinities and femininities. They reject the notion of an overall or global measure of the superordinate constructs as "a delusion".

In a third article which reviews the relationships of M and F to a variety of other sex role phenomena, they go a step further, advocating the "disentangling" of instrumentality and expressiveness from the global concepts of masculinity, femininity, and androgyny. In this paper, they distinguish between the types of traits measured by the PAQ and role expectations or role attitudes. Role expectations have to do with the kinds of behavior that others seem to expect of the individual and the associated consequences of those kinds of behavior. Role attitudes are the beliefs about the legitimacy of these expectations. All three are seen as the roots of behavior that is sex role related, but no one by itself determines behavior entirely.

Spence and Helmreich argue that it is more appropriate to regard these measures narrowly as trait measures of so-

cially desirable instrumental and expressive characteristics. As measures of sex roles the BSRI and PAQ are said to have no face validity and minimal construct validity. This is a radical departure from the history of these measures, and a significant de-emphasizing of their purported relationship to sex roles. It may in fact go too far in restricting the implications of work done with these instruments. In essence, these two instruments were originated in ratings based on sex-typed desirability and sex-role stereotypes. To divorce them from these constructs entirely seems unjustified. Nevertheless, Spence and Helmreich have articulated the inadequacy of the BSRI and PAQ scores to stand for global estimations of the superordinate constructs of M and F. Implicit in this view is the need for alternative approaches which examine theoretically relevant and empirically distinct aspects of masculinity and femininity.

In their 1979 rebuttal to the critique of Locksley & Colten (1979) they make this very point, and describe briefly their own efforts to extend the analysis of M and F dimensions into other domains.

> A profitable way to proceed, we have suggested, is to identify the principle components or classes of psychological phenomena related to gender and to devise methods to measure them so that their interrelationships may be determined (Spence and Helmreich 1979, p. 1039).

Even before Spence and Helmreich advanced the idea that masculinity and femininity were multi-faceted, they them-

selves as well as other investigators were beginning to look at other aspects of sex role attitudes, self-concept, and behavior. The measures which will be discussed below probably represent only a few of the possible choices which the researcher might have. These measures however are those which are most directly relevant to the current effort to build a new, behaviorally based measure related to masculinity and femininity.

Attitudes Towards Women Scale (AWS)

As previously mentioned, the AWS is a measure of pro-femininist attitudes, i.e. high scores indicate egalitarian attitudes toward the rights and proper roles of women in American society. An example of an item is, "A woman should not expect to go exactly the same places or to have quite the same freedom of action as a man." (Spence & Helmreich 1978) This is rated on a scale from strongly agree to strongly disagree. Stereotype ratings on PAQ items are more strongly related to AWS scores than are self ratings on the PAQ. This seems to suggest that the degree to which individuals perceive the sexes as different is more closely related to their "progressivism/conservatism" on women's rights, than it is to their own level of M and F as measured by the PAQ. Orlofsky (1981) has referred to the AWS as a measure of sex role "attitudes", but this is somewhat misleading. The AWS represents attitudes about sex roles in

general. It is not however, the same as a measure of "masculine or feminine attitudes" in the same sense as was intended by, say, Terman and Miles.

Sex Role Identity Scale (SRIS)

Storms (1979) describes a short six-item scale which is designed to examine the "sex-role identity" of the respondent, as opposed to the sex role orientation. Ratings for M items and F items were moderately to strongly correlated in a negative direction. This scale is said to represent a "global self concept" of one's masculinity and femininity. A sample item, which would be rated on a 31 point scale, is: "In terms of the typical image of what is masculine in our society, how masculine is your personality?" The endpoints of the scale range from "Not at all masculine" to "Very masculine".

Extended Personal Attributes Questionnaire (EPAQ)

In 1979, Spence, Helmreich and Holahan reported the development of an extended PAQ with items added which dealt with negative characteristics as well as the previously explored positive traits. Other investigators have also been intrigued by negative aspects of M and F (cf. Kelly, Caudill, Hathorn & O'Brien 1977). The EPAQ produces three additional scores beyond the previously described M, F, and MF scores. The first, M-, is a eight-item measure of socially

undesirable characteristics generally thought of as more typical of males. It is said to represent the relative absence of positive feminine (commnunal) qualities. The items include: arrogant, boastful, egotistical, greedy, dictatorial, cynical, looks out only for self, and hostile.

The female negative characteristics are broken down into two subcategories. The first cluster, F_c-, describes the absence of male-valued instrumental qualities: i.e., spineless, servile, gullible, and subordinates self to others. The second cluster, F_{va}-, (for verbal passive-aggressiveness) subsumes the items whiny, complaining, fussy, and nagging.

A complete review of the findings requires more space than is available here. However, they noted negative correlations between the M+ scale and the F- scales as well as between the F+ and M- scales. Only mild relationships were observed between the respective sex typed positive and negative scales (e.g. M+ and M-), further evidence in the view of the authors for the multi-dimensionality of M and F (Spence, Helmreich & Holahan 1979).

Sex Role Behavior Scales-1,2 (SRBS)

Orlofsky (1981) introduced a 160-item version of the Sex Role Behavior Scale (SRBS-1). An expanded 240-item version (SRBS-2) was subsequently described in a second article by Orlofsky, Ramsden, and Cohen (in press). The Sex Typed

Activities Test, the development of which is detailed in this paper, is most closely related in terms of theoretical rationale to the Sex Role Behavior Scales, of any of the tests discussed here. Both are attempts to extend the dualistic notion of M and F measurement into the behavioral realm. Both are based on the general stereotypes held by both males and females. Both produce separate M and F scores. There are however significant differences between the two which will be apparent as the two are described.

In the first place, the Orlofsky scales are quite long. The SRBS-2 has 240 items: 80 MV or male-valued items; 80 FV or female-valued items; and, 80 MF or sex-specific items. Each of these three scales contains four subtests: Recreational Interests, Vocational Preferences, Social and Dating Behaviors, and Marital Behaviors. Consequently, for a particular respondent, the entire test delivers 12 subtest scores and 3 overall scores. It should be apparent that although this test incorporates the quasi-dualistic approach to measurement of Spence and Helmreich's PAQ, it is an omnibus measure in terms of content. The three overall scores combine a series of subtests of various content into overall scores.

The SRBS attempts to reinstate some of the same content domains used by the first generation measures, employing the M/F/MF arrangement of scales. Subsequently, overall scores are created across four conceptually separate domains on the

assumption that these domains all covary to a high degree. In Appendix A, the SRBS scales in both versions are more thoroughly reviewed and some possible objections to them are raised. In particular, the variability of the relationships between the 12 individual subtests casts doubt on their utility as a single test. Also, the creation of overall M, F, and MF scores does not seem justified, despite high overall reliabilities, because of the poor inter-item correlations which obtain across the component subtests. Third, the decision to use three scales, instead of two or perhaps four is questioned.

An interesting observation about the early version of the SRBS (SRBS-1) can be made. While the MF scale is made up of negatively correlated groups of masculine and feminine items, the separate M and F scales correlate positively. In overly simplified form this seems to indicate that the more masculine one is, the more feminine one is, at least with regard to interests and social roles. This is a curious finding, which is given relatively little discussion. Suffice to say, it raises once again the issue of the lack of a theoretical explanation for the relationship between masculine and feminine dimensions.

Male-Female Relations Questionnaire (MFR)

Spence, Helmreich and Sawin (1981) have developed a questionnaire which assesses an individual's personal preferences with regard to sex roles, as distinguished from their overall philosophical attitudes about sex roles which are measured by the Attitudes Towards Women Scale. This new questionnaire in many ways is similar to the AWS. It is not referred to as a masculinity/femininity test, but it seems to assess a preference for traditional masculine role preferences in males and traditional feminine role preferences in females. And, this particular test is discussed by Spence and Helmreich as an example of their attempt to expand inquiry about sex roles into alternative domains (Spence & Helmreich 1979).

The questionnaire itself parallels two of the subscales of the Orlofsky Sex Role Behavior Scale (Social and Dating Behavior, Marital Behavior) but with significant differences in structure. Again, the SRBS-2 is arranged so as to be administered to either males or females and yields an M, F, and MF type of scoring. In contrast, Spence et. al. chose to build separate questionnaires (one for males, one for females) asking about role preferences. Many of the items on the male and female versions are parallel in form. For example, on the Marital Scale of the MFR the male item is "One of my wife's jobs should be to help me in my work by taking pressure off me at home." The corresponding female item is

worded: "One of my jobs should be to help my husband in his work by taking pressure off him at home." Spence, Helmriech & Sawin 1981).

Each version of the MFR contains three separate subscales. The first and third of these are similar for both sexes and are entitled Social Interaction and Marital Roles, respectively. The Social Interaction scale measures a general personal preference for interacting with the opposite sex in a manner consistent with traditional sex roles where the male is the leader, and the woman more supportive. The Marital Roles scale measures the preference for a traditional male-female marital relationship where the male retains a certain authority and is the major provider for the family and the female is in charge of the home and takes principal responsibility for the children. The content of these scales, though largely parallel, does vary in terms of nuance. Reading over the items on the two alternate forms, the roles do seem to vary in a subtle way depending on whether a male or a female is describing him- or herself.

The second subscale in each case expresses more explicit differences between the sexes. For males, the second subscale is called Expressivity and measures a preference to have or to be perceived as having a high degree of emotional self-control. For females, the second subscale is called Male Preference and assesses a general preference to relate to masculine males who are not easily dominated by females.

The assignment to subscales is largely the result of factor analysis. The final scales each have 16 Social Interaction items, 4 Expressivity/Male Preference items, and 10 Marital Roles items. The scales are all satisfactory in terms of reliability. They are also moderately to highly correlated with one another, although the correlations are lower for the 4-item sex-dependent scales which have lower reliabilities due to their brevity.

Scores on these scales are also highly related to the AWS scales but show modest to negligible relationships with the subscales of the EPAQ, with self-esteem, and with the measure of comfort and sex role preference described in Helmreich, Spence & Holahan (1979) which was reviewed earlier.

The Present Inquiry

The movement into new domains of M and F displays both a wide variety of areas of investigation as well as a wide variety of measurement approaches. Some measures, such as Storms measure of sex role identity and Orlofsky's MF scale, are traditional in form-- bipolar, unidimensional, and global. Some are unidimensional but separate for the sexes, such as the Male-Female Relations Questionnaire. Others, for example the SRBS-2, attempt to apply the dualistic model across a wide variety of sex role phenomena. The EPAQ tries to extend the trait approach to the negative traits that are sex stereotyped.

The variety of these approaches graphically illustrates the movement of masculinity and femininity research toward a more diverse, less homogeneous strategy. It also points up the absence of guiding theory in this area. How are decisions made to attempt to measure some aspects of M and F in a dualistic fashion (as in the BSRI) while other aspects are measured either globally or separately for the two sexes? Is it necessary or desirable to have the same basic structure for all tests in this area? On the other hand, is it necessary to apply different structural models to each isolatable domain that requires investigation? What does it mean if M and F dimensions are sometimes negatively correlated, sometimes orthogonal, and sometimes positively correlated? What phenomena are subtypes of masculinity and femininity, and which are to be excluded from that realm of phenomena? These are some of the perplexing issues which one encounters in an overview of these efforts to examine and measure psychological attributes related to sex roles. Further discussion of these issues is deferred until the results of the STAT development process can be considered together with these other measures.

Some theoretical points

An unequivocal definition of psychological masculinity and femininity does not yet exist. To produce and explicate such a definition would no doubt require a separate disser-

tation in itself. Any researcher, though, operates from certain preconceptions about the phenomena studied. These assumptions and hypotheses often go unspoken, or at least they go unwritten. Prior to summarizing this review of previous research and explicating the hypotheses which governed the development of the Sex Typed Activities Test, I would like to explain, from an informal perspective, my own approach to masculinity and femininity, and the working definitions which were employed.

It is clear that individuals differ in regard to many aspects of personality which relate to sex roles. While one man likes to go hunting, another prefers to go to art museums, even fashion shows. One woman likes to wear frills and cosmetics, while another may prefer to wear jeans and flannel shirts most of the time. These phenomena are not necessarily subtle. They appear to be robust, and yet they have proved elusive insofar as measurement is concerned. Clearly a central difficulty lies in the complexity of sex roles in our society. Many different men feel "masculine", but may share very little in terms of activities, interests, or even traits. Self-concepts may place different emphases on different aspects of the personality, so that the very characteristics which are seen as evidence of one woman's "femininity" are thoroughly irrelevant to another woman who considers herself just as feminine, but for her own completely different reasons.

It seems to me that we can say that psychological masculinity is an orientation to things male, while psychological femininity is an orientation to things female. This definition may seem to be too vague, but the definition is no more broad than the constructs. To give examples of what is meant by "things male and female", consider the following seven categories: demeanor, interests, traits, activities, skills, role enactments, and vocations. I'm not sure, as were Terman and Miles, that there are characteristically masculine and feminine attitudes, although the media attention placed on the role of the female voter in the 1982 elections suggests that there might be. The basis for including the seven categories listed derives in part from the survey of previous tests in the preceding sections.

This definition incorporates the notion of M and F which have become increasingly prevalent: that they are separate and that they represent a variety of domains. Within each domain, there are things (or items) which are more characteristic of one sex than of the other, as well as items which are irrelevant to the sex of the individual. Bem has argued that these things which we call masculinity and femininity are a "hodgepodge" thrown together by historical accident. I cannot agree. There is a general order to these phenomena which results from the differences, on average, between the dispositions of males and females which are observable from birth onward (cf. Maccoby & Jacklin 1974) in

interaction with the demands and reinforcers within the particular culture. This point of view emphasizes neither inborn qualities nor learning in isolation but rather the interaction of the two. The variety of aspects of life we call masculine (or feminine) in our culture, even though they may differ on the basis of class, race, geography, or other factors, are united by cohesive themes which should be discernible and, indeed, measurable.

What then does it mean to say that a person is psychologically masculine? That is to say, how should the individual who scores high on a particular measure of masculinity differ from one who scores low? In the first place, according to the definition it makes no sense to speak of an individual being masculine in psychological terms without some reference to the aspect of masculinity being discussed. Intuitively, this is not difficult to grasp. The highly assertive and independent individual (M traits) may still be interested in fashion (interests) while carrying himself in a way that is about average for males (demeanor). But the prototypical masculine individual would be one who takes particular interest in masculine activities, such as those involving sports or mechanics, dresses with an eye to function rather than fashion, tends to be more interested in fulfilling a traditional male role in the family, and so forth. This kind of rigidly consistent orientation to things male would seem to be the exception rather than the

rule. To the extent that the individual fails to conform to the conventional stereotype regarding a particular aspect or domain of masculinity, we would be justified in saying that he or she (for women can also gravitate toward this model) is less masculine with regard to that attribute than someone who conforms to a high degree.

High feminine individuals are those who are most oriented to things female. In spite of a decade of attacks on sex roles, most people still have fairly clear notions about what those roles are supposed to be. The woman's role has shifted probably more than the man's. Female things, to give a few examples, are concerns around child rearing, domestic activities such as cooking, an interest in design or fashion, or fine arts. To say so is not to imply that these are exclusively female, or that females are better equipped for such things. It merely indicates the general attitudes of society about sex stereotypes. Within a given domain, high feminine individuals should be seen to conform to the stereotype, or to identify with the stereotype to a greater degree than low feminine individuals.

These sample descriptions are not mutually exclusive, since either applies to both males and females. Any person differs from others in regard to these two orientations. In cognitive terms, we may say that each person has a schema for each sex which varies in its importance and complexity. It is also true that not all people have personally signifi-

cant schemata for the same things (Markus 1977). For some, sex role influences their world-view and their cognitive processing to a much higher degree than is true for other individuals. The relative M and F scores for an individual reflect not only the relative degree to which an individual conforms to a stereotype but also the degree to which it is important to the individual to distinguish between the two and identify more with one than with the other. Such a score is probably composed of three elements, the first being a person's general level of comfort at performing various behaviors; the second a relative contrast between sex appropriate and sex-inappropriate tasks or behaviors, and the third, a residual or error component. This is where the notion of sex-role orientation, broadly defined, becomes important. It is a phenomenon which is studied indirectly through experimental designs which use both M and F scores as independent variables. Both males and females grow up with both masculine and feminine models. The notion of sex role orientation describes the relative "pull" of two kinds of sex role attributes--those that most people feel are characteristically and/or desirably male, and those that are characteristically or desirably female.

It is true that individuals do have the capacity to introject or internalize the characteristics of both sexes. But the degree to which they are willing to do so depends in large part upon the domain of which we are speaking. This

is part of the reason why I distrust the term "androgyny" which seems to imply a general mixing of sex roles, which is not prevalent in this or any other culture. Androgyny as measured by positive abstract terms is most likely to seem plausible precisely because of its high level of abstraction from the dimorphism of behavioral sex roles. It can be argued that individuals of different sexes may give themselves quite similar ratings on a set of items such as those which comprise the BSRI even though the individuals are quite different in terms of masculinity and femininity. This can occur for two reasons. First, as noted earlier, there is the possibility pointed out by Locksley & Colten (1979) that individuals may make ratings which are based on comparing oneself only to members of the same sex. This amounts to a "cognitive adjustment for sex" which is more easily made where the items themselves are abstract in nature. Second, the thesis is advanced here that the universe of masculine and feminine dimensions of personality goes beyond that which is reflected by traits alone, in contradiction to Bem's operational definition. Consequently, sex-specific negative characteristics, as well as sexual-family roles, sex identity, and sex role behaviors represent areas in which the less ambiguous aspects of sex role are salient, and androgyny is less likely to appear. These suggestions will be discussed at greater length following the presentation of the results of the present study.

As methods of distinguishing between masculine vs. non-masculine or feminine vs. non-feminine individuals with regard to a particular content domain become validated, the true picture of androgyny will result. It may not closely resemble the idealized abstraction advanced during the 1970's. And its advantages and influence may not seem as great. Nevertheless, it will shed some light on the advantages or disadvantages of being more or less sex-typed in various ways.

M and F are generally thought of as qualities of personality or "traits". The most common approach to measuring these has been paper-and-pencil measures which measure self-attributes or self-concepts about particular traits represented by item content. In the case of M and F, these self-attributions can reflect notions about the self across all of the domains which we have listed. The development of a multi-dimensional approach to masculinity and femininity provides some significant dilemmas. What should be the relationships between M and F across a number of domains? What does it mean, for example, if a particular M score is correlated with the F score for a particular domain more highly than it is correlated with another M score from a different domain? If M and F are negatively correlated, should they necessarily be placed on a single bipolar dimension? These sorts of questions underscore the frequently deplored problem of an absence of theory in this area.

Masculinity and femininity, then, are seen to be broad classes of personality phenomena which are differentially valued on the basis of sex. These characteristics are differentiated from the "sex role identity" of the individual (Constantinople 1973; Storms 1979) although levels of M and F may be related to sex role identity in a manner that is as yet unspecified. Constantinople has suggested a probable relationship between these characteristics or attributes, collectively called sex-role orientation, and sex role identity, which means "both the cognitive and affective factors which reflect both self-evaluation and the evaluation of others as to one's adequacy as a male or female (1973, p. 391). Indeed, the importance of masculinity and femininity as culturally defined and the likelihood of having a schema about these attributes is probably enhanced in those individuals whose sex role identity is most in question.

The Sex Typed Activities Test

The history of M and F measurement has been traced in an effort to highlight those issues which are of greatest interest from the standpoint of psychometric theory. In this section, the reasoning behind the creation of a new measure based on behavioral items will be discussed, and specific hypotheses about the test development process will be articulated.

The shift from the first to the second generation of sex role orientation measures was marked by several notable changes in item selection, theoretical rationale, and structure. The most salient of these was the breaking-down of the single M-F construct into the two separate M and F constructs--the dualistic model. The replacement of the bipolar by the dualistic approach, however, was accompanied by an appeal to an equally dualistic theoretical formulation, a restriction of item content to traits, and the use of stereotypes as a basis for the creation of the M and F test items.

But the gradual development of measurement approaches from the omnibus approach to a more refined multi-dimensional methodology has entailed a new set of complex issues. In particular, the investigation of the whole concept of androgyny confounded the issue of test validity. In spite of the fact that separate measures of M and F had been developed, researchers continued to think about M and F as interlocking concepts which in turn influenced their choices about data analysis techniques and conceptual models in deleterious ways. In addition, criticism of the BSRI in particular centered on the lack of solid unidimensionality for each of the two subscales. More important issues involved questioning what, in psychological terms, masculinity and femininity are, and whether current efforts were adequate to measure them. The phenomena of masculinity and femininity have un-

ravelled so that it can no longer be assumed that a questionnaire which measures a small part of M and F such as traits alone, represents all important or conceivable aspects.

A recognition has been emerging that if we are to successfully study psychological phenomena related to sex roles we must extend our ability to measure masculine and feminine dimensions over a broad range of relevant domains. Previous measures of M and F have only begun to explore the possible avenues of empirical assessment of sex roles in personality. This review has provided ample discussion of the importance of developing alternative measurement devices. The Sex Typed Activities Test, or STAT, was designed here specifically to address this issue. It was hypothesized that other important aspects of M and F had been left to be explored in a systematic and quantitative fashion. The question then became which of the potential domains would be the most profitable and how the test should be constructed, to best reflect the underlying psychological dimensions.

Based on what has been studied and learned over the past decade, the decision was made to explore a behavioral aspect of masculinity and femininity: that of masculine and feminine activities. There are a number of good reasons for thinking this might prove to be of value. First of all, as Pedhazur and Tetenbaum (1979) point out, there had been a blurring of the distinction between traits and behavior es-

pecially in Bem's work. A measure directed at more accurately assessing the behavioral repertoire of an individual in terms of sex-linked dimensions seemed to be a natural next step. The work of the androgyny theorists pointed up the fact that sex typed behavior was an important variable to be studied in its own right. Bem's research was aimed at assessing the relative degree of flexibility towards sex typed behavior evidenced by different sex role groups (Bem & Lenney 1976). Helmreich et. al. (1979) replicated the Bem and Lenney study, once again looking at the relationship of M and F traits to behaviors, but arriving at different conclusions. It was thought that perhaps a paper-and-pencil measure could be developed which could more directly tap the aspect of masculinity and femininity represented by the dependent variables in these experiments. Helmreich et. al. argued that the PAQ traits were not necessarily good predictors of behavioral sex roles, but they also implied that those roles are an important variable in themselves. In any case, since it is always preferable in terms of psychometric theory to tie the measure as closely as possible to the phenomena that are to be predicted, it makes sense to try to address the issue of sex typed behavior directly through the content and structure of the test.

The Bem and Lenney study and the Helmreich et. al. replication both looked at preferences for everyday tasks and comfort on those tasks in a mild deception experiment.

Likewise, the STAT uses a number of everyday tasks of activities as items. In terms of responses, however, the use of preferences did not seem to make sense outside the experimental situation. Therefore, the decision was made to use comfort on a variety of behaviors in the hope that that this might reflect something about the person and the rigidity or flexibility of the sex role schemata that he or she employs.

The Sex Typed Activities Test is based on the premise that many stereotypes exist concerning the appropriate behaviors or behavioral roles for males and females. To the degree that individuals are sex-typed they should report feeling more comfortable with that domain of behaviors that is sex appropriate, and uncomfortable with behaviors that are socially defined as inappropriate for their sex. Conversely, the cross sex typed individual--the feminine male or masculine female--should feel more comfortable with cross sex typed behaviors and less comfortable with sex appropriate behaviors relative to members of their own sex. An individual's score on the STAT M scale indicates a relative degree of comfort on stereotypically masculine everyday activities or tasks.

General stereotypes about sex and behavior provided the basis for item selection for the STAT. As we have seen, both Bem and Spence and Helmreich have argued for this approach with regard to trait items. Bem required that items be rated by both male and female raters and that both must

agree on the sex typed social desirability of the item in order for it to be used. Spence and Helmreich used different item selection procedures in terms of the types of ratings, but adhered to the same general approach. This was seen as an advance over the use of actual sex differences. Stereotypes represent a general cognition about any domain of items. The approach used to measure masculinity and feminine orientations to activities assumes in effect that these stereotypes about the appropriateness of a certain group of behaviors also govern the self-concepts and behaviors of individuals. The study of sex stereotypes made by Rosenkrantz et. al. (1968) suggested that indeed males and females often differ in their self-ratings in predictable ways on traits that are collectively judged to differ for the typical man and woman, reflecting to some degree the veridicality of those stereotypes. As noted, others have been critical of this approach. Pedhazur and Tetenbaum argued that it is atheoretical, and Locksley and Colten have argued that global stereotypes may not be used to govern individual behavior. In addition, questions have been raised with regard to attribution: do the stereotypes mean the same things when self-ratings are made as when some general referent is being rated? Despite these objections, it was felt that assessment of general stereotypes about the types of everyday behaviors employed as items would provide a reliable, meaningful, and unbiased source of items for a meas-

ure of sex typed behavior. Consequently, the following hypothesis was formed before developing the test:

Hypothesis I: It was hypothesized that a group of items could be selected that both male and female raters would see as significantly more comfortable for one sex than the other.

Using behaviors that are stereotyped by sex differs in a number of ways from using traits. In the first place, social desirability, the criterion for choosing positive traits for the BSRI and PAQ, is not an issue with behaviors. In the second place, behaviors are not abstractions like traits, but instead are governed by social norms and sanctions. Although there was never any question that M and F would be treated as separate constructs, the question arose whether it would be necessary or desirable, to build separate scales for males and females. The matter of comparability between sexes was not seen as an issue since the primary purpose of sex role orientation measures is to discriminate between levels within sex. It became apparent at the planning stage that many potential items might differentiate between levels of masculinity or femininity within one sex but not within the other. For example, a less feminine woman might report feeling uncomfortable or awkward "wearing pantyhose" or "wearing lipstick", while a more feminine woman would report feeling very comfortable doing either. But such an item is irrelevant to measuring femininity in males,

almost all of whom would feel extremely awkward doing either. The decision was made to begin by investigating stereotypes and then to try to build an instrument that could be used for either sex. It was hoped that if enough items were collected which could potentially apply to either sex, the total scores would demonstrate sufficient variability to accurately discriminate between levels of masculine and feminine orientation without having to use separate scales. This can be framed as a hypothesis:

Hypothesis II: It was hypothesized that a single test could be created for use with both sexes.

As with other measures which treat masculine and feminine dimensions separately, certain kinds of empirical relationships between masculinity, femininity and gender were anticipated.

Hypothesis III: It was hypothesized that males would show generally higher masculinity scores than females and that females would show generally higher femininity scores than males.

Although it has been repeated that no specific theory exists which predicts what the precise relationship should be between masculinity and femininity scores, previous scales have shown generally orthogonal relationships between the two constructs in both sexes of respondents.

Hypothesis IV: It was hypothesized that STAT M and STAT F scores would also be orthogonal to one another both for males and females.

As with any new measure, it was desired to demonstrate the reliability and the internal consistency of the test. Following the model presented by Pedhazur and Tetenbaum it was thought that the most desirable outcome for the test in terms of factor analysis would be the demonstration of two relatively large factors which included masculine and feminine items respectively. This would indicate the separation of the two scales and the unidimensionality of each scale.

Hypothesis V: A two factor structure will derive from this test conforming to the division of activities into male and female categories and accounting for a large portion of the common variance among the items.

ITEM SELECTION AND TEST DEVELOPMENT

In this chapter, we will examine the results of the stereotype ratings which led to the selection of items for the Sex Typed Activities Test. Within the context of this chapter, the first two of the five hypotheses described in the previous chapter will be examined. These stated that first, a group of items was sought which would be seen by both males and females as stereotypically more comfortable for one sex than the other. Second, if possible, it was desired that a single test be created which could be used by both male and female respondents. This means that items chosen for the final scale should be relatively comfortable for either sex although stereotypically regarded as more comfortable for one sex than the other. In this chapter, a comparison of the stereotype ratings by male and female judges is made. The process of item generation and selection to meet these criteria is then explained.

Method

Item generation

The item selection procedure for the Sex Typed Activities Test began with the collection of potential items. A handout was created which asked for examples of everyday behaviors that might be considered masculine, feminine, or neutral. A facsimile of this handout is given in Appendix B, titled "Suggestion Form." This handout was given to coworkers, graduate students, and undergraduates. Thirty-seven people contributed suggestions. When duplications were eliminated, a list of over 282 potential items resulted. A reduced list was developed by eliminating on a rational basis those items which were too broad in scope, too narrow or esoteric, and those which were too closely tied to one sex or another. The 209 items which were retained made up the pool from which the STAT items were drawn. The eventual selection of STAT items depended directly on the ratings for the stereotypical features of these items.

Subjects.

The stereotype ratings were made by 101 female and 59 male undergraduate students in Introductory Psychology and Introduction to Personality courses in exchange for course credit. Unfortunately, twenty cases had to be eliminated due to a problem with the machine-scored answer sheets. Consequently, analyses were based on the data from the remaining 53 male and 87 female judges (N = 140).

Procedure.

Questionnaires were administered in class. Each subject or judge was to rate each of 209 behavioral items on the following seven-point Likert scale:

1	2	3	4	5	6	7
Very uncomfortable or very awkward					Very comfortable; not awkward at all	

There were three groups within each sex. One group rated the typical American male; another, the typical American female; and the third, the typical American adult. No subject rated more than one target. A facsimile of the questionnaire is reproduced in Appendix B ('Stereotype Rating Form'). This procedure produced six sets of ratings for each item due to the crossing of the sex of rater factor with the target (M,F,A) factor.

Results

The item selection process involved three separate steps. First, a preliminary analysis of item stereotypes was made. Second, items were retained from the preliminary list on the basis of whether they met minimal criteria for comfort for either sex. Third, items were deleted where clear disagreement about the stereotype could be shown between male and female judges.

Item stereotypes.

In order to determine whether males and females actually agreed on general stereotypes about activities and their sex-typing, it was first necessary to divide the total item pool into identifiable clusters. A preliminary division of the 209 initial items into broad masculine, feminine, and neutral activities was accomplished by simply pooling male and female respondents' ratings and computing 209 separate ANOVA's. An alpha level of .01 was used as the basis for assigning items to the Neutral, Masculine, or Feminine groups. The purpose of the statistical tests was to sort the items into such groups and not to test hypotheses in the conventional sense. Therefore, the implications of "false positives" is relatively inconsequential.

Where no significant difference appeared among the target conditions (M,F,A), the item was relegated to the neutral pool. Where significance beyond the .01 level was acheived and the mean rating for the "typical male" was highest, the item was assigned to the masculine group. Where the female mean rating was highest, and significance acheived, the item was assigned to the feminine group. In this way, 70 masculine, 79 feminine, and 60 neutral items were distinguished. Tables 2, 3, and 4 list these items with their respective significance levels and the amount of variance accounted for by the analysis of variance (eta^2). Inspection of these tables reveals that the initial categor-

ization did an adequate job of distinguishing broad classes of behaviors that relate to male vs. female family roles, tasks, and activities.

A feature of these tables deserving attention is the order of the observed means, listed in the last column of these tables under the title 'Trend'. Here, one sees a subtle contrast between masculine and feminine items. Even though quite often the neutral adult rating would fall much closer to one of the sex-typical mean ratings than to the other, there is a striking consistency to the order of the means on the feminine items (Table 3), which is lacking on the masculine items (Table 2). On the feminine items, the order of the means is always the same: the highest being the female target mean, the lowest the male target mean, and the adult mean falls somewhere in between these two. In contrast, with the masculine items, (Table 2) there are 23 instances out of 70 where the order varies and the lowest average rating is for the adult, not the female. Parenthetically, subsequent t-tests revealed that in all these cases the male target and female target means did differ significantly justifying the assignment of these items to the 'F' pool. Nevertheless, such a difference between the results for masculine and feminine items would not appear to be due to chance alone. Further analyses will shed some light on this observation.

TABLE 2

Preliminary Selection of Masculine Items

F-tests, Significance levels, and proportion of total variance accounted for (eta-squared)

	Item	F	p	eta^2	Trend*
7	Lifting weights	42.97	.0000	.42	MAF
15	Replacing a washer in a leaky faucet	39.66	.0000	.40	MAF
14	Using an electric drill	39.66	.0000	.39	MAF
23	Replacing the plug on an electric cord	37.37	.0000	.37	MAF
196	Chopping firewood	35.25	.0000	.37	MAF
126	Picking up a hitchhiker	30.36	.0000	.34	MAF
175	Using a power saw	31.31	.0000	.34	MAF
138	Changing the car's oil	30.76	.0000	.34	MAF
142	Building a dog house	28.86	.0000	.32	MAF
189	Changing a car's air filter	27.38	.0000	.31	MAF
52	Changing a tire	26.68	.0000	.30	MFA*
173	Installing a window air conditioner	26.28	.0000	.30	MAF
109	Repairing a toaster	25.74	.0000	.30	MAF
31	Assembing a bicycle for a child	23.20	.0000	.28	MAF
101	Using a snowblower	23.57	.0000	.28	MAF
153	Cleaning the rain gutters on a house	23.19	.0000	.27	MAF
151	Reading Playboy	20.63	.0000	.25	MAF
30	Reading the sports page	20.63	.0000	.25	MAF
209	Building a simple table	19.95	.0000	.25	MAF
179	Watching football on TV	19.83	.0000	.24	MAF
38	Going camping alone	19.22	.0000	.24	MAF
34	Getting rid of a dead mouse	18.21	.0000	.23	MAF
97	Watching basketball on TV	19.18	.0000	.24	MAF
53	Playing poker	18.08	.0000	.23	MAF
47	Pruning a tree limb	17.83	.0000	.22	MFA*
56	Using a tiller to plow up a garden	17.81	.0000	.22	MFA*
65	Changing a fuse	17.63	.0000	.22	MAF
202	Riding a motorcycle	17.35	.0000	.22	MFA*
133	Building shelves	17.46	.0000	.22	MAF
78	Playing softball	16.53	.0000	.21	MFA*

	Item	F	p	eta^2	Trend*
166	Driving a boat	16.58	.0000	.21	MAF
105	Going fishing	16.24	.0000	.21	MAF
121	Going to a bar alone	15.50	.0000	.20	MAF
176	Checking the oil in a car	14.85	.0000	.19	MAF
125	Buying a new car	13.90	.0000	.19	MAF
172	Painting the house	13.40	.0000	.18	MAF
157	Going on a trip alone	13.07	.0000	.18	MFA*
117	Playing pool	12.42	.0000	.17	MAF
106	Driving a car with a stick shift	12.51	.0000	.17	MFA*
80	Using a hammer	12.58	.0000	.17	MAF
207	Trimming a hedge	12.16	.0000	.16	MFA*
102	Driving a pick-up truck	11.96	.0000	.16	MFA*
156	Climbing a tall ladder	11.86	.0000	.16	MFA*
17	Driving a sports car	11.79	.0000	.16	MFA*
76	Using a screwdriver	11.66	.0000	.16	MAF
152	Swearing	11.07	.0000	.15	MAF
92	Building a model plane	10.89	.0000	.15	MAF
208	Pumping your own gasoline	10.33	.0001	.14	MFA*
116	Buying a used car	10.01	.0001	.14	MAF
122	Drinking a beer	9.67	.0001	.14	MAF
183	Opening a tight jar lid	9.82	.0001	.14	MAF
145	Playing touch football	9.67	.0001	.13	MAF
16	Mowing the lawn	9.30	.0002	.13	MAF
87	Reading the business page	8.90	.0002	.13	MFA*
161	Picking up the tab in a restaurant	8.64	.0003	.12	MFA*
192	Buying car insurance	8.53	.0003	.12	MFA*
171	Shovelling a sidewalk	8.46	.0004	.12	MAF
37	Starting a fire in the fireplace	8.07	.0005	.12	MAF
35	Computing your income tax	7.95	.0006	.11	MFA
88	Jogging	6.97	.0013	.10	MFA*
82	Sharpening a knife	6.63	.0018	.10	MFA*
57	Shaking hands	6.56	.002	.10	MAF
185	Playing catch with a kid	6.37	.0023	.09	MAF
184	Driving a car	5.97	.0034	.09	MFA*
162	Handling family finances	5.77	.004	.09	MFA*
204	Planting a tree	5.52	.0051	.08	MAF
66	Writing a check	5.14	.0072	.08	MFA*
114	Letting your spouse cook dinner for you	4.83	.0096	.07	MFA*
79	Making a bank deposit	4.81	.0097	.07	MFA*

Note: Table 2 compiled over 128 subjects.
Degrees of freedom for the SS_B are 2.
Due to missing data, the degrees of freedom for the SS_W varies between 120 and 126.

Items in this table are arranged in descending order according to the value of eta^2, which represents the proportion of total variance (both linear and non-linear) held in common with the independent variable.

*Trends represent the order of the means for the typical male, the typical female, and the typical adult conditions in descending order.

TABLE 3

Preliminary Selection of Feminine Items

F-tests, Significance levels, and proportion of total variance accounted for (eta-squared)

	Item	F	p	eta^2	Trend
84	Buying eye make-up	77.43	0	.45	FAM
112	Shaving your legs	72.31	0	.55	FAM
147	Having a shower party	51.21	.0000	.46	FAM
74	Crocheting	48.48	.0000	.45	FAM
46	Crying over a TV show	43.06	.0000	.42	FAM
200	Wearing high-heeled shoes	42.44	.0000	.41	FAM
129	Going shopping for hours	37.97	.0000	.38	FAM
170	Embroidering	36.27	.0000	.37	FAM
137	Using hairspray	36.06	.0000	.37	FAM
148	Planting a flower garden	36.50	.0000	.37	FAM
61	Shopping for children's clothing	35.07	.0000	.36	FAM
113	Buying linens	34.98	.0000	.36	FAM
26	Getting your hair curled	32.22	.0000	.17	FAM
103	Carrying a packet of Kleenex	32.42	.0000	.22	FAM
130	Hugging a friend	31.38	.0000	.34	FAM
64	Planning a menu	30.67	.0000	.33	FAM
195	Buying new dishes	29.46	.0000	.33	FAM
9	Knitting a scarf	28.39	.0000	.32	FAM
167	Mending socks	28.71	.0000	.32	FAM
174	Dusting a table	28.84	.0000	.32	FAM
190	Changing a baby's diaper	27.95	.0000	.31	FAM
96	Sewing	27.96	.0000	.31	FAM
18	Putting the flowers in a vase	27.03	.0003	.31	FAM
41	Packing for other family members	23.85	.0000	.28	FAM
13	Setting the table	23.98	.0000	.28	FAM
40	Baking a cake from a mix	23.96	.0000	.28	FAM
85	Dyeing your hair	23.08	.0000	.28	FAM
136	Taking dictation	22.84	.0000	.27	FAM
141	Cleaning house	22.82	.0000	.27	FAM
110	Buying a wedding gift	22.64	.0000	.27	FAM
139	Picking flowers	22.19	.0000	.27	FAM
33	Re-potting a plant	21.98	.0000	.26	FAM
93	Bathing a baby	21.63	.0000	.26	FAM
140	Reading a clothing magazine	21.06	.0000	.26	FAM

	Item	F	p	eta^2	Trend
58	Babysitting for money	20.69	.0000	.25	FAM
44	Wrapping a present	20.24	.0000	.25	FAM
8	Ironing a shirt	19.41	.0000	.24	FAM
123	Sorting laundry	19.32	.0000	.24	FAM
178	Making dinner for company	19.18	.0000	.24	FAM
150	Reading Playgirl	17.16	.0000	.22	FAM
169	Helping a child get ready for school	17.44	.0000	.22	FAM
99	Shampooing a child's hair	17.30	.0000	.22	FAM
135	Changing sheets	16.60	.0000	.21	FAM
146	Making the bed	15.95	.0000	.20	FAM
124	Washing clothes	14.99	.0000	.19	FAM
6	Re-arranging the furniture	14.80	.0000	.19	FAM
199	Giving a bottle to a baby	14.49	.0000	.19	FAM
11	Doing crafts	13.99	.0000	.19	FAM
72	Planning a party	14.40	.0000	.19	FAM
29	Scrubbing a floor	13.76	.0000	.18	FAM
22	Crying in private	12.83	.0000	.17	FAM
107	Folding clothes	12.36	.0000	.17	FAM
132	Writing letters	12.12	.0000	.16	FAM
182	Cleaning a stove	12.06	.0000	.16	FAM
198	Reading to a child	11.40	.0000	.16	FAM
104	Getting your hair styled	11.25	.0000	.15	FAM
39	Weeding a flower bed	11.38	.0000	.15	FAM
188	Washing the dishes	10.99	.0000	.15	FAM
4	Cleaning the bathtub	10.92	.0000	.15	FAM
163	Getting up with a baby at night	10.57	.0001	.15	FAM
25	Comforting a child	10.62	.0001	.15	FAM
20	Shopping for clothes	10.52	.0001	.15	FAM
205	Defrosting the refrigerator	10.15	.0001	.14	FAM
203	Mopping the floor	10.05	.0001	.14	FAM
3	Buying a gift for your mother	9.91	.0001	.14	FAM
45	Sending an anniversary card	10.14	.0001	.14	FAM
160	Sunbathing	9.66	.0001	.14	FAM
149	Replying to an invitation	9.45	.0002	.13	FAM
5	Typing a letter	8.87	.0003	.13	FAM
197	Reading a gossip column	8.68	.0003	.12	FAM
119	Taking a child to the dentist	8.19	.0005	.12	FAM
158	Going to the PTA	7.99	.0005	.11	FAM
98	Picking up a child at school	6.39	.0023	.09	FAM
86	Talking to a kid's teacher	6.33	.0024	.09	FAM
194	Weeding a garden	6.15	.0028	.09	FAM
193	Using a blow-dryer	6.01	.0032	.08	FAM
73	Taking a bath	5.73	.0042	.08	FAM
120	Reading the Society page	5.59	.0047	.08	FAM
71	Making coffee	5.23	.0066	.08	FAM

Note: Table 3 compiled over 128 subjects.

Degrees of freedom for SS_B are 2.
Due to missing values, degrees of freedom for the SS_W vary from 120 to 126.

Items in this table are arranged in descending order according to the value of eta^2, which represents the proportion of total variance (both linear and non-linear) held in common with the independent variable.

* Trends represent the order of the means for the typical male, the typical female, and the typical adult, in descending order.

TABLE 4

Preliminary Selection of Neutral Items

F-tests, Significance levels and Proportion of total variance accounted for (eta-squared).

	Item	F	p	eta^2	Trend
1	Reading science fiction	1.29	.28	.02	
2	Combing your hair	4.49	.01	.07	FAM
10	Getting the mail	.77	.46	.01	
12	Walking the dog	1.18	.31	.02	
19	Visiting a friend in the hospital	1.56	.21	.02	
21	Barbecuing ribs	1.35	.26	.02	
24	Peeling an orange	2.04	.14	.03	
27	Eating a steak	2.84	.06	.04	
28	Listening to music	2.04	.13	.03	
32	Watching television	.87	.42	.01	
36	Setting up appointments	2.41	.09	.04	
42	Packing your suitcase	1.63	.20	.03	
43	Being a volunteer	2.69	.07	.04	
48	Laughing at a cartoon	.01	.98	.00	
49	Brushing your teeth	1.02	.36	.02	
50	Cutting hair	4.70	.01	.07	FMA
51	Writing checks to cover bills	3.33	.04	.05	MFA
54	Balancing a checkbook	3.02	.05	.05	
55	Opening a door for someone else	3.64	.03	.06	MAF
59	Raking leaves	2.99	.05	.05	
60	Planting a vegetable garden	3.75	.03	.06	FMA
62	Washing the car	4.77	.01	.07	MFA
63	Sweeping the driveway	2.04	.13	.03	
67	Going dancing	3.39	.04	.05	FAM
68	Looking for a new job	3.18	.04	.05	MFA
69	Carrying a handkerchief	.54	.58	.01	
70	Feeding the dog	.90	.41	.01	
75	Reading a murder mystery	.43	.65	.01	
77	Going to the laundromat	3.38	.04	.05	
81	Using deodorant	.75	.47	.01	FMA
83	Taking a snapshot	2.71	.07	.04	
89	Choosing a paint color	2.23	.11	.03	
90	Watching the news on TV	1.65	.20	.03	
91	Talking about sex	.57	.57	.01	
94	Move a couch	1.57	.21	.02	
95	Using cologne	2.13	.12	.03	

	Item	F	p	eta^2	Trend
100	Playing Monopoly	.26	.78	.00	
108	Reading the newspaper	.29	.75	.00	
111	Playing Bridge	3.58	.03	.05	FMA
115	Riding a bicycle	.33	.72	.01	
118	Going to a party	3.29	.04	.05	MFA
127	Loading a car for a trip	4.76	.01	.07	MFA
128	Making ice cubes	.07	.93	.00	
131	Painting a door	2.86	.06	.04	
134	Going to the movies	1.31	.27	.02	
143	Playing with dominoes	.29	.75	.00	
144	Locking a door	.87	.41	.01	
154	Taking prescription medicine	2.02	.14	.03	
155	Taking a shower	1.28	.28	.02	
159	Buying a record	2.02	.14	.03	
164	Taking out the garbage	1.19	.31	.02	
168	Cleaning out the garage	3.39	.04	.05	MAF
177	Oversleeping	.36	.70	.01	
180	Going to the dentist	4.20	.02	.06	FMA
181	Making a phone call	2.43	.09	.04	
186	Hanging pictures	2.71	.07	.04	
187	Cleaning a fish tank	3.02	.05	.04	
191	Playing tennis	1.75	.18	.03	
201	Sharpening a pencil	.97	.38	.01	
206	Asking someone for help	4.04	.02	.06	FMA

Note: Table 4 compiled using data from 128 subjects.
Degrees of freedom for SS_B are 2.
Due to missing values, degrees of freedom for the SS_W vary from 120 to 126.

*Trends represent means in descending order for the typical male, typical female and typical adult conditions. Trends are listed only where $p < .05$.

The division of the initial item pool into masculine, feminine, and neutral items affords the opportunity for an examination of the level of agreement between the sexes about the relative degree to which particular items are related to gender and/or gender role. Two different analyses were completed to detect the level of agreement about the stereotypical features of the 209 items and the general degree of similarity between males and females regarding their sex stereotypes about these behaviors. In the first of these analyses, the correlations between the average ratings given the 209 items for the different target groups are examined to determine the relative degree of agreement between male and female raters about each of the items. In the second analysis, average ratings for each subject on the masculine and feminine items are analyzed for systematic differences.

Mean ratings were computed for the six groups of judges for each of the 209 questionnaire items. Table 5 provides the correlations among the ratings given by the six groups of judges for each of the 209 items. The correlations between the male and the female raters for the typical male (r = .93), and the typical female (r = .88), are both very high indicating a high degree of concordance between the male and the female raters about which items would be more comfortable for males and which should be more comfortable for females. In contrast, correlations within each sex for the

male and female targets are virtually orthogonal, though in both cases, negative in sign.

The ratings for the neutral adult target appear to be strongly related to the sex of the rater. Between male and female raters, the correlation for the adult target was .56. However, the adult ratings made by male judges are strongly correlated to both the ratings made for the typical male by male raters (r = .85) and also to the ratings made for the typical male by the female raters (r = .77). In symmetrical fashion, the mean ratings on the items for the adult target given by females most clearly resembled the ratings made for the female target, whether the people making the ratings were female (r = .89) or male (r = .80).

The correlation of .56 for the adult ratings made by males and females represents a significantly lower level of agreement about the items than the other correlations mentioned. The lowest of the correlations listed above, .77, was compared to the .56 level and found to be significantly higher, $t(206) = 15.41$, $p < .0001$. This suggests that as a general rule, the average item ratings given by male raters for the typical adult resemble a general stereotype about men, while the ratings given by females for the typical adult more closely resemble general stereotypes about women. Furthermore, the ratings given by a particular sex (about the adult target) are more closely correlated to general stereotypes for that particular sex than are the ratings

TABLE 5

Intercorrelations of Mean Ratings for 209 Items

	Male Raters		Female Raters		
	Typical Female	Typical Adult	Typical Male	Typical Female	Typical Adult
Male Raters					
Typical Male	-.17**	.85***	.93***	.05	.26***
Typical Female		.18**	-.28***	.88***	.80***
Typical Adult			.77***	.37***	.56***
Female Raters					
Typical Male				-.10	.14
Typical Female					.89***

*p <.05.
**p <.01.
***p <.001.

given by the two sexes for the same target, when the sex is unspecified.

In order to better examine the relationships between sex of rater, target object, and the type of item, mean ratings were computed over the 70 masculine and 79 feminine items for each of the 140 respondents, yielding a single M rating and a single F rating for each rater or judge who completed the questionnaire. The summary means and standard deviations of these average masculine-item and feminine-item ratings appear in Table 6. The range of the means varies from a minimum of 3.47 (slightly below the scale midpoint of 4.0), up to 6.04, which approaches the upper boundary of the scale. Of the 12 means, only three fall below the midpoint of the comfort rating scale which would seem to indicate that the items on these lists were perceived to be fairly comfortable across the board. These averages were subjected to a 2 x 3 x 2 analysis of variance (Sex of Rater x Male, Female or Adult target x Type of Item (masculine/feminine)). The results of this analysis are reported in Table 7. They reveal two significant main effects, two first order interactions, and a second order interaction which subsumes the lower order effects. These effects will be described separately.

The first of the two main effects was observed for target (rating group), $\underline{F}(2,132) = 3.79$, $\underline{p} = .026$. The mean comfort ratings across both masculine and feminine items

TABLE 6

Mean ratings on Masculine and Feminine Items

	n	MEAN	STANDARD DEVIATION
MASCULINE ITEMS			
Typical Male			
Male Raters	17	5.86	0.67
Female Raters	29	6.04	0.89
Typical Adult			
Male Raters	18	4.90	0.84
Female Raters	28	4.16	0.95
Typical Female			
Male Raters	18	3.93	0.94
Female Raters	28	4.56	1.14
FEMININE ITEMS			
Typical Male			
Male Raters	17	3.82	0.71
Female Raters	29	3.47	1.07
Typical Adult			
Male Raters	18	4.04	0.86
Female Raters	28	5.00	1.12
Typical Female			
Male Raters	18	5.36	0.99
Female Raters	28	5.90	0.64

TABLE 7

Analysis of Variance

Source	SS	df	MS	F	p
Sex	2.69	1	2.69	2.66	.11
Target	7.68	2	3.84	3.79	.03*
S x T	5.15	2	2.58	2.54	.08
Error	133.62	132	1.01		
Item-type	6.19	1	6.20	11.68	.001**
S x I	2.11	1	2.11	3.99	.048*
T x I	151.24	2	75.62	142.56	.001***
S x T x I	15.15	2	7.57	14.28	.001***
Error	70.02	132	0.53		

Note:
Sex indicates the sex of rater.

Target indicates the rating condition. Separate groups rated the typical male, female, and adult.

Item type indicates whether the score is an average of the 70 preliminary masculine items or the 79 preliminary feminine items. (M v F)

* $p < .05$
** $p < .01$
*** $p < .001$

were significantly higher for the typical female (M = 4.94) than for the typical adult (M = 4.53). The average rating for the typical male (M = 4.80) fell between the other two means. A Newman-Kuels analysis demonstrated the comparisons between the middle group (typical male) and each of the higher and lower group means (female and adult) to be non-significant. A highly significant main effect was also observed for the Type of item, $F(1,132) = 11.68$, $p = .001$. Masculine items (M =4.91) were rated significantly more comfortable than feminine items (M = 4.6).

The first of two lower order interactions was observed between the Sex of rater and the Type of item. A test of simple effects indicated that while ratings made by females for the masculine and feminine items across all 3 target groups were comparable, $F(1,132) = .03$, n.s., male raters gave significantly lower comfort ratings to the feminine items than to the masculine items, $F(1,132) = 9.04$, $p < .005$.

The second of the simple two-factor interactions occurred between Target (typical male, female, adult) and Type of item (masculine vs. feminine). As indicated by Figure 1, the typical male was rated high on comfort for masculine items and low on feminine items. The reverse pattern also occurred with the typical female rated high on comfort for feminine items and low on comfort for masculine items. The typical adult ratings fell in between the extremes. This

overall pattern is not surprising in light of the fact that the basis for assigning items to the masculine and feminine subgroups was a significant difference in mean ratings among the target groups.

The meaning of all of these subsidiary effects is altered in light of the overall interaction among all three factors (Sex of rater, Target, and Type of item) which was highly significant, $F(2,132) = 14.28$, $p < .001$. In Figure 1, the individual means have been numbered to facilitate the discussion of the pattern of results. A Newman-Kuels analysis over all twelve means was conducted. The results indicate an interesting patterning of mean ratings depending on the sex of the rater and the sex for which the ratings were being made. They indicate that there is overall agreement about the masculine and feminine items and their relationship to sex, the basis for further item selection, but that some differences in perspective do exist.

The mean ratings given for the typical male were consistent for both male and female raters. By judges of either sex, the typical male was seen as significantly more comfortable with the stereotypically masculine activities ($p < .01$). Mean ratings for the masculine items given by males and females (Means #10,#12) were not significantly different. However, the mean ratings for the typical male on the feminine items did differ significantly depending of the sex of the rater (#1,#2). Female raters estimated the

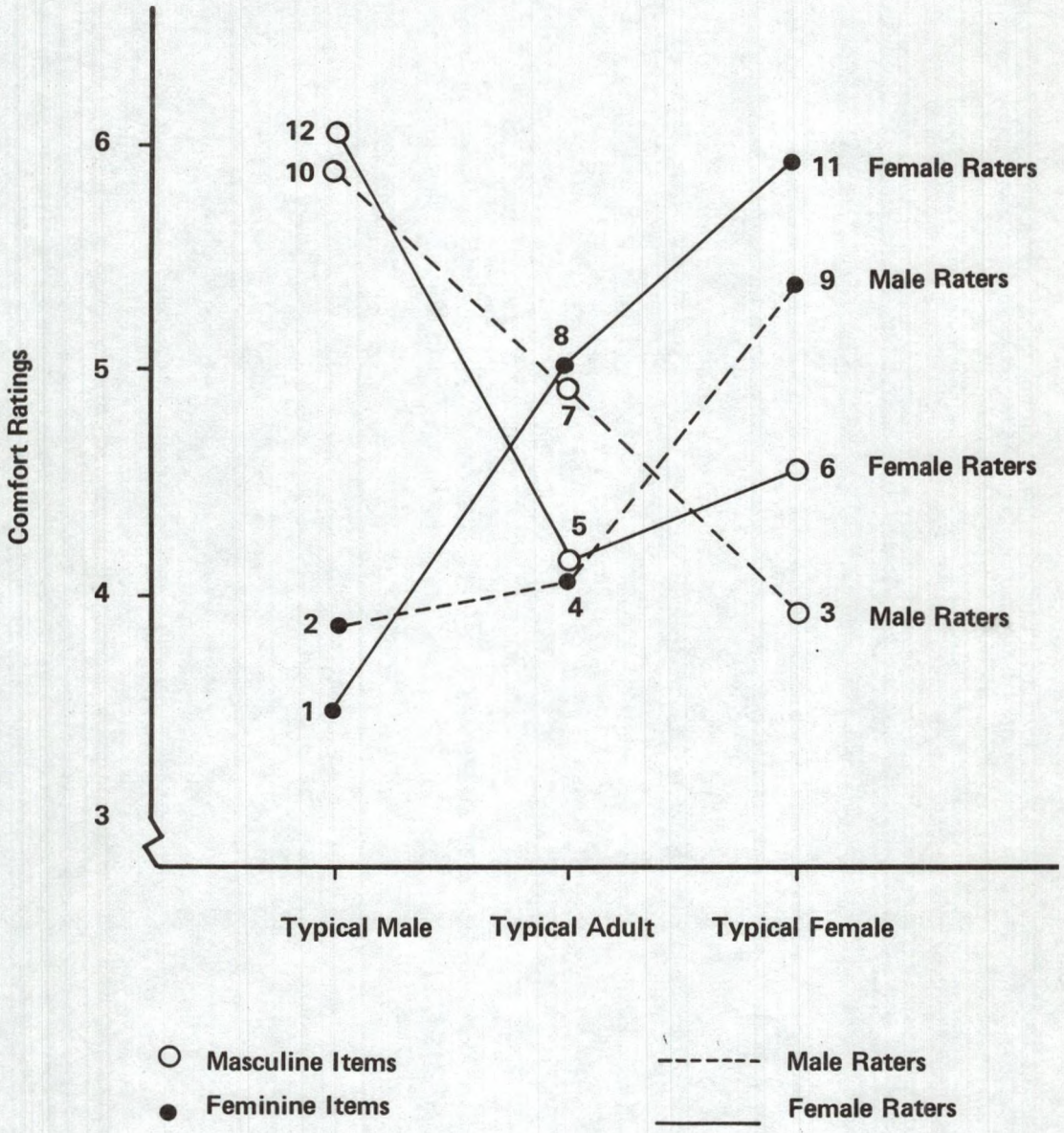


Figure 1. Three-way analysis of variance: Mean stereotype ratings for masculine and feminine items for three targets.

typical male to be less comfortable with feminine behaviors than male raters.

As we have already noted regarding the matrix of intercorrelations among the average item ratings (Table 5), males gave ratings for the typical adult which resembled the male ratings for the typical male (r = .85). Female ratings for the typical adult similarly resembled those made for the typical female (r = .89). This tendency to identify the neutral adult target with one's own sex can be seen graphically in Figure 1. Comfort ratings for the typical adult are symmetrical. The ratings are relatively high when subjects are rating those (sex-congruent) items which are most comfortable for their own sex, relatively low when rating the items more comfortable for the opposite sex (sex-incongruent items). The male rater/masculine item mean for the typical adult condition does not differ significantly from the female rater/feminine item mean (#7,#8). Neither do the sex-incongruent means (i.e., males rating feminine items, females rating masculine items) differ to a significant degree (#4,#5). However, sex congruent and sex incongruent means do differ from one another. This pattern illustrates the general pattern of responding to the adult condition as if the "adult" were of one's own sex. Adult ratings, do however, show some regression from the more extreme sex-typed ratings.

The ratings for the typical female (#3,#6,#9,#11) are all significantly different from one another at the .01 level. Both sets of judges saw the typical female as more comfortable with regard to the feminine items than with regard to masculine items. However, males gave significantly lower ($p < .01$) ratings to the typical female on both masculine and feminine items than did female judges. In other words, females saw the typical female as much more comfortable with either masculine or feminine activities than did their male counterparts.

Preliminary Discussion of Stereotype Ratings

Probably the single most important piece of information to be gained from this analysis is the observable contrast between the overall patterns of responses given by males and females. Male judges gave the predictable "X" pattern: a sequence of means for the masculine items which had the typical male most comfortable, the adult less so, and the typical female least comfortable; the X is completed by the reverse sequence for the feminine items. Females also had the F-A-M descending order for the feminine items, indicating that their stereotype is one of relative comfort for females on such items and relative discomfort for males. On the masculine items, however, the ratings of females for the three targets differed not only in terms of degree, but also in terms of the order of the means. The mean for the typi-

cal female on masculine items (# 6) is actually higher than the corresponding mean for the typical adult (# 5) to a significant degree ($p < .01$). Perhaps it is not surprising that females rate the typical female as more comfortable with the masculine items than do the male raters. But if we assume that the tendency of the adult means to cluster towards the center reflects the recognition that the adult could be of either sex, it might be surprising that in this case female raters appear to see either sex as more comfortable than the typical adult of undisclosed sex.

In general, the ratings given to the neutral adult reflect the stereotypes about the rater's own sex. Knowing this, it would appear that on the masculine items the female judges adjusted their ratings upwards, as if to contradict the stereotype. This seems to be the most likely explanation of these data, although other possibilities might be entertained. The idea that the sexes simply have different notions about the competencies of the two sexes on a variety of activities and tasks would seem to be contra-indicated by the lower rating given the adult target by female judges, but it might account for a certain amount of the difference. The other possibility, that of a chance difference, is similarly dubious. A chance difference might lead to no difference between the adult and female means, but a difference in an unexpected direction seems less likely.

If a different pattern of ratings for masculine and feminine items is observed depending on the sex of the raters, a pertinent question can be raised about the interpretation of the data. The differences in ratings made by males and females could either represent actual differences in stereotype beliefs or systematic differences in the reporting of stereotype beliefs. It seems plausible, looking at the overall pattern of results, that the tendency of female rater/female target ratings to be higher than expected represents an effort on the part of the female raters as a group to give the typical member of their own sex the "benefit of the doubt" on items that are sex-incongruent. The general similarity of adult ratings to same-sex ratings encourages such an interpretation: in every other case adult ratings resemble the stereotypes for the raters' own sex but are closer to the middle of the range.

Of course, such an interpretation is only hypothetical; it would require further experimental investigation to demonstrate how much of the difference relates to actual stereotype differences versus how much relates to differences in the reporting of stereotypes.

In spite of this unexpected interaction and the problem of interpreting it, the analysis performed above gives significant support to the hypothesis that males and females do agree about which activities are stereotypically more comfortable for one sex than the other. Correlational data in-

dicated high levels of covariance between the ratings made by the two sexes on the 209 items for the typical male and the typical female. Analysis of variance does illuminate a potential difficulty in using such ratings, however. The interaction observed implied that females tend to give the "typical female" surprisingly high ratings on the overall group of masculine items, while males tend to give the typical female lower ratings on both the masculine and feminine items than do female judges. This interaction chiefly concerns the masculine items. There are still significant differences between female judges' evaluations of the typical male and the typical female on the bulk of these items. These differences in stereotype are simply not as great or as consistent as those reported by males.

Item Selection

In spite of the observed differences between the pattern of ratings given by males and females, the general stereotype ratings were judged to distinguish sufficiently between clusters of items that are more appropriate or comfortable for males and those that are more comfortable for females. In the process of creating a self-rating scale, or pair of scales, this is the essential requirement. Both male and female raters saw the masculine items as more comfortable for females than males. This is tautological in view of the way in which the two groups of items were gener-

ated. But it does indicate that the pooling of male and female raters did not do any injustice to the data and served as an adequate method of dividing these items into meaningful clusters.

Unlike the PAQ or the BSRI, the Sex Typed Activities Test is based on stereotyped notions of relative comfort with everyday behaviors rather than notions of the desirability of traits. However, the STAT was designed to resemble these earlier tests in two ways. First, separate scales for masculine activities and feminine activities were planned. Second, for reasons of simplicity, a strategic decision was made to follow the lead of these two models in creating a single test for both males and females. The M and F scales on the PAQ and the BSRI were intended to contain only positive traits; that is, traits that could reasonably be endorsed by members of either sex. In the case of the PAQ, the problem of traits which were positively valued for one sex but relatively negative for the other sex was resolved by creating a third, bipolar, Masculinity-Femininity subscale which was composed of such items. This procedure was also adopted for use with other types of items by Orlofsky et. al. (in press) for the SRBS-2.

In the case of the STAT, a preliminary decision was made to create only two scales, M and F. If subsequent analyses dictated the creation of a third scale, it could be constructed and added to a revised form of the test. Since

only one test was to be used for either sex, it was also decided that only items which were relatively comfortable for either sex would be included. This was intended to be analogical to the use of only positive-valued traits on the BSRI and PAQ. Consequently, after the items were divided into Masculine, Feminine, and Neutral pools, the second step in test building involved distinguishing between the items which were relatively comfortable for either sex and those that were not.

An inspection of the individual item mean ratings for the male, female, and adult targets was made for the items on the masculine and feminine lists. These item means are tabled in Appendix B. As a preliminary measure, items on these lists which had an eta^2 value of .10 or less were deleted, to eliminate the least sex-typed of the items. Eta^2, again, is an indicator of total variance accounted for by an independent variable, in this case by sex stereotype.

On the masculine items, the typical male is usually rated high while considerable variance exists for the typical female. On the feminine items, the typical female is generally rated high while considerable variance exists for the typical male ratings. Consequently, all items were deleted which had an item mean of less than the scale midpoint (4.0) for the "inappropriate" sex. In other words, masculine items retained were those on which the mean rating for the typical male was significantly higher than the mean rat-

ing for the typical female and the mean rating for the typical female was at least 4.0 (moderately comfortable). Feminine items retained at this stage were those where the typical female mean rating was significantly higher than that of the typical male, and the mean rating for the typical male target was at least 4.0.

Interestingly, although there were originally more feminine items (79) than masculine items (70), after this procedure that pattern was reversed. Only 23 of the potential 79 feminine items were retained while 34 of the potential masculine items were retained. It appeared as if the raters felt that males would experience greater discomfort performing female activities than females would experience in performing male activities. Since it was deemed important to have a larger pool of feminine items at this stage, the criterion for the minimum typical male mean on the feminine items was relaxed. When a minimum of 3.5 was used instead of the 4.0 criterion (half a point below the scale midpoint on the comfort scale), a total of 38 feminine items were retained. In all, 72 items were judged to meet both the overall significant difference criterion and the minimal comfort criterion.

Another parenthetical observation relates to the fact that these items seemed to fall on a continuum with regard to their degree of relationship to sex. This fact is reflected in the significance levels obtained when contrasting

the male and female target, and also by the relative value of eta^2. In reading down the list in Tables 1 and 2, which are sequentially arranged according to the value of eta^2, one can sense the gradations in sex typing. It should also be noted that many of the "strongest" items showed high levels of eta^2 precisely because they were comfortable for one sex and not for the other, and mean ratings for the two sexes were therefore highly discrepant. As a consequence, in choosing items that were appropriate for either sex, many of these items were simply deleted at this stage. The implications of this type of selection process will be discussed in the final chapter which examines the overall results of the self-ratings in conjunction with the stereotype ratings.

The third and final step in selecting the items for the STAT was to eliminate any items on which the male and female raters showed clear disagreement about sex-stereotypy. In the creation of the BSRI, Bem (1974) performed multiple t-tests separately for the two sexes and selected only items which were significantly different for both sexes. In contrast, the preliminary division of items here has depended on ratings taken from all of the raters, i.e., with the two sexes pooled. While this permitted a more general analysis of the stereotype ratings, it includes in the M and F pools certain items about which males and females may fundamentally disagree, insofar as sex stereotypes are concerned. With this in mind, the 34 M and 38 F items retained after the

second step were subjected to a series of 2 X 2 (Sex of rater X Male/Female Target) Analysis of Variance procedures. In this case, both the "Adult" target ratings and the Neutral items were completely ignored. The results of these analyses are reported in Tables 8 and 9. As shown, the F-values for the Target main effects were all significant. All but one were significant below the .01 level, and the vast majority were significant beyond the .001 level of alpha.

Disagreement between the sexes about the stereotype items, however, was indicated by the F-value for the interaction term. There appeared to be many more disagreements on the F items than on the M items. An alpha level for the interaction term was set at .01 and items were deleted from the final list if they were significant beyond this level. This led to the deletion of three of the 34 masculine items and 11 of the feminine items. The final scales were to contain 31 masculine and 27 feminine items. Unfortunately however, after the STAT form was printed it was discovered that clerical errors had resulted in the unintentional retention of one masculine and two feminine items (#15, M; #22, #40, F) which should have been deleted from the final list of items. Three other items (#208, M; #39, #146, F) should have been included but were deleted in error. Given the number of items on the final form of the test these few errors were not considered sufficiently important to justify a

TABLE 8

M-items: Target by Sex of Rater Interactions

F-values for the Sex-of-Rater and Target group interactions as well as the main effects. Degrees of freedom are 1,95 in each of the 34 ANOVA's. Starred (**) items were deleted from the final STAT form.

	SEX		TARGET		S x T	
Item	F	p	F	p	F	p
16	8.56	.004	16.44	.000	3.96	.050
17	.60	.441	19.36	.000	.80	.373
30	16.61	.000	4.51	.000	1.05	.308
35	2.47	.119	4.24	.042	.04	.844
37	8.10	.005	22.20	.000	1.43	.235
47	3.65	.059	24.46	.000	.24	.629
53	1.98	.163	33.13	.000	.01	.943
65	3.27	.074	53.33	.000	1.04	.311
76	14.19	.000	27.37	.000	8.38	.005**
78	3.27	.074	32.30	.000	.03	.865
80	8.16	.005	33.08	.000	4.17	.044
87	4.03	.047	7.47	.007	.01	.918
88	2.14	.147	3.82	.053	.32	.571
97	6.73	.011	43.09	.000	.18	.675
102	2.08	.152	29.34	.000	.04	.834
105	25.36	.000	38.99	.000	5.14	.026
106	4.26	.042	26.44	.000	6.64	.001**
117	2.42	.123	29.02	.000	1.38	.243
122	4.92	.029	17.70	.000	1.98	.163
125	.59	.442	30.52	.000	.96	.330
145	.02	.886	22.38	.000	.20	.653
156	3.40	.068	15.36	.000	.02	.883
161	16.52	.000	13.17	.000	5.55	.021**
166	2.45	.121	33.70	.000	.30	.583
171	13.11	.000	18.03	.000	1.38	.243
172	11.55	.000	34.34	.000	.88	.352
176	5.99	.016	33.95	.000	2.96	.089
179	7.30	.008	52.51	.000	.00	.956
183	6.55	.012	17.95	.000	1.29	.258
192	.95	.333	18.79	.000	1.04	.311
202	3.75	.056	40.21	.000	.05	.830
207	1.78	.185	16.33	.000	.09	.769
208	7.63	.007	17.91	.000	4.83	.030
209	3.93	.050	46.09	.000	3.15	.079

TABLE 9

F-items: Target by Sex of Rater Interactions

F-values for the Sex-of-Rater and Target Group interaction as well as the main effects. Degrees of freedom are 1,95 in each of the 38 ANOVA's. Starred (**) items were deleted from the final STAT form.

	SEX		TARGET		S x T	
Item	F	p	F	p	F	p
3	.35	.553	17.99	.00	7.42	.008**
5	4.49	.028	10.20	.002	.00	.978
6	1.78	.186	21.18	.000	1.25	.267
11	.04	.849	44.57	.000	5.06	.027
13	2.86	.094	68.79	.000	.05	.818
20	.00	.947	35.48	.000	.33	.565
22	1.19	.277	43.49	.000	8.94	.004**
25	4.92	.029	28.72	.000	4.33	.040
33	.88	.350	74.73	.000	5.86	.017
39	1.47	.229	29.70	.000	5.16	.025
40	.09	.76	88.33	.000	7.39	.008**
44	1.93	.168	47.57	.000	5.49	.021
45	.00	.999	27.71	.000	6.88	.010**
58	.14	.712	68.10	.000	.39	.532
72	.49	.486	22.15	.000	.01	.936
93	2.51	.117	53.13	.000	1.05	.309
99	9.11	.003	67.66	.000	3.80	.054
104	.07	.787	20.64	.000	.45	.504
107	2.21	.14	40.25	.000	8.33	.005**
110	.23	.632	70.70	.000	4.57	.035
119	1.21	.275	14.86	.000	3.24	.075
123	.01	.942	43.88	.000	8.65	.004**
124	.26	.613	29.24	.000	3.36	.070
132	.00	.99	28.57	.000	.27	.605
135	1.43	.235	39.29	.000	5.52	.021
146	2.16	.145	34.18	.000	1.90	.171
149	5.98	.016	21.10	.000	2.74	.101
158	.06	.804	18.53	.000	1.34	.249
160	.00	.973	16.39	.000	.77	.383
163	3.90	.051	18.33	.000	1.59	.210
169	6.19	.015	37.34	.000	.54	.464
174	.05	.820	62.30	.000	8.04	.006**
188	1.16	.285	20.99	.000	3.36	.070
195	2.41	.124	85.56	.000	9.08	.003**
198	1.63	.204	26.92	.000	8.81	.004**
199	.00	.981	28.99	.000	7.01	.010**
203	.14	.710	25.13	.000	6.28	.014
205	.67	.415	28.06	.000	6.97	.010**

second printing. However, the error is noted to explain discrepancies that appear between Tables 8 and 9 and the final listing of items.

Summary. The item selection process was constituted by three steps: (1) the division of the items into masculine, feminine, and neutral pools; (2) the elimination of items that were relatively uncomfortable for either sex; and (3) the elimination of items which evoked significant disagreement on the part of the male and female raters. In this manner the final 58 items were selected and placed on a self-rating form using the same seven-point comfort scale with instructions to make self-ratings on the full complement of masculine and feminine items. It was expected that the result would be a test using behavioral items, analogous to the Bem Sex Role Inventory or the Personal Attributes Questionnaire, which would provide a measure of masculine and feminine dimensions of personality separately and independently. The analysis of the stereotype ratings which has been discussed at some length raised certain issues about the symmetry of sex-stereotype ratings by males and females--a source of possible difficulties in using stereotypes for the item selection process for masculinity and femininity scales. As pointed out at the beginning of this chapter, however, the litmus test of the item selection process is not in the analysis of the stereotype ratings itself. It is in the way the items perform when used as a test.

TEST STANDARDIZATION AND CORRELATES

The Sex Typed Activities Test is a self-rating instrument using 58 pre-selected masculine and feminine activities as items. Subjects are asked to rate the relative degree of comfort or discomfort they might experience in performance of these behaviors. In this chapter, the relative success of this approach will be analyzed. Both the internal characteristics of the test and the inter-relationships of the test with other theoretically related psychological dimensions are examined in detail.

To complete the initial form of the questionnaire, a group of questions regarding the respondent's background and other selected variables of interest were added to the STAT. Independent measures of masculinity, femininity, self-esteem, and social desirability were also administered to a subset of the large number of individuals who completed the initial STAT questionnaire form. The report of the complete pattern of results will follow a brief description of the method used to collect the data.

Method

For the test standardization phase of this project, described in this chapter, subjects were asked to rate themselves on a seven-point scale identical to that used by the stereotype raters during the item development stage. The endpoints of the scale were respectively labelled "Very uncomfortable or very awkward" and "Very comfortable; not awkward at all". Thirty-one masculine and 27 feminine items were listed in random order, for a total of 58 items. After the individual had completed the self ratings, two separate scores were computed by adding the ratings on the masculine (M) items and the feminine (F) items. Low scores indicated less comfort that high scores. A facsimile of the questionnaire is found in Appendix C, entitled "Original Sex-Typed Activities Test".

On the same form, subjects were also asked a number of questions concerning their age and sex, the size of their town of origin, the size of their families, whether their mothers worked outside the home, and so forth. In addition, they were asked to make some global ratings of the degree to which they saw themselves as masculine or feminine, their political orientation, and the degree of religiosity.

During three consecutive semesters, this questionnaire was administered on a number of occasions to undergraduate psychology students at the University of North Dakota. It was administered in class, in special testing sessions, or

in some cases distributed in class to be returned by the students at a later class session. In all, 662 female and 516 male undergraduates completed the original questionnaire.

There are two subsamples from this group who completed additional tests of interest. One group of 53 males and 79 females (N = 132) completed the Personal Attributes Questionnaire or PAQ (Spence & Helmreich 1978) in addition to the STAT. Another group, consisting of 37 males and 98 females (N = 135) completed a battery of tests in addition to the STAT questionnaire in specially arranged testing sessions. The tests which composed the battery are as follows:

Personal Attributes Questionnaire (PAQ). (Spence & Helmreich 1978) The PAQ is a self-rating instrument measuring sex-role orientation which uses trait descriptions for items. The 24 item scale produces three scores for each respondent: M, F, and M-F.

Bem Sex Role Inventory (BSRI). (Bem 1974) Similar to the PAQ, this measure of sex role orientation requires that the subject rate himself on a number of traits. In this case, the 20 Neutral items were not used, resulting in a 40 item form of the test. The scales produce an M and an F score for each individual.

Sex Role Behavior Scale-2 (SRBS-2). (Orlofsky, Ramsden, & Cohen, in press) The SRBS-2 is a large 240 item test which asks the subject to make self-ratings on four different areas: Recreational Interests, Vocational Preferences, Social and Dating Behaviors, and Marital Behaviors. On the first three subtests, the subject is asked to rate the degree to which certain interests or behaviors are "characteristic of me" on a five point scale. On the fourth subtest, Marital Behaviors, subjects are asked to rate the degree to which a particular behavior or role is "characteristic of me" as opposed to "characteristic of my spouse". Each of these four subtests yields three scores: M, F, and M-F. In this way, the SRBS-2 resembles the PAQ (upon which the SRBS-2 is consciously modeled) while using a completely different kind of item domain. In addition, subtest scores are summed to yield an overall M, F, and M-F score for each individual. For reasons of economy, the M-F scale was deleted in this administration and only the 160 items composing the M and F scales were given to the participants. In the abbreviated form administered in this study, the SRBS-2 yielded four M-subtest scores, four F-subtest scores, and overall M and F scores, for a total of ten possible scores for each respondent.

Marlowe-Crowne Social Desirability Scale. (Crowne & Marlowe 1964). The Marlowe-Crowne, a 33 item, True-False

test, measures a response set in subjects such that items are consistently answered in a socially desirable manner. High scores indicate a strong tendency to respond in a socially desirable way. Correlations between this scale and other personality scales permit an assessment of the degree to which a test may be invalidated by such a response pattern.

Texas Social Behavior Inventory (TSBI). (Helmreich & Stapp 1974; Helmreich, Stapp, & Ervin 1974). This is a 16-item measure of self-esteem and social competence which has been widely used in conjunction with the BSRI and the PAQ in previous research. High scores indicate high levels of self-esteem. This test was included because of the highly publicized and reliably reported relationship between masculinity (as measured by trait-based measures) and self-esteem (Taylor & Hall 1982).

Connecticut sample

As a precaution against the possibility that the type of items used on the Sex Typed Activities Test might be affected by a geographical bias, a second sample of questionnaire responses were obtained from undergraduate participants from a different section of the country. In this case, a cohort of students completed the STAT questionnaire in a group testing session at the University of Connecticut

in Storrs, Connecticut. Fifty males and 83 female undergraduates enrolled in introductory psychology courses completed the form, providing a basis for comparison between respondents from two very different regions of the country.

Results

The analysis of the data collected from the Sex Typed Activities Test questionnaire was guided by two interlocking concerns. Of course, normality of distributions, reliability, and internal coherence for each of the separate subscales were principal goals. But equally important was the investigation of the degree to which the results obtained would conform to rudimentary notions about the constructs of masculinity and femininity. The first goal implied that each of the subscales would be consistent, unified, and reliable. The second would imply, for example, that males would have higher masculinity scores than females, who would in turn have higher femininity scores. All of the results reported in this chapter pertain to one or the other of these principal concerns.

Within this chapter, the last four of the five hypotheses advanced in Chapter III can be addressed by the pertinent analyses. Hypothesis II, which suggested that a single test with two M and F subscales can be developed for use with subjects of either sex, has already been partially supported in the previous chapter, but will be considered fur-

ther in this one. Mean ratings will be examined to test the assertion of Hypothesis III that males should show higher M scores than females and that females should show higher F scores than males. Similarly, the obtained pattern of interscale correlations will relate directly to Hypothesis IV which stated that within each sex the M and F scores would be orthogonal. Finally, the factor analysis presented at the conclusion of the first part of this chapter will address Hypothesis V, which argues in favor of a two factor structure for the test. As these data for the internal attributes of the test are discussed, both reliability and validity issues will be addressed. In the latter part of this chapter, external relationships between overall STAT M and F scores and independent measures of masculinity, femininity, male and female sex roles and behaviors, and self-esteem will be explored. These may broadly be said to concern concurrent or convergent validity.

Descriptive statistics.

As noted, the North Dakota sample consisted of 516 males and 662 females, while the Connecticut sample was considerably smaller, consisting of 50 males and 83 females. There are certain contrasts between the two groups aside from the regions of the country from which they originate. The Connecticut sample for example was somewhat younger, 18.3 years of age compared to the average 19.9 for the Dako-

ta sample. The members of the Connecticut group also tended to come from slightly larger towns, although in both groups only a very small percentage came from very large cities. The typical Dakota student came from a town with fewer than 10,000 population. Fully one-third came from farms or from towns with fewer than 2,000 population. The average Connecticut student came from a town with a population between 10,000 and 50,000, and only 15% came from towns of fewer than 2,000 inhabitants.

The North Dakota students also tended to come from larger families. The average number of siblings was 2.5 for Connecticut and 3.33 for North Dakota. The comparison of ranges is interesting. The largest of the Connecticut families in this sample was ten children, including the respondent. The largest of the North Dakota families was 18 children, though needless to say this was hardly typical. Fully 14% of the North Dakota sample had at least five brothers and sisters; the comparable figure for Connecticut is less than 5%.

All students were asked to respond to a set of four questions about their family life as it was when they were growing up. The percentages of students endorsing each alternative are listed in Table 10. Inspection of this table indicates that the samples appear to be roughly equivalent. As a footnote, complaints were sometimes heard from the Dakota students that the manner in which the items were worded

made it difficult to apply them in all cases to the farm situation. In particular, it was pointed out that "being a farm wife is a full time job!"

The data presented in Tables 11 and 12 are taken exclusively from the ratings made by the Dakota sample. Listed separately by sex, the means of the self-ratings for each of the 31 masculine items and the 27 feminine items appear in these tables. For reference purposes, the item numbers on the original 209 item stereotype questionnaire (Appendix B) appear in parentheses after the STAT item number. Inspection of these means reveals that on average, respondents tended to see these items as more comfortable than uncomfortable. In only one case did the activity receive a mean rating below the scale midpoint (4.0) for both sexes. Both males and females apparently felt they would be uncomfortable "going to the PTA" (Item 18, F subscale). This item was also the one with the lowest mean rating (M = 3.04, male subjects). No item received an average rating of less than "3" on the scale ranging from "1" to "7". The highest average rating for an item was for female respondents on the feminine item #47, "Baking a cake from a mix" (M = 6.56, female subjects).

TABLE 10

Family background: North Dakota and Connecticut Students

This table lists the percentages of respondents checking the description alternatives given for four questions.

	All of the time %	Most of the time %	Occas-ionally %	Never %
North Dakota				
My parents lived together.	88.7	5.9	3.7	1.7
My father worked full time.	89.9	8.1	.9	1.1
My mother worked full time.	15.2	22.6	21.9	40.2
My mother held a part time job.	6.5	16.9	34.9	41.6
Connecticut				
My parents lived together.	83.5	11.3	3.8	1.5
My father worked full time.	89.5	8.3	1.5	.8
My mother worked full time.	10.5	21.1	20.3	48.1
My mother held a part time job.	6.8	15.0	37.6	40.6

TABLE 11

Means for the Masculine Comfort Items

Item*			FEMALES		MALES	
			M	SD	M	SD
1	(53)	Playing poker	3.36	1.80	4.79	1.88
3	(65)	Changing a fuse	3.27	1.92	5.72	1.64
4	(47)	Pruning a tree limb	3.52	1.90	5.27	1.67
6	(37)	Starting a fire in the fireplace	5.00	1.68	5.97	1.31
9	(35)	Computing your income tax	3.48	1.87	4.11	1.84
11	(30)	Reading the sports page	5.04	1.91	5.93	1.69
12	(17)	Driving a sports car	5.46	1.73	6.23	1.32
14	(16)	Mowing the lawn	5.94	1.51	5.92	1.40
15	(15)	Replacing a washer in a leaky faucet	3.20	1.94	5.32	1.70
17	(209)	Building a simple table	3.67	1.87	5.33	1.54
19	(207)	Trimming a hedge	4.35	1.77	5.22	1.50
20	(206)	Riding a motorcycle	4.25	2.12	5.56	1.78
22	(192)	Buying car insurance	3.50	1.75	4.35	1.66
25	(183)	Opening a tight jar-lid	5.48	1.53	6.09	1.23
27	(179)	Watching football on TV	4.98	2.02	6.23	1.47
28	(176)	Checking the oil in a car	4.71	2.01	6.32	1.25
30	(172)	Painting the house	5.13	1.61	5.16	1.62
33	(171)	Shovelling a sidewalk	5.52	1.55	5.71	1.42
35	(166)	Driving a boat	4.44	1.99	5.84	1.51
36	(156)	Climbing a tall ladder	4.05	2.02	5.13	1.86
38	(145)	Playing touch football	4.92	1.86	6.17	1.34
41	(125)	Buying a new car	4.38	1.86	4.94	1.73
43	(122)	Drinking a beer	5.22	2.12	5.79	1.83
44	(117)	Playing pool	4.64	1.75	5.93	1.32
46	(105)	Going fishing	5.09	1.83	5.88	1.62
49	(102)	Driving a pick-up truck	5.68	1.70	6.32	1.29
51	(97)	Watching basketball on TV	4.70	1.98	5.29	2.00
52	(88)	Jogging	4.94	1.87	5.33	1.73
54	(87)	Reading the business page	3.70	1.74	4.18	1.72
57	(80)	Using a hammer	5.41	1.51	6.22	1.19
58	(78)	Playing softball	5.21	1.75	6.06	1.47

* Item numbers are listed as the items appear on the STAT form. Numbers in parentheses are the item numbers on the original stereotype rating list. (Appendix B)

TABLE 12

Means for the Feminine Comfort Items

Item*			FEMALES		MALES	
			M	SD	M	SD
2	(104)	Getting your hair styled	5.49	1.54	4.26	1.65
5	(110)	Buying a wedding gift	5.58	1.38	3.72	1.56
7	(119)	Taking a child to the dentist	5.13	1.69	4.42	1.80
8	(124)	Washing clothes	6.28	1.15	4.80	1.79
10	(132)	Writing letters	6.05	1.36	4.61	1.68
13	(135)	Changing sheets	6.21	1.27	4.78	1.68
16	(149)	Replying to an invitation	5.49	1.43	4.59	1.53
18	(158)	Going to the PTA	3.70	1.96	3.04	1.57
21	(160)	Sunbathing	5.85	1.63	5.02	1.84
23	(163)	Getting up with a baby at night	4.73	1.91	3.44	1.85
24	(169)	Helping a child get ready for school	5.76	1.52	4.05	1.80
26	(188)	Washing the dishes	6.21	1.32	4.85	1.71
29	(203)	Mopping the floor	5.73	1.52	4.66	1.64
31	(5)	Typing a letter	5.50	1.66	4.12	1.75
32	(6)	Rearranging the furniture	6.10	1.21	5.01	1.57
34	(11)	Doing crafts	5.48	1.65	4.63	1.68
37	(13)	Setting the table	6.44	1.03	4.86	1.55
39	(20)	Shopping for clothes	6.27	1.29	4.59	1.71
40	(22)	Crying in private	6.01	1.49	3.89	2.07
42	(25)	Comforting a child	6.16	1.21	4.82	1.67
45	(33)	Re-potting a plant	5.29	1.64	4.14	1.71
47	(40)	Baking a cake from a mix	6.56	.99	4.44	1.83
48	(44)	Wrapping a present	6.50	.94	4.55	1.56
50	(58)	Babysitting for money	6.18	1.37	4.03	1.88
53	(72)	Planning a party	5.22	1.53	4.63	1.63
55	(93)	Bathing a baby	5.25	1.70	3.10	1.77
56	(99)	Shampooing a child's hair	5.41	1.63	3.33	1.79

*Item numbers are listed as the items appear on the STAT form. Numbers in parentheses correspond to the order of the items as they were listed on the original stereotype questionnaire (Appendix B).

Distribution of scores

An unexpected finding appeared regarding the distribution of raw scores (or mean ratings) on the M and F scales (see Table 13). The distribution of scores for the male sample was skewed to a significant degree on the masculine activities scale, but not on the feminine activities scale. The mirror image also occurred: female scores were significantly skewed on the F scale, while normally distributed on the M scale. Thus, the absence of normality occurred chiefly on the sex-congruent scale in each case but not on the sex-incongruent scale. In the case of each sex, the distributions of the total scores were skewed and peaked toward the high end of the scale.

It seems reasonable to assume that the underlying dimension is normally distributed within each sex. Since the distributions are skewed toward the high end of the scale on the sex-congruent scales, it would appear that the item selection process failed to include enough items at the high end of the range to adequately distinguish between high levels of masculine comfort, typically seen in males, and high levels of feminine comfort, typically seen in females. As was noted in the previous chapter, many of the "strong" items were excluded from the final scales, because they implied discomfort for the opposite sex, and it was desired to create a test that could be used by members of either sex. The lack of normality on the sex-congruent subscales would

appear to be a serious shortcoming to the test as it now exists, but it would also appear to be one that could be rectified in a straightforward manner, by adding items which discriminate between high levels of the two variables.

TABLE 13

Skewness and Kurtosis for STAT scores

MALE SUBJECTS (n = 516)						
	Skew	Z	p	Kurtosis	Z	p
STAT M	-1.29	-12.02	.0001	2.98	14.07	.0001
STAT F	.12	1.13	n.s.	- .36	1.74	n.s.
FEMALE SUBJECTS (n = 662)						
	Skew	Z	p	Kurtosis	Z	p
STAT M	- .001	- .01	n.s.	- .06	.34	n.s.
STAT F	- .93	9.84	.0001	2.05	10.94	.0001

Of course, the addition of a number of items on which members of the opposite or "incongruent" sex are uncomfortable could distort the very distributions which are normal at present, the sex-incongruent ones. Consequently, it may be necessary to enlarge only the sex-congruent scales while leaving the sex-incongruent scales intact. That is, extra

items would be added to the Masculine subscale, but only for males, and at the same time, extra feminine items would be added, but only for females. This implies two slightly different forms of the test for males and females.

Hypothesis II, which stated that a single test could be developed for use with either sex may require revision in the event that the addition of the high items distorts the scores for the sex-incongruent scales. The abandonment of Hypothesis II does not invalidate the methodology developed here in any fundamental way. On the other hand, it does raise questions about the basis upon which one decides whether the same or different groups of items should be used to measure masculinity and femininity in males as opposed to females.

Multiple Regression Analysis

In choosing the item selection and item analysis procedures for this study, considerable attention was paid to the development of other dualistic masculinity and femininity measures. In addition, the procedures used by Pedhazur and Tetenbaum (1979) to re-examine the items comprising the Bem Sex Role Inventory were scrutinized, so that questions which might be posed about the items of the Sex Typed Activities Test might be answerable in advance. One of the procedures that Pedhazur and Tetenbaum used to analyze the self-ratings on the BSRI was a Stepwise Discriminant Function analysis

using sex of respondent as the dependent variable and the 40 BSRI traits as predictor variables. In that case, the results were used to show that 38 additional variables did little to improve the prediction of sex over and above the prediction based only on the two items Masculine and Feminine. However, multivariate analysis also constitutes a way of looking at the relationship between the test items and the respondent's sex. Furthermore, it is in some ways to be preferred to performing a large number of separate t-tests. Consequently, a similar procedure was applied to the items of the STAT.

Separate stepwise multiple regression analyses were completed for all 58 STAT variables (Table 14), for the masculine items only (Table 15), and for the feminine items only (Table 16). Taken as a whole, the test predicts the sex of the individual at a high level. The multiple R for the 53 items included in the final regression equation was .85, and the amount of common linear variance (R^2) was .72. This compares to an R^2 value of .79 reported for the full 40 BSRI items by Pedhazur and Tetenbaum (1979, p. 1009). Five variables in this case failed to meet very basic SPSS criteria for inclusion in the final equation (Kim & Kohout 1975, p. 345). A summary table for the items in the equation is provided in Appendix C.

In contrast the R^2 values for the two subscales were .47 for the 29 masculine variables included in the final

TABLE 14

Multiple Regression: Overall

Stepwise Multiple Regression analysis predicting Sex of Respondent from the 58 STAT questionnaire items. Five variables were excluded from the final equation.

Source	SS	df	MS	F	p
Regression	208.52	53	3.93	54.29	.001
Residual	81.46	1124	.07		

Multiple R = .85 R^2 = .72

TABLE 15

Multiple Regression: Masculine variables

Stepwise Multiple Regression analysis predicting Sex of Respondent from the 31 masculine STAT questionnaire items. Two items were excluded from the final equation.

Source	SS	df	MS	F	p
Regression	136.68	29	4.71	35.29	.0005
Residual	153.30	1148	.13		

Multiple R = .68 R^2 = .47

TABLE 16

Multiple Regression: Feminine Variables

Stepwise Multiple Regression analysis predicting Sex of Respondent from 27 Feminine STAT items. All 27 variables were included in the final equation.

Source	SS	df	MS	F	p
Regression	173.78	27	6.44	63.71	.0005
Residual	116.19	1150	.10		

Multiple R = .77 R^2 = .60

equation, and .60 for the 27 feminine variables, all of which were included in that final predictive equation. Individual results for the items are listed in Appendix C.

These results indicate that the items which were originally selected on the basis of differences in stereotypes do in fact differentiate between males and females in terms of their own self-ratings. Inspection of the individual standardized coefficients as the variables were introduced into the equations suggests that at least with regard to the prediction of sex, many fewer variables could be used. This fact was one of several taken into consideration in the subsequent development of two abbreviated scales, which will be described later in this chapter.

Comparison of Male and Female Scores

To examine the masculine and feminine item ratings in the aggregate, total M and F scores were computed and these were converted to mean ratings to allow easier comparison between the M and F scales which contain different numbers of items. The average item ratings for male and female respondents, presented in Table 17, would appear to be very similar for the North Dakota and Connecticut groups. The impact of regional differences on total STAT scores would appear, at least in this instance, to be negligible.

TABLE 17

Average Item Ratings on STAT M and F Subscales

	North Dakota			Connecticut		
	n	M	SD	n	M	SD
31 STAT Masculine Items						
Males	516	5.56	.84	50	5.54	.73
Females	662	4.59	.83	83	4.42	.86
27 STAT Feminine Items						
Males	516	4.32	.92	50	4.54	.98
Females	662	5.73	.73	83	5.77	.77

As expected (Hypothesis III), males showed higher masculinity scores than females and females showed higher femininity scores than males. The mean self-ratings were subjected to a 2 X 2 Analysis of Variance (Sex X Masculinity/Femininity) with one between-subjects factor and one within-subjects factor. The results, presented in Table 18, indicate a classical interaction, such that males score higher on masculinity than females and that females score higher on femininity than males. There was a significant main effect for sex ($F = 28.61$, $p < .001$) in combination with a highly significant interaction with the within factor, masculine versus feminine scales, $F = 2102.31$, $p < .001$. In addition, there was a significant difference between the M and F mean scores, $F = 4.18$, $p < .05$, but given the large number of subjects involved (N = 1176), this is not a meaningful difference. The R^2 value for the repeated measure factor, masculine vs. feminine items, is less than .0007.

In addition to the cross sex comparisons, males had significantly higher masculinity than femininity scores, $t(515) = 30.39$, $p < .001$. Females had significantly higher femininity scores than masculinity scores, $t(661) = 34.72$, $p < .001$. These data indicate that the STAT scores vary in a manner that is related to sex in a theoretically predictable manner, and support Hypothesis III.

TABLE 18

Analysis of Variance

Source	SS	df	MS	F	p
Sex	27.83	1	27.83	28.61	<.001
Error	1141.98	1174	.97		
M/F	1.63	1	1.63	4.18	.042
Sex X M/F	820.57	1	820.57	2102.31	<.001
Error	458.23	1174	.39	.39	

N = 1176

Sex = Sex of Respondent
M/F = STAT M and STAT F scores, repeated measures factor.

Reliability

Both the M and F subscales of the Sex Typed Activities Test proved to be highly reliable as can be observed in Table 19. Alpha coefficients for the M scale for the two samples range from .87 to .92. For the F scale coefficient alpha values range from .88 to .91.

Since coefficient alpha represents both the number of items on a test and the average inter-item correlations in combination, the latter are given in Table 19 as a more direct indicator of internal consistency. In this case, the

TABLE 19

Reliability coefficients for STAT M and F subscales

	Coefficient alpha	Average inter-item correlations
North Dakota		
Males (N = 516)		
STAT M	.91	.27
STAT F	.91	.26
Females (N = 662)		
STAT M	.87	.18
STAT F	.88	.23
Connecticut		
Males (N = 50)		
STAT M	.88	.20
STAT F	.91	.26
Females (N = 83)		
STAT M	.88	.19
STAT F	.89	.24

inter-item correlations range from .18 to .27 for the M scale and .23 to .26 for the F scale, suggesting adequate levels of inter-relationship among the individual items which make up the two scales.

Relationships between STAT variables

Hypothesis IV suggested that the scores for particular individuals of either sex would vary independently, i.e., they would be statistically orthogonal. The precedents for this hypothesis are the two trait-based measures of M and F, the BSRI and the PAQ. Where scores vary independently it is possible to classify individuals into a typology including mixed as well as "pure-type" sex role orientation groups in the manner advocated by androgyny theorists. In direct contrast, a negative correlation between masculinity or M scores and femininity or F scores would suggest a traditional view of the underlying constructs as mutually opposing.

In the data collected with the STAT, however, this hypothesis of orthogonality was not supported. Neither were the scales negatively correlated, as might have been predicted by the traditional view. Instead, in the case of both male and female subjects, and in both the Dakota and Connecticut samples, scores on the M scale are positively correlated to scores on the F scale. For males, the correlation between M and F scores was .44 for the North Dakota sample and .33 for the Connecticut sample. For females, the correlation between the M and F scores was .42 for the North Dakota sample, and .44 for the Connecticut sample. All correlations are highly significant.

The finding of a positive correlation between M scores and F scores is conceptually perplexing but it is not with-

out precedent. It seems to suggest that, in one sense, the more "masculine" one is, the more "feminine" one is also. However, if stated in terms more precisely reflecting the operational definition used here, it is less perplexing: that is, the more comfortable one is with a wide variety of masculine activities, the more comfortable one is with a variety of feminine activities. Orlofsky reported a similar, even higher, correlation between overall M and F scores on the Sex Role Behavior Scale-1 (Orlofsky 1981). Helmreich, Spence and Holahan (1979) obtained self-ratings for comfort on four female, four male, and four neutral behaviors and found an average correlation of .67 among the three. Nevertheless, this finding disconfirms the original hypothesis that masculine and feminine comfort scales would resemble trait-based scales in form. Furthermore, it raises questions about either the underlying dimensions, or the methodology for measurement, or both. These questions will be deferred and considered more fully in the following chapter.

Social Desirability and STAT scores

The presence of a strong tendency for subjects to respond to questionnaire items in a socially desirable manner can obscure or minimize the ability of a test to represent the underlying dimension it seeks to measure. To examine the possibility that the STAT scores are subject to such a tendency, a subgroup of the North Dakota sample completed

the Marlowe-Crowne Social Desirability Scale. For 98 females, the simple correlation for the Marlowe-Crowne and the M scale was .14 (n.s.), which approaches virtual independence. The correlation between the F scale and social desirability is .27 ($p = .003$) suggesting a small percentage of the variance (7%) is held in common. For the 37 males, the correlation with the M scores was .21 (n.s.) and with the F scores, .32 ($p = .027$). It would appear that there is little, if any, relationship between self-reported comfort on masculine activities and the tendency to respond in a socially desirable manner. The responses to feminine activities do covary with a "social desirability set" to a mild degree in either sex. Thus, the F scale is largely, but not entirely, free from such influence.

Factor Analysis

The final matter to be dealt with regarding the internal properties of the Sex Typed Activities Test is the dimensionality of the test items. The responses of the 516 males and 662 females from North Dakota were subjected to separate factor analyses. In some cases the results of these analyses were very similar between the sexes. However, it was discovered that pooling males and females introduced spurious variance related to the sex of respondent which resulted in a highly interpretable but distorted factor structure. In all cases, the SPSS principal components

factoring procedure with iterations was used (Nie, Hull, Jenkins, Steinbrenner, & Bent 1975), and squared multiple correlations were used as initial estimates of the communalities.

It was originally hypothesized (Hypothesis V) that the M and F subscales would form the basis of two large factors. Consequently, within each sex, an initial factor analysis was computed extracting only two factors. As Comrey (1978) has pointed out however, the extraction of too few factors can lead to an amalgamation of smaller factors into larger ones, distorting the analysis. Consequently, a number of alternative solutions were examined concurrently and a five factor solution was judged the most interpretable and informative. The results of both sets of analyses will be reported.

A two factor solution. The two factor solution, it was hypothesized, would have a masculine factor and a feminine factor accounting together for a large amount of the variance observed for the 58 items. To examine this hypothesis, a two factor solution was derived and subjected to both varimax and oblique rotations. Under the oblique condition, the two factors correlate below .30 for either males or females. The choice to present the results of the varimax or orthogonal solution was based on the belief that, although the test scores for the M and F scales are correlated, the underlying dimensions are, or should be, orthogonal. Where

the varimax solution reflected a more dramatic division of the M and F items between the two factors, the oblique solution was similar but provided a less consistent division of the items into sex-typed groups. Having noted this, only the varimax solution will be presented for the two factor case. It is explicitly noted that this solution provides the best or strongest evidence to support the hypothesis, and is presented for precisely that reason.

A clear differentiation of masculine and feminine items occurred when two factors were extracted and rotated using a varimax technique. These two factors respectively accounted for 53% and 58% of the common variance for the male and female respondents. The factor variance for Factor I is 12.37 and 9.43 for the males and females respectively, accounting for most (71%) of the variance explained by the two factors in each case. For Factor II the corresponding factor variances are 4.99 and 3.80. The tendency for F items to load on Factor I in each case and for M items to load on Factor II is easily visible in Table 20. All factor loadings above the minimum criterion of .40 have been marked by an asterisk. With few exceptions, the factors correspond in terms of meaning to the division between masculine and feminine activities on which the test is based. This is true of both the male and female samples.

Further evidence of the relationship between the dual subtest structure of the Sex Typed Activities Test and the

TABLE 20

Factor Loadings for Two-factor solution

		MALES		FEMALES	
	Factor #	I	II	I	II
Female Comfort Items					
2	Getting your hair styled	.21	.02	.29	-.03
5	Buying a wedding gift	.47*	.04	.31	.42*
7	Taking a child to the dentist	.50*	.17	.49*	.08
8	Washing clothes	.53*	.14	.56*	.03
10	Writing letters	.30	.00	.40*	.05
13	Changing sheets	.64*	.18	.67*	-.02
16	Replying to an invitation	.49*	.24	.42*	.14
18	Going to the PTA	.53*	-.07	.41*	.16
21	Sunbathing	.23	.28	.21	.13
23	Getting up with a baby at night	.58*	-.02	.55*	.08
26	Washing the dishes	.60*	.14	.62*	-.02
29	Mopping the floor	.62*	.25	.62*	.14
31	Typing a letter	.45*	.16	.39	.15
32	Rearranging the furniture	.48*	.22	.54*	.16
34	Doing crafts	.45*	.16	.38	.22
37	Setting the table	.66*	.16	.67*	-.01
39	Shopping for clothes	.33	.11	.30	.01
40	Crying in private	.34	-.03	.23	.06
42	Comforting a child	.55*	.01	.45*	.03
45	Re-potting a plant	.59*	.14	.44*	.31
47	Baking a cake from a mix	.54*	.14	.51*	.03
48	Wrapping a present	.57*	.12	.55*	.10
50	Babysitting for money	.50*	.09	.42*	.04
53	Planning a party	.26	.32	.30	.26
55	Bathing a baby	.65*	-.08	.55*	.10
56	Shampooing a child's hair	.67*	-.05	.59*	.08
Male Comfort Items					
1	Playing poker	-.02	.38	-.13	.34
3	Changing a fuse	.23	.54*	.05	.46*
4	Pruning a tree limb	.34	.43*	.13	.42*
6	Starting a fire in the fireplace	.23	.61*	.19	.42*

9	Computing your income tax	.33	.23	.15	.27
11	Reading the sports page	-.09	.50*	.10	.38
12	Driving a sports car	.01	.64*	-.01	.43*
14	Mowing the lawn	.36	.47*	.48*	.25
15	Replacing a washer in a leaky faucet	.36	.55*	.08	.51*
17	Building a simple table	.22	.46*	.09	.49
19	Trimming a hedge	.41*	.43*	.30	.43*
20	Riding a motorcyle	.05	.56*	-.10	.50*
22	Buying car insurance	.28	.34	.13	.37
25	Opening a tight jar-lid	.34	.55*	.41*	.28
27	Watching football on TV	-.11	.58*	.05	.41*
28	Checking the oil in a car	.15	.69*	.02	.52*
30	Painting the house	.39	.45*	.39	.43*
33	Shovelling a sidewalk	.35	.49*	.48*	.31
35	Driving a boat	.08	.55*	.02	.53*
36	Climbing a tall ladder	.15	.48*	.12	.31
38	Playing touch football	.03	.60*	.07	.49*
41	Buying a new car	.11	.31	-.05	.44*
43	Drinking a beer	-.04	.40*	-.00	.18
44	Playing pool	-.02	.57*	.04	.50*
46	Going fishing	.07	.49*	.22	.41*
49	Driving a pick-up truck	.20	.65*	.08	.42*
51	Watching basketball on TV	-.05	.43*	.18	.30
52	Jogging	.18	.44*	.16	.33
54	Reading the business page	.38	.33	.23	.39
57	Using a hammer	.15	.67*	.35	.45*
58	Playing softball	-.03	.60*	.13	.49*

*An asterisk indicates that this loading exceeds the minimum criterion of .40.

TABLE 21

Mean factor loadings for M and F items

	Factor I	Factor II
Males		
Masculine items	.16	.50
Feminine items	.50	.10
Females		
Masculine items	.14	.39
Feminine items	.46	.09

two factors extracted is given in Table 21 which lists the average loadings of the complete group of 31 M items and 27 F items on each of the two factors. For both the male and female analyses, the mean of the loadings of the F items on Factor I is much higher than the mean of the M item loadings. Factor II has much higher average loadings for the group of M items than the group of F items. Consequently, it would appear that when two factors are extracted and rotated in an orthogonal pattern, the dimensions that appear are clearly sex-linked and correspond more-or-less directly to the subscale structure of the overall test.

A Five factor solution. A more complete picture of the dimensions which underlie the STAT questionnaire in its present 58 item form can be gained by extracting and rotating five separate factors. While two factors accounted for

about half the common variance, five factors account for 81% of the common variance in both sexes. Fewer factors led to less interpretable structure, and more than five factors led to excessive fragmentation. The five factors are very similar regardless of whether an oblique or a varimax rotation is performed and the correlations between the oblique factors are low. Therefore, once again, only the results of the varimax rotation will be reported here. The factor solutions for the two correlation matrices produced by male and female self-ratings are also quite similar.

The complete listings of factor loadings are found in Appendix C. For the reader's convenience, the variables which have factor loadings above .40 for each of the five factors have also been listed in separate tables for males and females (Tables 22 and 23). These items have been listed in descending order according to their correlations with the factor (factor loadings). The relative proportions of variance explained by the five factors are presented in Table 24.

For the most part, the resulting factors can readily be classified as masculine or feminine in nature, reflecting again the structure of the test. The first of two factors concern sex-typed work activities or responsibilities and in each analysis the sex-appropriate group of items emerges as the first factor. The "Domestic drudgery" factor (Factor I for females, II for males) is chiefly composed of household

TABLE 22

Items Loading on Five Factors: Male Responses

Item		Type	Loading
Factor I: Male Tasks			
3	Changing a fuse	M	.73
15	Replacing a washer	M	.70
6	Starting a fire in the fireplace	M	.65
4	Pruning a tree limb	M	.62
57	Using a hammer	M	.62
17	Building a simple table	M	.61
28	Checking the oil in a car	M	.57
35	Driving a boat	M	.56
12	Driving a sports car	M	.52
19	Trimming a hedge	M	.52
49	Driving a pick-up truck	M	.52
20	Riding a motorcycle	M	.51
36	Climbing a tall ladder	M	.47
25	Opening a tight jar-lid	M	.44
Factor II: Domestic drudgery			
26	Washing the dishes	F	.70
14	Mowing the lawn	M	.65
29	Mopping the floor	F	.65
33	Shovelling a sidewalk	M	.64
13	Changing sheets	F	.62
37	Setting the table	F	.59
30	Painting the house	M	.50
8	Washing clothes	F	.49
25	Opening a tight jar-lid	M	.44
32	Rearranging the furniture	F	.43
19	Trimming a hedge	M	.42
Factor III: Sports Interests			
27	Watching football on TV	M	.77
11	Reading the sports page	M	.72
58	Playing softball	M	.69
38	Playing touch football	M	.68
51	Watching basketball on TV	M	.66
44	Playing pool	M	.51
52	Jogging	M	.46

Factor IV: Child-care

55	Bathing a baby	F	.81
56	Shampooing a child's hair	F	.80
24	Helping a child get ready for school	F	.78
23	Getting up with a baby at night	F	.75
42	Comforting a child	F	.56
7	Taking a child to the dentist	F	.47
18	Going to the PTA	F	.45

Factor V: Feminine enjoyment

39	Shopping for clothes	F	.56
21	Sunbathing	F	.48
53	Planning a party	F	.49
48	Wrapping a present	F	.46
5	Buying a wedding gift	F	.42
45	Re-potting a plant	F	.42
31	Typing a letter	F	.42
16	Replying to an invitation	F	.41

TABLE 23

Items Loading on Five Factors: Female responses

Item		Type	Loading
Factor I: Domestic drudgery			
13	Changing the sheets	F	.70
26	Washing the dishes	F	.70
37	Setting the table	F	.70
29	Mopping the floor	F	.69
8	Washing clothes	F	.60
48	Wrapping a present	F	.57
33	Shovelling a sidewalk	M	.56
14	Mowing the lawn	M	.54
47	Baking a cake from a mix	F	.50
32	Rearranging the furniture	F	.49
30	Painting the house	M	.49
Factor II: Male Tasks			
3	Changing a fuse	M	.67
15	Replacing a washer in a leaky faucet	M	.66
4	Pruning a tree limb	M	.60
19	Trimming a hedge	M	.58
17	Building a simple table	M	.55
28	Checking the oil in a car	M	.53
22	Buying car insurance	M	.44
6	Starting a fire in the fireplace	M	.42
30	Painting the house	M	.42
57	Using a hammer	M	.40
Factor III: Child-care			
55	Bathing a baby	F	.82
56	Shampooing a child's hair	F	.79
24	Helping a child get ready for school	F	.73
23	Getting up with a baby at night	F	.71
42	Comforting a child	F	.61
18	Going to the PTA	F	.43

Factor IV: Sports Interests

27	Watching football on TV	M	.72
38	Playing touch football	M	.66
11	Reading the sports page	M	.63
58	Playing softball	M	.59
51	Watching basketball on TV	M	.57

Factor V: Feminine enjoyment

53	Planning a party	F	.54
12	Driving a sports car	M	.52
39	Shopping for clothes	F	.52
41	Buying a new car	M	.52
21	Sunbathing	F	.42

tasks that are stereotypically held to be more comfortable for females. The principle underlying this factor seems to have more to do with the type of work per se than with sex. Certain domestic M-typed tasks appear on this factor, although it is chiefly a feminine factor.

TABLE 24

Factor variance for the Five Factor Solutions

Factor	Factor Variance	Pct of Variance	Cumulative %
	FEMALES		
1	9.56	47.0	47.0
2	3.93	19.4	66.4
3	2.78	13.7	80.1
4	2.47	12.1	92.2
5	1.58	7.8	100.0
	MALES		
1	12.48	51.9	51.9
2	5.13	21.3	73.3
3	2.58	10.7	84.0
4	2.21	9.2	93.2
5	1.64	6.8	100.0

The next factor to be discussed is entitled "Male Tasks" and appears as the first factor for males and the second for females. All of the items which load on this

factor are masculine-typed items. This factor differs between males and females since, in addition to tasks, it includes some recreational items (e.g., Riding a motorcycle), but only for males.

It can be observed at this point that one of the fundamental components of the STAT is sex-typed work activity. In each case these two work factors account for over two-thirds of the variance that can be accounted for by the full five factors (Males, 73.3%; Females, 66.4%). However, the task focus of the test is supplemented by three other principal factors. Factor III for females, and Factor IV for males are entitled "Child-care" and represent a cluster of activities related to the raising and care of children. These are all stereotypically feminine activities and this factor is clearly an F factor. In contrast, Factor III for males (Factor IV for females) is a supplementary M factor entitled "Sports Interests". These items describe a degree of comfort in activities that are athletically oriented.

Finally, Factor V is a small factor which differs for males and females. This factor is given the tentative title of "Feminine enjoyment". For males it is an exclusively F factor, including only items which are judged stereotypically to be more comfortable for females. It seems to be a residual F category separated from the domestic and child-care tasks. For females, Factor V is less work oriented than it is for males and includes two M items (Driving a

sports car, Buying a new car) which seem to fit in with the other items by virtue of being extremely enjoyable or positive activities, which parallel the sports interests of males. In sum, there appear to be two fundamentally masculine components to this test and three feminine components. The M components are male tasks and sports interests. The F components are domestic drudgery, child-care, and feminine enjoyment.

Although the anticipated factor structure of two very large factors appeared in the two factor solution the careful examination of the larger solution suggests two things. First, the division of the items into M and F categories does in fact permeate the underlying structure with the single exception of the Domestic drudgery factor. (Feminine enjoyment, the most minor factor, also shows some overlap for female respondents only.) Second, each of the subtests can be decomposed into identifiable elements or clusters. This means that a careful examination might be made of the need for developing subscales conforming more closely to the underlying structure.

Sex-typed comfort and sex-typed traits

If the Sex Typed Activities Test is to be regarded as a supplementary measure of masculinity and femininity, analogous in purpose to the BSRI and PAQ, then its relationship to other masculine and feminine phenomena is significant.

Correlations between the STAT M and F scores on the one hand, and the subscales of the Bem Sex Role Inventory and Personal Attributes Questionnaire on the other were examined as a method of providing evidence about the concurrent validity of the scales.

The level of intercorrelation which would be predicted depends on one's view of masculinity and femininity. If a more general construct called masculinity or femininity underlies both trait endorsement and comfort on sex-typed items, large correlations between M scales and between F scales would be expected. This assumption, that all varieties of masculine and feminine characteristics would be highly correlated among themselves, seemed to be operative when the omnibus, bipolar, M-F measures were in vogue. As noted, the assumption that masculinity and femininity are tightly intercorrelated across content domains has recently come into question (Spence & Helmreich 1980; Spence, Helmreich, & Holahan 1979).

Overall, STAT M and STAT F scores were correlated to the appropriate trait subscales and orthogonal to the sex-crossed trait measures. However, the levels of these correlations varied considerably. The evidence for the concurrent validity of the STAT subscales, vis a vis the trait measures, is presented in Table 25. For both sexes of respondents, STAT M scores are significantly and positively correlated to the BSRI Masculinity scale, the PAQ Masculini-

TABLE 25

Intercorrelations of STAT scores with Trait M/F scores

		STAT M		STAT F	
	n	r	p	r	p
BSRI M					
Males	34	.41	.008**	.08	.32
Females	92	.53	.000***	.20	.03*
BSRI F					
Males	34	.02	.45	.29	.046*
Females	92	.14	.09	.37	.000***
PAQ M					
Males	88	.27	.006**	.03	.38
Females	174	.47	.000***	.11	.08
PAQ F					
Males	88	-.07	.25	.21	.024*
Females	173	.03	.37	.19	.007*
PAQ MF					
Males	88	.27	.005**	-.01	.45
Females	174	.32	.000***	-.15	.023*

*p < .05
**p < .01
***p < .001

STAT = Sex Typed Activities Test
BSRI = Bem Sex Role Inventory
PAQ = Personal Attributes Questionnaire

ty scale, and the PAQ M-F scale (scored in a masculine direction). STAT M scores also appear to be orthogonal to PAQ F and BSRI F scores. Comfort on feminine sex-typed activities is positively correlated to the endorsement of feminine traits (PAQ F, BSRI F). With the exception of females who showed a very mild correlation between STAT F and BSRI M scores, STAT F scores are also orthogonal to the M scales based on traits.

The magnitude of these correlations generally tends to be small to moderate. The amount of variance held in common between the masculine comfort scores and trait masculinity scores ranges from 7% to 28%. The common variance between comfort and trait scores that are feminine tend to be smaller, ranging between 4% and 14% in this sample.

To put these correlations into perspective, it is pertinent to report the correlations between the trait measures themselves. For males, the correlations between the two trait measures (BSRI and PAQ) were .65 for the masculinity subscales, and .60 for the femininity subscales. For females, the two trait masculinity measures correlated .73 while the two trait femininity measures correlated .69. Given the similarity of content between the BSRI and the PAQ, and the fact that they are both measures of the domain of traits, their intercorrelations define the upper boundary against which the correlations with the STAT M and F scales can be compared.

It is significant that in spite of the unexpected correlation between the STAT M and F scales, the pattern of intercorrelations with other M and F measures is theoretically consistent with the test rationale. This would seem to provide support for the validity of the test as a sex role orientation measure.

Sex typed interests, roles and behaviors

The Sex Typed Activities Test is composed of behavioral items which reflect stereotypes about male and female social and work roles, interests, and activities. As a result, it would be anticipated that strong relationships might be apparent between the STAT M and F scores and scores obtained on other measures of similar non-trait dimensions related to sex role. One such published test is the Sex Role Behavior Scale, which has been described earlier and is discussed in Appendix A.

The general pattern of correlations is seen as partial support for the claims of the STAT to concurrent validity. The multiple problems and shortcomings of the SRBS-2, taken as a whole, however, make it difficult to use it as a criterion for validity. In my view, the SRBS-2 may be reliable but it is not meaningful for two reasons. First, overall scores combine subtest scores in a way that is not statistically defensible: subtest scores are often more highly correlated to their cross-sex subscale of similar content (M

with F) than to the other subtest scores along the same M or F dimension. Second, the three way structure of the test (M, F, M-F) does not make good theoretical sense. A better test might be composed of the same items if they were restructured into two subtests (M and F) that are slightly different for males and females within each content domain. Nevertheless, the correlations are provided here since the SRBS-2 is the closest instrument to the STAT that exists.

Table 26 lists all of the correlations between the subjects' scores on the SRBS female-valued scales and their scores on the two STAT subscales. Again the overall pattern is encouraging although the magnitude of the correlations is unimpressive. Only two of the STAT F/SRBS F correlations exceed .30 and none exceed .37. Consequently, all of these correlations are actually smaller than the correlation between the STAT M and F subscales, which were intended to measure different dimensions.

Correlations between masculine comfort scores and male-valued Recreational Interests and Social and Dating Behaviors scores are more significant and of greater magnitude. However, comfort on male stereotyped activities is not predictive of male vocational interests or marital behaviors. This is congruent with the rather variable pattern of correlations observed among the subscales of the SRBS-2 themselves, which are discussed in Appendix A. Of particular interest is the negative correlation between the Marital

TABLE 26

Intercorrelations: STAT M & F with SRBS-2 F scales

	STAT M		STAT F	
	r	p	r	p
Sex Role Behavior Scale-2 Female-valued Subscales:				
Recreational Interests				
Males	-.13	(.22)	.35	(.017)*
Females	.01	(.47)	.22	(.014)*
Vocational Interests				
Males	-.19	(.13)	.29	(.043)*
Females	-.06	(.28)	.26	(.006)**
Social and Dating Behaviors				
Males	.15	(.18)	.37	(.011)*
Females	.06	(.27)	.17	(.05)*
Marital Behaviors				
Males	-.25	(.07)	-.15	(.195)
Females	-.22	(.02)	.20	(.026)*
Overall Composite scores				
Males	-.21	(.10)	.23	(.085)
Females	-.14	(.09)	.29	(.002)**

*p <.05
**p <.01
***p <.001

TABLE 27

Intercorrelations: STAT M and F with SRBS-2 M Subscales

	STAT M		STAT F	
Sex Role Behavior Scales- 2: Male-valued Subscales:				
	r	p	r	p
Recreational Interests				
Males	.43	(.004)**	.16	(.174)
Females	.54	(.000)***	.12	(.13)
Vocational Preferences				
Males	.11	(.26)	.19	(.13)
Females	.12	(.12)	.04	(.34)
Social and Dating Behaviors				
Males	.28	(.044)*	.25	(.06)
Females	.46	(.000)***	.03	(.19)
Marital Behaviors				
Males	-.05	(.39)	-.12	(.23)
Females	.12	(.12)	.15	(.07)
Overall Composite Scores				
Males	.24	(.07)	.12	(.24)
Females	.40	(.000)***	.18	(.04)*

*p < .05
**p < .01
***p < .001

Behavior M and F subscales and the STAT M and STAT F scores. These scales are totally orthogonal with one exception (the correlation of .20 for females between STAT F and the Marital F scores). This is not surprising in view of the fact that for the males in our sample, the Marital M and F subscales were actually negatively correlated to Recreational Interests and Vocational Preferences subtest scores as well. It is because of such discrepancies in the SRBS-2 itself that overall composite scores are not viewed as meaningful since the scales are composed of large numbers of items with relatively low levels of communality.

Clearly the relationships between different varieties of sex-role characteristics are complex and far from unitary even within the masculine and feminine "families". The strong possibility exists that the correlations would be stronger if the measures themselves were improved to reflect more accurately the underlying dimensions which they attempt to assess. However, at present, it appears that a multidimensional view of masculine and feminine personality dimensions might be preferred to a model of two highly intercorrelated dimensions which affect a variety of aspects of personality.

Self-esteem

The relationship between levels of self-esteem and levels of masculinity and femininity has been a focus of substantial discussion (Taylor & Hall 1982). In the validation component of this study, the Texas Social Behavior Inventory was administered as part of the overall battery. The expected strong relationship between self-esteem as measured by this instrument and trait masculinity scores was replicated. Correlations to the Bem Sex Role Inventory scores will be used to illustrate. For females, (n = 92) the correlation between the TSBI scores and BSRI-M scores was .71, and between the TSBI and BSRI-F, .22. For males (n = 33), the corresponding correlation for the M scores was .53, and for F scores, .01 (n.s.). These correlations are within the same range as has been previously reported in the literature.

In contrast, the correlations between self-esteem and sex-typed comfort as measured by the Sex Typed Activities Test were markedly lower. STAT M scores were correlated to TSBI self-esteem scores at the level of .23 for males (n = 36) and .36 for females (n = 98). STAT Femininity scores were correlated with TSBI scores at .15 (n.s.) for males and .19 for females in this sample. All correlations are significant unless otherwise noted. It would appear that within this domain the relationship between masculine scores and self-esteem is quite limited.

The Abbreviated STAT

When all of the data had been analyzed for the 58-item version of the test, derived from the stereotype ratings, an abbreviated version was developed to permit rapid administration. This version consists of 36 of the 58 items: 18 M and 18 F items. The selection of the 36 final items depended upon a combination of factors: redundancy, intercorrelations, multiple correlations, and factor loadings were among the major criteria. These items are listed in Table 28. Reanalysis of the data gained from the North Dakota sample indicates that these 36 items perform as efficiently as the full 58 item Sex Typed Activities Test. For males, the M scores for the full and short versions correlated .96, while the F scales correlated .98. For females, the M scores correlated .96 and the F scores, .97. Obviously, there is little to be gained by using the full scale when the correlations with the shortened scales are this high.

The reliabilities of the subtests are not adversely affected by the reduction, either. For the abbreviated STAT the alpha coefficients range from .81 to .88. The average inter-item correlations range from .20 to .30. Similarly, the correlations between subscale scores are comparable to those of the full scales. Reanalysis indicated that for females, the two are correlated at .33, and for males, at .39. It seems clear that for most research applications the form of the test to be used would be the abbreviated form, composed of the items listed in Table 28.

TABLE 28

Items of the Abbreviated STAT

Masculine Activities	Feminine Activities
Changing a fuse	Taking a child to the dentist
Pruning a tree limb	Washing clothes
Driving a sports car	Writing letters
Replacing a washer in a leaky faucet	Changing sheets
Building a simple table	Replying to an invitation
Trimming a hedge	Getting up with a baby at night
Opening a tight jar-lid	Washing the dishes
Watching football on TV	Typing a letter
Checking the oil in a car	Rearranging the furniture
Driving a boat	Doing crafts
Playing touch football	Setting the table
Buying a new car	Shopping for clothes
Drinking a beer	Comforting a child
Playing pool	Re-potting a plant
Driving a pick-up truck	Baking a cake from a mix
Watching basketball on TV	Wrapping a present
Using a hammer	Babysitting for money
Playing softball	Shampooing a child's hair

OVERVIEW AND DISCUSSION

The present study was an attempt to apply the dualistic paradigm of masculinity and femininity to an unexplored domain of item content: day-to-day activities. In view of the widespread acceptance of dualistic trait measures, and the controversy which surrounds the implications of varying degrees of M and F traits for behavior, this seemed an appropriate extension of contemporary research trends.

The results of this effort have provided significant evidence that it is indeed possible to apply similar item selection procedures to domains other than traits and derive reliable masculine and feminine scales. This study therefore can be cited as support for the notion of a variety of interrelating domains of sex-linked personality characteristics which share varying degrees of communality (cf. Spence & Helmreich 1979, 1980).

As an attempt to differentiate among individuals who are more comfortable with stereotypically masculine or feminine activities, the Sex Typed Activities Test is a qualified success. Scores on this test reliably discriminate between those who report greater comfort with a range of masculine and feminine activities and those who report less comfort. Reflecting a multi-dimensional perspective on mas-

culinity and femininity, this test has been proposed as a supplement to measures which assess sex role orientation in other domains, not as a complete measure of masculinity and femininity, per se. The thorough validation of such an approach will take additional time and study.

Originally, it was projected that an instrument could be developed which would be in many ways analogous to the trait M and F measures, but which would simply tap a separate domain. However, in the process of attempting to replicate the procedure used to develop M and F tests in the trait domain, it was discovered that the underlying model of masculinity and femininity may be different for different content areas. It is not possible to assume that the structure of either the BSRI or the PAQ will be appropriate across all possible content areas. At the same time, it was also discovered that certain general principles may apply to the measurement of different kinds of masculine and feminine personality characteristics.

The original goals for the Sex Typed Activities Test were termed "hypotheses", again underscoring the investigative nature of test-building in personality research dealing with sex roles. The first of these hypotheses concerned the selection of a group of items which were seen by both males and females as stereotypically more comfortable for one sex than the other. Such a group of items was found, and this formed the basis for the M and F subscales. Though males

and females as groups did not necessarily agree about how comfortable the typical male or female would feel, they did agree about which items were more comfortable for one sex, and which were more comfortable for the other.

However, at the same time, an interesting quirk appeared in the stereotype data. For the most part, general agreement about the level of comfort the typical male or female would feel in performing various behaviors was the rule between male and female judges. However, there was significant disagreement in one area. Male judges, as compared to female judges, tended to see the typical female as less comfortable overall, on either masculine or feminine activities. Female judges saw the typical female as more comfortable on the feminine items than did male judges. But they also saw the typical female as more comfortable with the masculine activities than the typical adult (sex-unspecified). This interactive effect is considered a significant finding of the stereotype phase of the study. It suggests that males and females differ in some important way: in the that they perceive stereotypes, in the way that they report them, or both.

It seems likely that women do see the world in less sex-stereotyped terms than men do. Men are more likely to underestimate the capabilities and the comfort of women on both sets of activities. But on the other hand, women may be motivated to disavow stereotypes which they see as pejor-

ative. The neutral adult ratings tended to be a reflection of the stereotypes that the raters had about their own sexes. Across the stereotype items, the comfort ratings of the male and female groups for the typical member of their own sex were highly correlated with the ratings made by members of the same sex for the typical adult. When this is taken together with the fact that the female judges rated the female target higher then the adult target on masculine activities, it seems to buttress the interpretation of the interaction made here.

The clear implication is that if stereotypes are to be used as the foundation for item selection in any content area, a thorough study needs to be made of sex differences in stereotyping beforehand and ratings for the "control category" of adult (sex-unspecified) should be included. By virtue of the high degree of concordance for the item means between the adult category and the sex of the stereotype raters, neutral ratings provide an important source of information about stereotypes that might otherwise be obscured.

Although the STAT proved to be highly reliable and revealed the appropriate pattern of means for males and females (cf. Hypothesis III), there are, as I see it, three important shortcomings to the test in its present form. These, I believe, all derive from the attempt to create a single test for use with both males and females (according to Hypothesis II). The first of these flaws was a failure

to obtain normal distributions on the sex-congruent subscales (M for males, F for females), which may have resulted from a truncation of the test itself at high levels of M and F. The second was the discovery of a positive correlation between the M and F subscales of the test, when the hypothetical model assumes that these are orthogonal dimensions. The third area of concern was the fragmentation of the factor structure, which failed to show a clear two-factor matrix as hypothesized (Hypothesis V).

To understand the ways in which these phenomena are related to one another it will be more expedient to explain the mechanics of M and F measurement through the device of a tentative theoretical model. The development of this model, which draws upon and organizes a number of contemporary theoretical notions, is largely original in this paper.

A Theoretical Model of M and F Measurement

Critics have argued that current masculinity and femininity measurement techniques are not based on theory and lack adequate definitions (Locksley & Colten 1979; Myers & Gonda 1982a,b; Pedhazur & Tetenbaum 1979). In the development of the Sex Typed Activities Test, a number of questions of a theoretical nature arose and certain patterns became discernible which would seem to apply regardless of the sphere of masculinity or femininity in which a researcher is interested. What follows is a description of certain basic

principles which underlie the attempts to measure masculinity and femininity with separate scales. These principles are described in terms of a general case, where a rating scale is used and subjects rate themselves on a series of items expressing masculine or feminine personality characteristics. It will not be possible in this discussion to trace all of the implications of the variables highlighted, but some of the more important ones will be stated. These implications are speculative or hypothetical, deriving from the overall theory, which must be tested in the context of actual applications of the model.

About definitions

The problem of measuring masculinity and femininity begins with the lack of a precise definition of those terms which can be translated into a measurement device. The terms masculinity and femininity are layman's terms indicating a belief in a particular quality of personality, inherent in the person rather than the observer. By definition, this quality, or more precisely these qualities, are regarded as different for males and females, in terms of how characteristic or how desirable they are. The premise for the development of measures of masculinity and femininity (regardless of whether they are bipolar or dualistic) would seem to be that individuals differ in some way having to do with sex role. Some people seem to have incorporated into

their personalities a strong identification with the sex role appropriate to their sex, while others show a milder identification, and still others embrace an "inappropriate" sex role.

If measurement devices are to be developed there must be consensus as to whether masculinity and femininity are one thing or two. Psychologists have spent a half-century attempting to define and operationalize these terms so that adequate discrimination could be made between discernible levels of masculinity and femininity. In their everyday sense, these two terms are seen as mutually exclusive, diametrically opposed, and, at the same time, inextricably bound to one another. But the attempts to measure a single construct called "M-F" have proved to be such a resounding failure that psychologists have since attempted to divide them, in concept as well as in terms of measurement technique. A number of commentators have argued against the bipolar M-F model on both rational and empirical grounds (Bem 1974; Constantinople 1973; Spence & Helmreich 1978; Storms 1979).

The use of separate M and F scales would seem to imply that the theoretical constructs, which the scales presumably represent, are separate entities. Very often, the traditional notion of M-F as a single construct is confused with the alternative notion of masculinity and femininity as separate qualities. There are multiple examples of this confu-

sion, many of them in the M/F and androgyny literature. For example, the PAQ contains not only separate M and F scales, but a bipolar M-F scale as well. This is evidence of the lack of a compelling theory in this area. A theory of masculinity and femininity measurement should provide some justification for deciding whether adequate measures will have one, two, three, or more scales.

Although most people believe that the level of masculinity-femininity is the property of the person, not the observer, this may not be the case. To the degree that such a thing as "overall M-F" exists, I would argue, it is a function of the observer's notions of sex role. It represents a perceptual compilation of a variety of observations made about the object by the subject. A distinction needs to be drawn between the psychological construct used by observers to organize the world around them (M-F) which is essentially unitary, and the orientation to sex role internalized by the person which is necessarily dual in nature (M and F). Most people, it may be argued, judge levels of sex-role orientation on the basis of a bipolar notion of M-F. Kelly (1963) has pointed out that many personal constructs employed by the person as a 'construer' of events are bipolar in nature. It may be due to the salience of this bipolar conception that psychologists first attempted to operationalize M-F in a bipolar fashion.

A distinction can be made between people as construers and as behavers. Although the study of how people perceive sex roles may be interesting in itself, this should be openly distinguished from the study of sex role characteristics in personality. What is relevant to the study of personality is the orientation to sex role which represents relative degrees of expression of masculine and feminine qualities.

Sex role orientation is one basis for making empirical distinctions among individuals. Any individual may be oriented toward male and female categories in a number of different areas of his or her life. The areas correspond more-or-less directly to the content used to assess masculinity and femininity in a systematic way. These may include (among other things): interests, vocations, demeanor, dress and appearance, role behaviors, traits, activities, and so forth. Since in these areas every individual has two models or prototypes from which to choose, the male and the female, it makes sense that each person would have the potential of gravitating toward either or both, although to varying degrees. It is as if such an individual grew up in a bi-lingual family and could choose a vocabulary that derived principally from one language or the other, or could be mixed for some purposes.

Where the person internalizes both models to a high degree, the label "androgynous" has been introduced as a way of describing the orientation to sex role. Unfortunately,

in its application this term has referred only to the endorsement of both masculine and feminine traits, and many other areas of life in which sex role orientation is expressed have been ignored.

In drawing a distinction between the phenomena of sex roles in person perception and sex roles in personality, it may be useful to point out that either is a valuable area of study. However, the confusion of the two can only lead to unfortunate results. In the area of personality, I would argue quite strongly against any encroachment of bipolar models in sex-role orientation measurement. Part of the difficulty in this area is semantic. It is difficult to think of what it means to be "non-masculine" without at the same time implying femininity. But if the underlying model a researcher has is bipolar, it can confuse the creation of measures, the analytic procedures and the interpretation of results (Taylor & Hall 1982). By recognizing the distinction I have described, the confounding of dualistic and unidimensional models can be diminished. In the following discussion, the focus will be limited to the study of sex roles in personality; masculine and feminine dimensions will be treated as separate and independent.

Background

In the early 1970's, two tests were introduced which fundamentally changed the field of masculinity and femininity measurement. These were the Bem Sex Role Inventory and the Personal Attributes Questionnaire. The use of separate scales to measure masculinity and femininity was the single most dramatic innovation incorporated into these tests, and the notion of M and F as separate entities has since achieved widespread acceptance as an advance over the bipolar, unidimensional model.

However, these two tests contained at least two other features which distinguished them from earlier attempts to measure similar dimensions. One of these, the restriction of item content, was a distinct improvement. The content of these two tests is strictly limited to trait items. Earlier tests had used virtually any kind of item on which sex differences in response could be obtained. This meant that in terms of content they were omnibus tests, which in part explains their poor performance as psychological measures. Another innovation was the abandonment of the strategy of choosing items on the basis of mere sex differences. Where previously any item was selected on which males and females gave different responses, the newer strategy was to use differences in sex stereotypes as the foundation for self-rating instruments. This innovation has been somewhat more controversial (Pedhazur & Tetenbaum 1979; Locksley & Colten 1979).

An unfortunate carry-over from the earlier work done on masculinity and femininity, however, was a tendency toward evaluative judgments. This was particularly true in the work of S. L. Bem, who advocated that there was an advantage to an androgynous or compounded sex role. This position she distinguished from the earlier value-laden notions of the past; those which, according to her argument, supported the view that sex-typing was a good thing in terms of mental health. A third value, however, can be distinguished: the value of investigating the implications of differing sex roles without advocacy.

The Person Continuum

In order to actually measure masculinity and femininity, the researcher is required to articulate a domain of potential items and relate that domain to the overarching concepts. Before discussing this step, however, it may be useful to describe what is known in the abstract about the hypothetical entities called masculinity and femininity.

The purpose of tests, of course, is to discriminate between individuals with varying degrees of some quality. It has already been stated that in terms of personality differences, masculinity and femininity have come to be regarded as separate qualities and measured by separate scales, or subscales. Consequently, the theoretical dimensions which tests attempt to approximate are either of one type or the

other, masculine or feminine. Two theoretical dimensions can be posited which refer to the range of persons from low levels of masculinity or femininity to high levels, and furthermore, it can be presumed that everyone has an appropriate location along each of these two dimensions. Since many of the arguments to be presented will be equally applicable to the M or the F dimension, it will save time to refer to the Feminine dimension as an example, for discussion purposes. The choice of femininity, illustrated in Figures 2 and 3, is entirely arbitrary since the principles to be discussed at this stage are so elementary as to apply equally well to either M or F. The model which will be discussed, is for the moment, stripped of its complexities and femininity will be considered a simple linear construct which implies that some individuals are more "feminine" than others.

If it is assumed that M and F are separate and statistically independent qualities of personality, and that both males and females have "true scores" on both of the hypothetical dimensions which correspond to their individual levels of M and F relative to all other persons, a few things can be inferred. First of all, by definition the scores are related to sex. Turning again to the illustration of F, we expect females to be more feminine than males. On a hypothetical dimension describing levels of femininity (F, in Figure 2), the mean score of females will therefore be higher than the mean score of males. (The M scale would

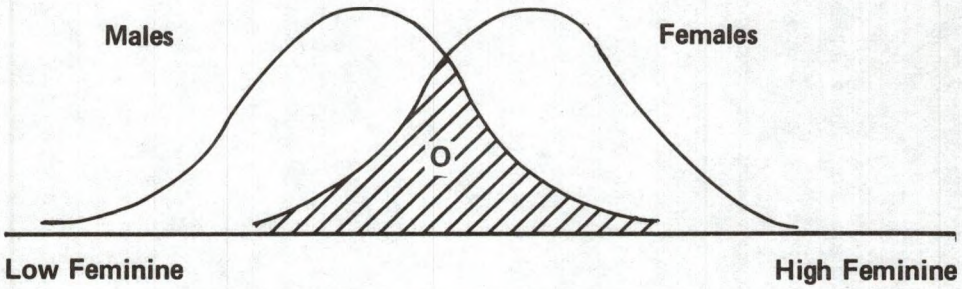


Figure 2. The theoretical distribution of "true scores" for males and females on the Femininity Dimension.

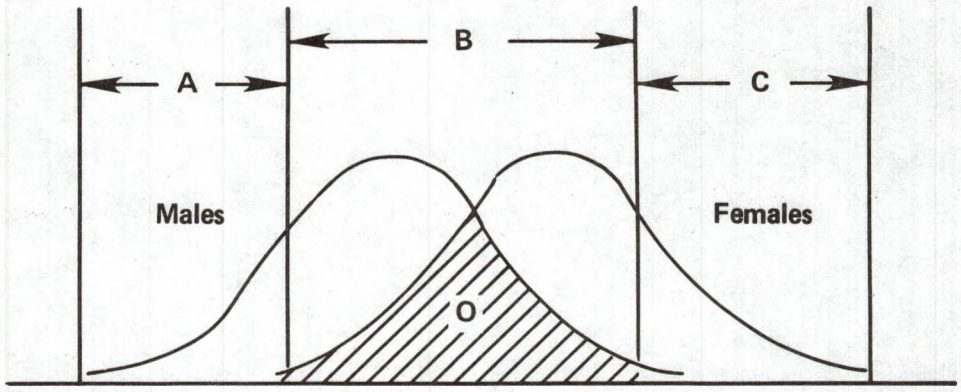


Figure 3. Three types of "true scores" on Femininity.

be a mirror image of this one.) On the other hand, we don't necessarily expect all females to be more feminine than all males. Whatever actually constitutes "femininity", a certain amount of overlap is to be expected.

In Figure 2, the distributions have been drawn as normal distributions. In part, this is an illustrative convenience, and for purposes of argument it will be postulated that within each sex, the distribution around the mean would be normal. The area of overlap has been designated Area O.

If this model is trisected as in Figure 3, some questions about the relationships within this model can be introduced. The first is an obvious question which concerns the size of "O" (the area of overlap) relative to the distributions. In other words, are males and females very different, somewhat different, or only slightly different with regard to femininity? The outside areas in Figure 3, labelled A and C, describe a range of individuals who are respectively either less feminine than almost all females, or more feminine than almost all males.

The second question to be raised about the model pertains to the variances of the two groups. It seems almost axiomatic that sex roles are reinforced differentially from very early on in life (Maccoby & Jacklin 1974). As a consequence, it might be suspected that each of the sexes will show greater variance on the sex-appropriate dimension which is encouraged, than on the inappropriate dimension which is

subject to social sanction both within and outside of the family. Consequently, the range of feminine qualities that boys show may be much smaller than the range shown by females. One possibility is that within a particular sex, the quality is normally distributed, but the underlying units may be smaller for the non-congruent sex, so that the shapes of the distributions when placed on a scale of "standard hypothetical units" will be quite different. Strictly speaking, sex differences in the observed variance are a matter of conjecture, since it would require more accurate measurement devices than are currently available to validate them.

Yet a third question may be posed about the model that has been described. It may well be asked if masculinity and/or femininity are not only quantitatively different for males and females, but <u>qualitatively</u> different as well. Here, for purposes of illustration, the two sexes have been aligned on a single dimension in the model, but this may or may not represent the reality of the situation. Are masculine women different from non-masculine women in the same ways that masculine men are different from non-masculine men? The dimensions called masculinity and femininity might be quite different for males and females, due to several things: the different role expectations placed on them; the reflectiveness of male and female roles; the different notions they hold about sex-roles, stereotypes and their importance; and/or, the relative extent or range of the do-

mains of characteristics and behaviors involved. It may be unwise to assume that masculinity and femininity are the same for both sexes and make direct comparisons between members of different sexes when the actual relationship of the distributions, or even the precise definitions of the constructs are yet unknown.

These three questions about the abstract quality called "true femininity" are all posed for the same reason. Depending on the answers, different strategies for test building are necessary and different types of measures will result. The purpose of creating a femininity test is not to differentiate males from females. It is to differentiate between levels of femininity in individuals, from high to low. As simplistic as this sounds, it is not always clearly understood, and its implications are not infrequently ignored.

To the degree that femininity is either quantitatively or qualitatively different, the constructs are different for males and females. As a result, the same masculinity and femininity tests may not be appropriate for both sexes. Separate versions of the masculinity and femininity subscales may therefore be necessary. This is another way of saying that the domains from which items are selected will be different, even if overlapping. Where a single test is used to discriminate between levels of masculinity and/or femininity regardless of the respondent's sex, the implica-

tion is that the theoretical domains are the same for either sex. This is the same as saying that, for instance, the quality of "masculinity" is exactly the same for males and females. This is the unstated assumption on which the Personal Attributes Questionnaire and the Bem Sex Role Inventory are built. In its present form, the STAT also reflects this assumption.

The Domains of Masculinity and Femininity

In the example portrayed in Figure 3, F is itself regarded as an undifferentiated construct. But the notion of a single femininity construct has proven difficult to operationalize. More than a few segments of our lives contain categories which are sex-linked. The definition of masculine and feminine personality characteristics which is proposed here describes them as orientations to these male and female categories. It would probably be useful if masculinity and femininity were thought of as families of constructs rather than as pure linear dimensions in themselves (Locksley & Colten 1979). Thus, the hypothetical model presented earlier is not held to resemble any real-world phenomenon called "true femininity" as such. Rather it is a model which can be applied across a number of content areas or domains. Implicitly, this is advocating a group or family of many "masculinities" and "femininities" which can be measured, where masculinity and femininity as such cannot be (Spence & Helmreich 1979).

In order to bring the discussion of femininity as a vague, undefined, and hypothetical abstraction closer to earth, it may be useful to draw an analogy to a more familiar kind of psychological measurement, such as intelligence testing. One of the ways in which the measurement of masculinity and femininity resembles the measurement of intelligence is that in both cases, the measures must contain some form of content that can be conceptually distinguished from the "pure" theoretical entity to be measured or assessed. IQ tests are made up of a variety of subtests, usually asking questions requiring general knowledge, visuo-spatial ability, verbal comprehension, and so forth. None of these things are intelligence as such, but they each reflect intellectual functioning. Similarly, to measure masculinity and femininity, the researcher must ask questions <u>about</u> something: e.g., traits, interests, hobbies, vocational preferences, and so forth. None of these are "masculinity" or "femininity", but all reflect the notion of an internalized orientation to sex role within that sphere. By analogy, the measurement of M and F in only a single domain is as incomplete as the measurement of only one mode of intelligence. The trend toward measuring different aspects of M and F has been traced in Chapters II and III.

The prototypical feminine person or masculine person may be so in "thought, word, and deed" but the linkage is unclear when separate modes are examined separately. Real

people are not prototypes. In order to generate items it is necessary to define in fairly narrow and specific terms the domain from which the items are to be derived. In the area of masculinity/femininity this has proven to be a forbidding task. In the first place, different individuals place different emphases on particular aspects of the phenomena. Therefore, such things as looks, deportment, interests, activities, vocations, and traits have different degrees of importance for different individuals.

The choice of a particular domain to study will depend upon the phenomena which interest a researcher, and will determine the content of the questionnaire to be used. Some domains may be more important for certain uses than other domains. For example, feminine interests may be more relevant to some third variable, such as creativity, than feminine traits. Masculine demeanor may be more important in the study of person perception than masculine traits or interests.

One of the problems with earlier M-F tests was that they mixed items from many of these domains haphazardly. Comtemporary approaches emphasize the delineation of specific domains. The reasoning behind these approaches can be described by using the hypothetical example of sex-typed interests, which have traditionally been considered relevant to M or F. If a representative sample of masculine interests could be assembled, it would be a fairly straightfor-

ward task to determine by questionnaire an individual's level of orientation toward such interests. In the same way, measures of masculine orientation could be found in other potential domains, each measuring some identifiable aspect of the family of measurable masculine dimensions. It may yet be demonstrated empirically that a domain such as "interests" is itself too vague, and will have to be reduced to more elemental dimensions: sports interests, mechanical interests, business interests, and so forth. This kind of breakdown may result from the factor analysis of complex tests covering large domains which actually subsume more basic unidimensional entities.

To be more clear on this point, the example of masculinity may be used. It may be that in order to distinguish masculine from non-masculine individuals, it is necessary to assess an orientation toward, say, athletics. This is considerably more specific than the kinds of domains discussed thus far. Indeed, items taken from different domains may turn out to be pertinent to this construct. Consequently, trait items, activity items, and interest items may all be relevant to this task. Such developments can only grow out of the preliminary work of measuring these other, perhaps less cohesive domains, individually.

To this point, researchers have chosen to employ different kinds of items to measure theoretically important phenomena related to masculinity and femininity. The BSRI

asks subjects to rate how often certain traits are characteristic of them on a scale which ranges from Never to Always. The PAQ gives opposing trait descriptions and has the subject rate himself on a five point scale between them. The Sex Role Behavior Scale requires that the subject rate how characteristic a particular interest, vocation, or role behavior is. The Sex Typed Activities Test asks for ratings on the level of comfort across a number of behaviors. All of these have in common the use of similar items divided into masculine and feminine subcategories. Each scale exemplifies the content area pertaining to sex roles that the researchers found worth investigating. The test structure varies however among the tests in ways which seem to be arbitrary to some degree.

It has been argued that the test structure for M and F measures depends on the degree of overlap between the male and female distributions. The greater the difference between males and females in terms of a feminine construct, the more likely that more than a single form of the test will be required to assess differences related to that construct. Furthermore, it has been proposed that there are a multiplicity of feminine constructs, for example, which differ in terms of definition and content. The task of the researcher then is to build tests which operationalize specific areas of sex-linked personality characteristics through the choice of questions, rating scales, and items.

When describing the hypothetical domain called F, it was stated that the concept was to be considered in the abstract, and stripped of its complexities. It may be observed, however, that if the abstract model described in Figure 3 is applied to specific, content-oriented domains, the degree of overlap between the male and female distributions will actually vary depending on the domain. Consequently, the general rule which has been articulated will serve as a guide to the structure of tests within different content domains. While there is no hard and fast rule which dictates whether a test should have two subtests or four subtests (parallel forms for males and females), the final form of the test will depend on the nature of the domain.

In the domains where males and females are very similar in terms of what distinguishes the feminine from the non-feminine, a single test will suffice. But, if F is a very different concept for males and females, in degree or in kind, it will be necessary to create separate tests, with items drawn from different domains. In terms of the model, Area 'O', which describes the overlap will be largest where the same kinds of things distinguish feminine from non-feminine individuals of either sex. Where very different sorts of questions must be asked of the two sexes, to divide the high from the low feminine, Area 'O' will be quite small.

It is the nature of certain categories to evidence a large overlap between the sexes, and others, a relatively

small overlap. Where the content of the test is largely abstract, as in the case of the BSRI and the PAQ, the area of overlap can be considered quite large because the trait descriptions can be adjusted in terms of the set of the respondent to fit either sex (cf. Locksley & Colten 1979). Where items are more behavioral as is the case with the STAT and the Sex Role Behavior Scale-2, social norms are more pressing and the degree of identification evoked by a particular item is more likely to be a function of the sex of the respondent. As a consequence, there is a greater differentiation between the overall levels of feminine orientation for males and females.

The question may be raised whether it actually requires separate sets of items to accomodate such sex differences. It is possible that, up to a point, the numerical levels of the self-ratings would suffice to produce different levels of average ratings in males and females. However, again, the purpose of the test is not to discriminate between males and females, but to discriminate among levels of M or F. In order to pursue this example, it is necessary to take a closer look at the items which make up M and F tests.

The Hypothetical Item Continuum

Up to this point, the model of "F" that has been used has referred only to the differences in levels of femininity across persons. Just as there are different kinds of indi-

viduals with regard to masculinity and femininity, there are also different kinds of items. Hence, it is possible to postulate a hypothetical dimension similar to the one used to describe femininity in persons. In this case, the dimension represents the gradations of different items. A parallel can again be drawn with intelligence testing for illustrative purposes. The items at the high end of the scale totally fail to discriminate among individuals with low levels of intelligence; by the same token, there are certain items which can discriminate among the low levels, but not at high levels. By analogy, items that relate to the masculine and feminine dimensions within a domain can also be visualized as having a hierarchical relationship to one another. Again, items could--at least theoretically--be scaled in terms of the level at which they are useful for discriminating between individuals.

What I am proposing, then, is a hypothetical dimension, or rather, pair of dimensions analogous to the hypothetical dimensions represented in Figure 3 which related to people. This time however the dimensions involved represent the spectrum of items. Again, at the simplest level, three levels of items can be discriminated: (1) those which could be used to discriminate between low levels of femininity observable only in males; (2) those which would discriminate among members of either sex; and (3) those which represent such high levels of feminine identification than they would

only discriminate among females. What this means is that at the extremes, certain variables will show no variance for one sex. So, for example, if we ask both males and females how comfortable they would feel "wearing pantyhose", on a 1 to 7 scale, virtually all males would endorse 1, or Uncomfortable, which means there would be no variance on that item. The females, on the other hand, might say they are fairly comfortable, but vary considerably in their individual responses. In other words, this might be the type of item which could help to discriminate among high levels of femininity, which are observable only in females.

To illustrate this hypothetical continuum of items, the relative distributions of ratings have been diagrammed for male and female respondents on three test items: one from the middle range, one from the low F range, and one from the high F range (Figure 4). These patterns illustrate the relative shapes of the distributions which might be projected for a hypothetical item taken from those levels. The items toward the middle of the continuum show differences between the male and female distributions, and the shape of the distributions is essentially normal. This reflects the expectation that middle level items would apply equally well to males or females, although there would be aggregate differences.

If there is in fact a difference in the male and female distributions on the overall dimension, however, as was il-

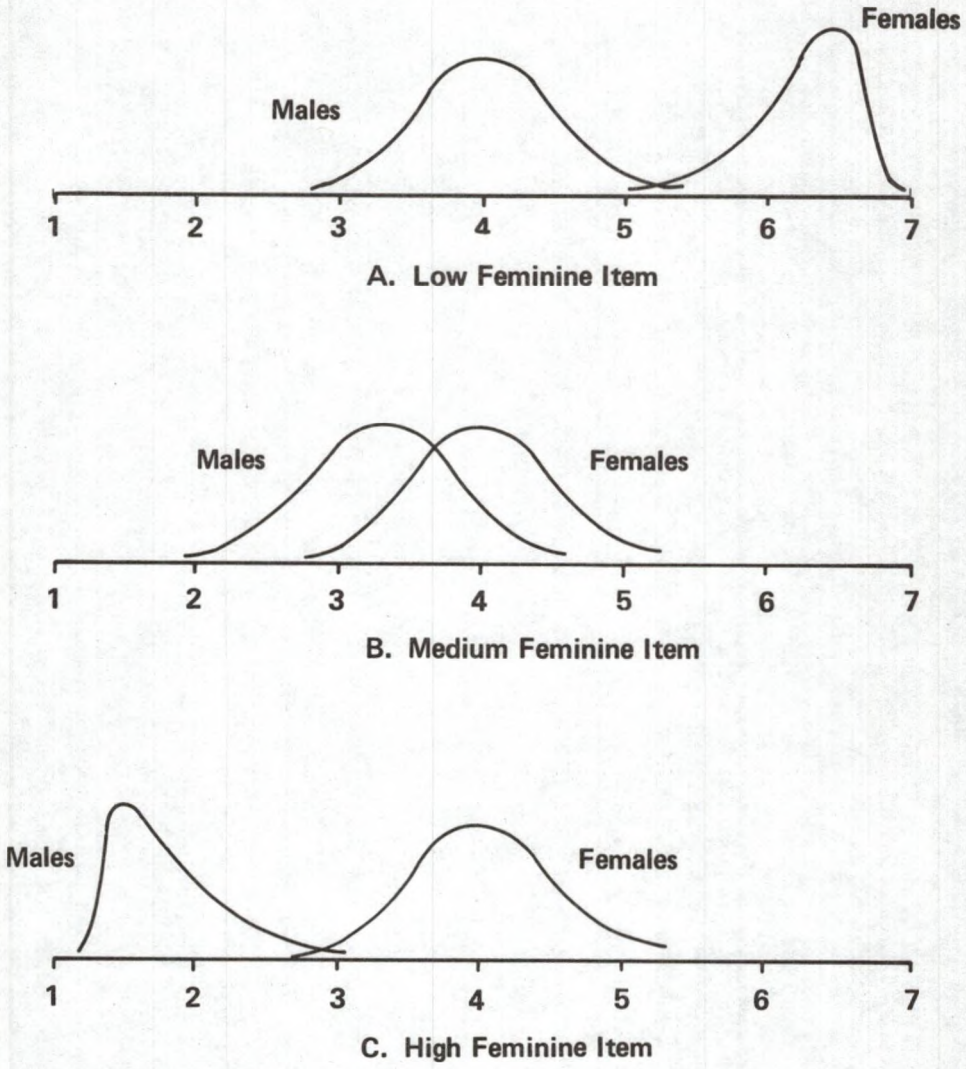


Figure 4. The distributions of male and female ratings on three items from the Hypothetical Item Continuum.

lustrated previously in Figure 3, then lower items on the hypothetical continuum, as illustrated in Figure 4 (A), would be less likely to discriminate between females, because these items reflect such low levels of femininity that almost any female would endorse them at a high level. Since males are in general less feminine, however, many of these low items would still serve to discriminate among levels of femininity in males. Consequently, the distribution for females has been drawn at the high end of the scale and highly skewed.

The reverse picture obtains as we give the respondents items reflecting higher and higher levels of F, as seen in the final drawing (3C). Here the items are more likely to discriminate between females, but less likely to discriminate among males, who show such low levels of femininity as a group that a ceiling effect takes hold. At the extreme ends of the continuum, the distribution of one sex or the other totally skews and flattens, as almost all members give themselves a rating of "1" or "7" on the item used for an example.

Thus we can draw an analogy between the continuum of persons and the continuum of items, with specific items corresponding to different levels of femininity observed across persons. The most significant feature of this model is the symmetry. All things being equal, it predicts a direct relationship between the placement of an item on the continuum

and the relative shapes of the distributions of male and female responses or self-ratings.

Good items on any test are those which elicit the most variance, provided that such variance is related to the task at hand--measuring a specified construct--and not unique or random. Items which fail to vary for one sex or the other are not necessarily bad items. They may be very good items for helping to discriminate among high feminine and low feminine members of one sex. But they are very bad items for the sex that shows little or no variance. To the extent that the sexes differ within a given content domain, the ranges of such sex-specific items are wider and, again, the more necessary it becomes to use separate versions of a test, or to create different tests to measure the same construct.

Thus, where many potential items entirely fail to discriminate within samples of one sex or the other, this does not mean they should be discarded. Such items may still be useful for one sex. If all potential items were to show variance for either sex, then it would not be necessary to consider using more than a single form of the test. Even in the case of tests based on adjective phrases describing traits, however, this is not true. Items like "Masculine" and "Feminine" result in such ceiling and floor effects. Pedhazur and Tetenbaum (1979) have noted the difference between these two items and the other test items on the BSRI.

Bem has also endorsed the viewpoint that these two items, by virtue of being closely correlated to sex, are the worst items on the BSRI (Bem 1979). According to the view presented here, this is not necessarily the case. They are not bad items, because they do elicit a certain amount of variance. But only the sex appropriate item should be administered to, or scored for, each sex. These two items are simply items which fall on the item continuum in a range corresponding to range 'C' on the person continuum in Figure 3.

The Actual Item Continuum

The hypothetical item continuum proposes that items can be systematically arranged in a hierarchical fashion according to the level of F or M which they represent. Turning again from the theory to practice, it will be shown that M and F items can indeed be scaled hierarchically, although the actual item continuum differs from the hypothetical model.

The idea that males and females will differ in regard to masculinity and femininity, illustrated in Figures 2 and 3, forms the basis for most conventional procedures for the selection of M and F items. The selection of items is achieved through comparisons of the actual item responses of males vs. females, actual heterosexual vs. homosexual responses, or stereotype ratings about males vs. females.

Criticisms of the use of actual sex differences have been cited in the review of the literature. But, in the current study, there was a strong veridical relationship between mean stereotype ratings and mean self-ratings for the final 58 STAT items (Appendix C). The items on which there were strong stereotype differences were also the ones on which there were strong differences between male and female self-ratings. It may also be pointed out that the use of stereotypes may involve its own pitfalls. This was evidenced by the ANOVA data presented in Chapter 3, which showed that some systematic differences may occur between the sex stereotypes reported by males and females. The item continuum model to be described here will apply to either self-ratings (test responses) or stereotype ratings. All that matters is that for a given set of items, there is one set of male ratings and one set of female ratings, which can be compared.

Whether actual differences or stereotype differences are used, items from a domain are sorted into categories: masculine, feminine, or neutral. Items on which there is no significant difference between the item means for the two groups (i.e., where the distributions are totally overlapping) are considered neutral. Items on which there is a significant difference, and the distributions are therefore somewhat separated, are classified as sex-typed depending on the direction of the difference. If males are higher, then

the item is masculine, and if females, feminine. Quite clearly, in making a categorical distinction, this procedure tends to obscure the fact that for any overall pool of items taken from a given domain, the degree of separation between the two distributions is a continuous variable, not a categorical one.

Unlike the hypothetical dimensions which are in theory separate and orthogonal, the Actual Item Continuum is bipolar, ranging from high masculine items to neutral to high feminine items. It is possible to visualize the dimension along which M and F items are scaled. In the middle of such a continuum are all those items which have nothing to do with sex. Examples might be an interest in TV, or how comfortable one feels "brushing your teeth", or any of the traits Bem calls "Neutral". By definition, these are items which show no distinction between the distributions of males and females, whether real self-ratings or stereotype ratings are discussed. At the endpoints of the continuum are items on which the distributions of scores are totally separated. Take for example, once again, the Bem items "Masculine" and "Feminine" which result in almost complete separation of the two distributions. Finally, there are all those sex-typed items which fall along the continuum between the neutral and the extreme items. These range from only mildly associated with sex to highly correlated.

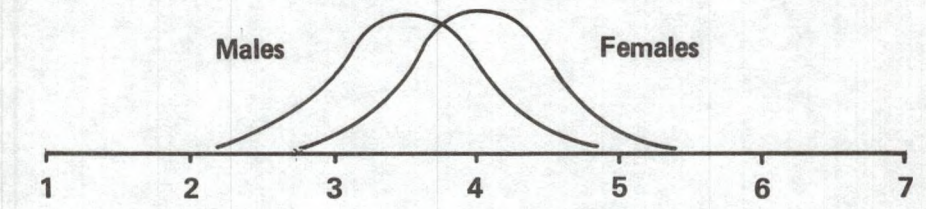
The relationship between the 'actual' and 'hypothetical' item continua resembles the relationship between a sample statistic and a population parameter. It is a way of approximating the model from the actual data collected within a content domain. Each half of the bipolar actual item continuum represents the corresponding hypothetical continuum. This distribution of items along a dimension ranging from F to Neutral to M has never been widely discussed. As a consequence, the implications of the continuum of items and its relationship to test structure have gone unrecognized. Items taken from anywhere on this scale may be useful M or F items within the context of the researcher's definition of the domain. This depends on communality among items. However, as we have seen, not all items will apply to both sexes. To the degree that the distribution of males and females within that domain are different, different items will have only limited usefulness for one sex or the other.

In contrasting the actual to the hypothetical item continuum, there are important differences. The relative distributions of males and females on three items representing low, medium, and high degrees of femininity on the hypothetical continuum were displayed in Figure 4. In Figure 5, it is possible to see how the actual model contrasts to the hypothetical model. Once again, the distributions of males and females have been graphed on a 1 to 7 scale, represent-

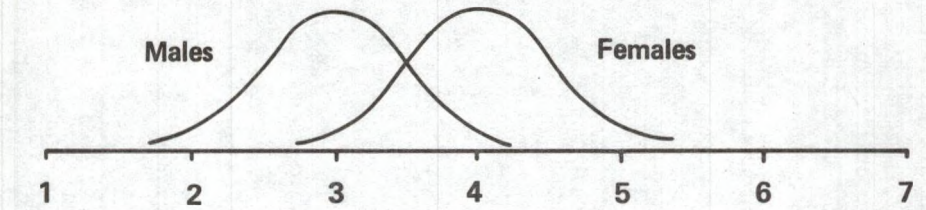
ing the range for three sample items taken from the feminine arm of the Actual Item Continuum. Just as in the hypothetical case, the middle range illustration shows a pair of overlapping normal distributions. As one approaches the endpoint of the scale representing the highest level of feminine items, again the females retain a normal distribution, but the male distribution tends to be skewed as males begin to uniformly endorse the item at low levels. High feminine items, to recap, are those which discriminate among levels of femininity found only in females, but not in males.

It is at the low end of the continuum that the Actual and the Hypothetical models differ. In the hypothetical model, the opposite pattern occurs when one goes down the scale to the low feminine items. However, in the actual item continuum model, as one moves down the scale, instead of a skewing of the female distribution, the distributions retain their shape and simply get closer together. At the extreme low end, the distributions converge and the item is classified as neutral. This is a way of saying that the item is orthogonal to sex and variance in the item will be the result of other sources of variance.

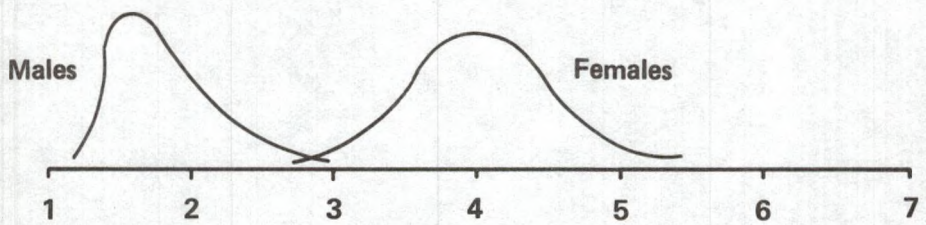
One way of thinking about the separation of the distributions is as a correlation to sex. The more separate, the stronger the correlation. In the hypothetical model, (Figure 4) that correlation would increase whether one moves up



A. Low Feminine Item



B. Medium Feminine Item



C. High Feminine Item

Figure 5. The distributions of male and female ratings on three items from the Actual Item Continuum.

or down the continuum. At either end one reached maximum differentiation. But in the actual item continuum, one half of the full continuum is used to represent the M dimension and one-half is used to represent the F dimension. Consequently, as one moves to the endpoint from the neutral point, the items are progressively more correlated to sex, and at the same time represent higher and higher levels of masculinity or femininity in that domain. Therefore, indicators such as the correlation with sex, the t-value between the means, and the associated probability levels are all rough indicators of the placement of the items along the scale.

The actual means of the distributions of course are not only the product of how feminine the item is, but also of other factors as well. The location at which an individual places his or her mark on the rating scale is determined by multiple factors. The actual rating represents the person's true score in combination with error variance. On the Sex Typed Activities Test, the principal source of extraneous variance pertains to how comfortable the person is in general, regardless of sex. For the BSRI or PAQ the extraneous or "method variance" derives from the willingness of subjects to endorse positive characteristics. Because, like intelligence testing, M and F measurement involves content, each domain has its own source of systematic error variance specific to that domain. Whatever the nature of that vari-

ance, it is significant because it accounts for the failure of the hypothetical model to apply. At low levels of F, the distributions for females do not skew toward the high end of the scale as predicted by the hypothetical model precisely because of the type of items which are used to measure masculinity and femininity. The responses of females at low levels continue to vary, but the basis of that variance no longer has much to do with femininity. The variance at those levels is principally related to extraneous factors or method variance.

Certain item selection procedures can bias the selection of potential items so that the final test items are chosen chiefly from the high or low ends of the item continuum. Such a bias can affect the normality of the subscales' distributions and the ways in which the scales interrelate. In the current study, it was originally projected that a single instrument would serve for either sex. Consequently, items were eliminated which were uncomfortable for the non-congruent sex on each scale. This had the systematic effect of eliminating items at the high end of the item continuum where the distributions were most clearly separated. The result was that the power of the test to discriminate between individuals at the high end of the person-continuum was compromised: the distributions of scores were skewed, but only for the congruent sex on each subscale.

When items are limited to the low range, the test necessarily incorporates more error variance relative to the M or F variance it is trying to maximize for measurement purposes. This has a number of deleterious effects in addition to the lack of normality.

The two-dimensional picture: Items

There are a number of instruments which measure M and F on separate scales. The word 'orthogonal' has often been used to describe the model upon which these measures are based. In reality, the degree to which M and F scores are correlated to one another on these different tests may vary considerably. The correlation between M and F scores, which should always be examined independently for the two sexes, can be a function of whether the items are chosen principally from the high or the low end of the actual item continuum. Tests can be classified according to the intercorrelations between M and F items which can be negative, positive, or uncorrelated.

Traditional M-F measures treat M and F items as if they were inversely correlated. Naturally, real inverse correlations between masculine and feminine items have therefore been seen as a good thing by researchers using these scales. An inverse correlation between groups of M and F items has also been marshalled as evidence of the need for an M-F dimension on the Sex Role Behavior Scale-2 by Orlofsky et. al.

(in press). Storms (1979) created a bipolar measure of "Sex Role Identity" which is basically composed of variations on the adjective items "Masculine" and "Feminine" from the BSRI. The M and F scales based on the Adjective Check List also show a negative correlation between M and F (Heilbrun 1976). In all of these cases, the theoretical explanation for having a bipolar scale is connected to the overall notion of masculinity and femininity as opposites. I have argued that this confounds M and F as personality dimensions with the schema which individuals use to judge masculinity/femininity in others. In general, M-F measures are in fact compound measures which relate to both the M construct and the F construct, but use only items from the high ends of the two continua. The fact that they are inversely correlated is not evidence that they belong on a single scale. Put very simply, to the extent that one chooses items that are highly correlated to sex, one selects items that will be inversely correlated to one another.

Examples of tests which show positive correlations between separate M and F scales are the Sex Typed Activities Test, already reviewed, and the Sex Role Behavior Scale-2. In this latter case, all of the items have been divided into four categories and three subtests developed. Masculine items and feminine items which are inappropriate for the opposite sex have been consigned to the bipolar M-F scale. All other items fall on the separate M and F scales. In

terms of the model, this means that all high M and High F items are placed on one scale because they are positively correlated to sex and negatively correlated to one another. The result is that the M and F subscales are composed only of items from the low M and low F ranges of the two continua. Predictably, these show the highest positive correlations between the M and F scales of any available measures as well as dismally low inter-item correlations among the F items and the M items, respectively.

The Bem Sex Role Inventory is the principal example of an orthogonal pair of M and F subscales. The PAQ M and F subscales are also largely orthogonal but sometimes, especially in the original 55-item version, have shown positive correlations (cf. Spence, Helmreich & Stapp 1975). According to this model, it seems likely that the item selection procedures in this case were such that a balance of high, medium, and low items on the item continua compose these scales. The PAQ M-F scale, by the way, is a special case. It is not composed of two groups of High F and High M items. Rather, since the individual items each contain two descriptors, one at each end of the rating scale, these are primarily items with a male-valued term on one end and a female-valued term at the other. As such, these are "bipolar scales in miniature". The factor analytic data presented by Helmreich, Spence, and Wilhelm (1981) show that these items load on both the M and F factors, evidence that they are complex items.

The selection of items from the continuum in a biased manner would also have implications for the factor structure on M and F measures. Where items are selected primarily from the low end of the continuum, the salience of alternative sources of response variance is enhanced. Consequently, the kind of two factor structure which might be desired in a test with M and F subscales is compromised as variables tend to sub-group according to other meanings which they incorporate in addition to sex-appropriateness. When items are selected primarily from the high end of the scales, a bipolar factor may appear which primarily reflects gender.

It can be seen that researchers must take great care to be certain that the appropriate range for each sex on M and F items is completely sampled. If item selection procedures systematically bias the choice of items, the test characteristics will fail to conform to the theoretical model of orthogonality and two-factor structure in a manner that may be predictable in advance.

Masculine, Feminine and Androgynous Persons

Another way of discussing the two-dimensional picture is to return to the idea of a continuum of persons, such as the one diagrammed for a single dimension in Figures 2 and 3. The dimension was divided into three areas, one covering persons who were more feminine than all males (C), one for persons who were less feminine than all females (A), and an

overlapping area (B). In Figure 6, the M and F dimensions have been mapped out in relation to one another according the view that these represent independently varying orientations to male and female categories. Area Q represents the individuals whose M and F scores are both within the overlapping range. Area P represents all other males, and area R, all other females.

Again, depending on the domain to which this model is applied, the size of the overlap will vary. Where the overlap portion 'B' along the individual dimensions is quite large, area Q will be larger, and it will be possible to describe individuals as androgynous in the sense of having high scores on both M and F dimensions. However, where Q is rather limited, the androgynous group will be limited in terms of how high their scores can be in absolute terms on the sex-incongruent subscale. Thus, androgyny is a concept whose usefulness may be more important in some domains than in others.

One of the implications of this model is that to the degree that the two sexes are very different within a particular domain, there is a limit imposed on how truly "androgynous" a given person is likely to be. In addition, the whole notion of various domains of M and F, which differ in terms of their empirical characteristics, suggests that outside of a very narrow range of meaning, the term androgyny may have limited descriptive power.

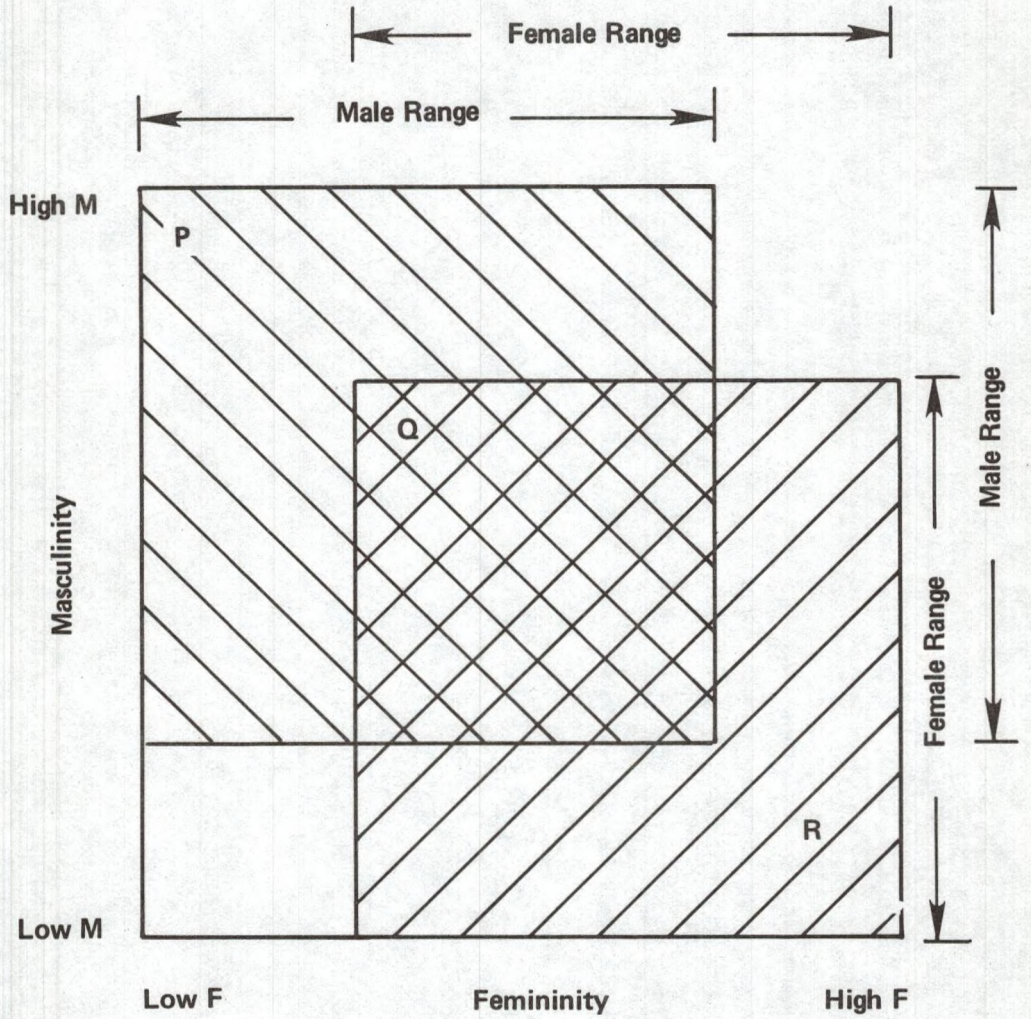


Figure 6. The potential ranges of scores on masculine and feminine dimensions for males and females.

Summary. Masculinity and femininity have been defined as orientations to male and female categories of behaviors, traits, and other personal characteristics. As such, they represent families of characteristics rather than unitary linear dimensions. If we assume that masculinity and femininity are separate dimensions within a variety of content domains, we can then explain why it is that particular item selection procedures will result in certain relationships among the variables. These predictions concern the scaling of items, internal distributions of scale scores, inter-score correlations, and factor structure. The model presented here, in that it is hypothetical, is subject to the empirical validation of its predictions across various item domains. However, the data presented here regarding the STAT, and other scales, do support these hypothetical relationships.

Limitation of space does not permit more extensive examination of all of the implications of the model. However, two things can be said for it. First, this model provides a more complete and flexible definition of the constructs based on the knowledge that has accumulated in the investigation of M and F as separate quantities. And second, it defines specific avenues of investigation which should result in a more coherent and comprehensive view of internalized sex roles.

The Theoretical Model and the STAT

The relevance of this model to the data presented in regard to the Sex Typed Activities Test can be summarized succinctly. Indeed this model is the offspring of the process of inquiry which began with the finding of a positive correlation between STAT M and F scales.

The model suggests that the item selection process in which stronger items were deleted from the test in order to make the same form of the test applicable to both male and female subjects, was in error. When this test was conceived, it was not clear that the comfort-domain might be very different from the trait-domain. The net effect of this inadvertent decision to eliminate strong items was to create scales that incorporated excessive "method variance" leading to correlated scales. This probably in turn affected the factor structure by virtue of the fact that other sources of variance became as important as M and F themselves. It also truncated the range of items in a way which led to skewed distributions, but only on the sex-congruent scales.

Clearly, this model, to the extent that it represents reality, can be used to improve and revise the STAT in terms of its structure and component items. Only after such revisions can the STAT be confidently used as a research tool, without reservation.

APPENDICES

Appendix A

THE SEX ROLE BEHAVIOR SCALE: A REVIEW

In 1981, Orlofsky first introduced a 160-item version of the Sex Role Behavior Scale (SRBS-1). An expanded 240-item version (SRBS-2) was subsequently described in a second article by Orlofsky, Ramsden, and Cohen (in press). The Sex Typed Activities Test, which is described in this paper, parallels the SRBS-2 in terms of theoretical rationale. Both are attempts to extend the dualistic notion of M and F measurement into the behavioral realm. Both are based on the general stereotypes held by both males and females. The strengths and weaknesses of the STAT have already been described at some length. This appendix has been added to examine the SRBS-2 more fully than could be done in the body of this paper. It also is an opportunity to present certain data pertaining to the SRBS-2 which relate it to the theoretical model presented in Chapter 6.

Borrowing its terminology from the early PAQ, The SRBS-2 contains 80 "male-valued" (M) items, 80 "female-valued" (F) items, and 80 "sex-specific" (M-F) items. Each of these 3 scales contains 4 subtests: Recreational interests, Vocational preferences, Social and Dating Behaviors, and Marital Behaviors. Consequently, for a particular respondent,

the entire test delivers 12 subtest scores and 3 overall scores. It should be apparent that although this test incorporates the quasi-dualistic approach to measurement of Spence and Helmreich's PAQ, it is an omnibus measure in terms of content. It combines a series of subtests of various content into overall scores.

The SRBS-2, it will be argued, is susceptible to both empirically-based and a priori criticism over two issues. The first concerns the decision to combine subtest scores (e.g. Recreational Interests, Marital Behaviors) into overall or total M, F, and M-F scores for each respondent. The second concerns the decision to have three scales, including a bipolar M-F scale, and to adhere to that breakdown for all of the subtests as well as total scores.

The fundamental basis for combining subtests into overall scores is the same as that for combining items into tests. It must be demonstrated that the components are unidimensional--they should be assessing various aspects of the same "thing". This means that to some minimal degree the components are intercorrelated. Using this criterion, there is little justification for summing the four separate subtest scores on the SRBS-2 to yield overall M, F, and M-F scores. Although it is true that reliability coefficients for the overall scores are high, alpha coefficients are composed of two elements: test length and average inter-item correlation. Consequently, on a very long test such as the

SRBS-2, high reliability is to be expected even with extremely low levels of relationship among the component elements.

Using data presented by Orlofsky (1981) concerning the reliabilities of the overall M, F, and SS (sex-specific, or M-F) scales for males and females (Table 29) this point can be graphically demonstrated. Using the Spearman-Brown formula (Nunnally 1978), average inter-item correlation coefficients have been computed for each of the three scales. For the two sexes, these range from .02 to .08--which suggests extremely minimal correlations across the entire range of items. Subtest reliabilities are not reported in the 1981 article because some of them were unacceptably low. It was largely because of this fact that some subtests were expanded and the test lengthened to 240 items in the revised version.

Orlofsky also presents reliability coefficients for the sexes combined. These will not be analyzed since combining the sexes often leads to inflated correlations as a result of confounding sex differences per se with variance on the variables of interest themselves (i.e., M and F). That is to say, when the two sexes are pooled for such analyses, "between-sex" variance is merged with "within-sex" variance on variables that are sex-differentiated by definition. This is not a defensible practice.

TABLE 29

Original SRBS-1: Reliabilities and Inter-item correlations

	Coefficient alpha*		Mean inter-item correlations	
	Males	Females	Males	Females
Male-valued (M)	.81	.82	.04	.06
Female-valued (F)	.78	.59	.03	.02
Sex-specific (MF)	.88	.87	.07	.08

*Note: Alpha coefficients taken from Orlofsky (1981). N= 95 males and 72 females.

Turning to the revised version of the test, the SRBS-2, a more complete picture of the problem can be observed. In Table 30, the reliabilities given are again taken from the work of the original authors, Orlofsky Ramsden, and Cohen (1981). The average inter-item correlations have again been computed and listed alongside the reliabilities for both male and female respondents. The mean subtest reliability is .79 for males and .74 for females. The reliabilities range from .49 to .91, and are for the most part acceptable in range. However, again, the average inter-item correlations for the overall M, F, and SS(MF) scales are quite low. For the males, these are .11, .07, and .06 re-

spectively; and for the females, the corresponding figures were .08, .07, and .06. This suggests that when all of the items on the M, F, and SS subtests are considered together the average level of intercorrelation among the individual items is negligible. Therefore, the combination of all subtest items into single overall scores may be much less desirable than simply using the four subtest scores separately.

As part of the Sex Typed Activities Test validation study, 37 males and 98 females completed the M-scale and the F-scale items from the SRBS-2 (see Chapter 5). Unfortunately, SS or M-F items were deleted in this case. A correlation matrix was then computed between the self-rating scores for all eight subtests separately for male and female respondents.

Tables 31 and 32 list the correlations among the content subtests for the M scale and F scale respectively. It can be clearly seen that although some subtests correlate at respectable levels, some are not correlated at all. The mean correlation for the male group among the four M subtests was .21. For females, the corresponding figure was .07 which is quite low. For the F subtests (Table 32) the average for the correlations within the matrix is .31 for males, and again, lower for the females, .17. It would be difficult to argue for the unidimensionality of these subtests, especially for female respondents.

TABLE 30

Revised SRBS-2: Reliabilities and Inter-item correlations

Subscales	n	Males		Females	
		Coeff. Alpha	Inter-Item r	Coeff. Alpha	Inter-Item r
Male-valued					
Overall M	80	.91	.11	.88	.08
Recreational Int.	14	.91	.42	.88	.08
Vocational Pref.	16	.75	.16	.80	.20
Social/Dating	12	.82	.28	.73	.18
Marital Beh.	38	.90	.19	.85	.13
Female-valued					
Overall F	80	.87	.07	.86	.07
Recreational Int.	10	.62	.14	.65	.15
Vocational Pref.	12	.79	.23	.74	.19
Social/Dating	18	.81	.19	.67	.10
Marital Beh.	40	.85	.12	.88	.15
Sex-specific					
Overall SS	80	.84	.06	.83	.06
Recreational Int.	16	.49	.06	.59	.08
Vocational Pref.	20	.63	.07	.59	.07
Social/Dating	16	.64	.10	.64	.10
Marital Beh.	28	.88	.20	.83	.15

Note: Alpha coefficients reported in the statistical appendix provided by Orlofsky, Ramsden, and Cohen for their article (in press).

TABLE 31

SRBS-2 Male Valued Subscales: Intercorrelations

	Vocational Preference	Social/Dating Behaviors	Marital Behaviors
Recreational Interests			
Males	.48***	.00	-.15
Females	.37***	.37***	.05
Vocational Preferences			
Males		.20	-.33*
Females		.26**	.04
Social/Dating Behaviors			
Males			.20
Females			.18*

Note: North Dakota sample: 37 males, 98 females.

*p <.05
**p <.01
***p <.001

TABLE 32

SRBS-2 Female Valued Subscales: Intercorrelations

	Vocational Preference	Social/Dating Behaviors	Marital Behaviors
Recreational Interests			
Males	.49***	.60***	-.15
Females	.44***	.40***	.27**
Vocational Preferences			
Males		.35*	-.04
Females		.36***	.21*
Social/Dating Behaviors			
Males			-.22
Females			.19*

Note: North Dakota sample: 37 males, 98 females.

*p < .05
**p < .01
***p < .001

The uneven and variable pattern observed here is comparable to the data reported by Orlofsky et. al. There would appear to be a contradiction in the rationale for the development of the SRBS and the use of total scores over these empirically unrelated subscales. In creating a test of male-valued and female-valued behaviors, Orlofsky (1981) justifies himself by observing that various aspects of masculine and feminine sex-roles are not necessarily tightly intercorrelated. Consequently he argued for the creation of a behavioral test of sex roles as opposed to trait measures of M and F. However, in compiling total scores from subtests with varying degrees of interrelationship he fails to follow this logic to its natural conclusion.

The logical extrapolation from low intercorrelations among the behavioral subscales of the SRBS-2 would be to use the subscales separately and dispense with the computation of overall scores. Orlofsky instead emphasizes the fact that the scales are slightly intercorrelated, drawing a conclusion that seems to be unjustified on the basis of the data. In his discussion, Orlofsky even makes note of the fact that some of these correlations between the various subtests composing the M or F scales were not significant:

> These findings suggest that individuals are at least partially consistent in their sex typing across behavior areas. They further suggest that the overall scales have utility as general indices of individual's adherence to sex-stereotyped behaviors. However, the small magnitude of some of these inter-area relationships suggests that sex role behaviors are certainly not unidimensional.

> Thus, the area subscales can and should be used separately as well as collectively. (emphasis added) (Orlofsky, Ramsden & Cohen; in press)

Clearly, one cannot have it both ways. Either the scales are unidimensional and the overall scores are meaningful or they are not unidimensional and must be treated as independent scales measuring different 'things'. "Partial consistency" among the subtests is not the same as unidimensionality.

Correlations between M and F scores

In their famous paper on convergent and discriminant validity, Campbell and Fiske (1959) argue that validity is not only a matter of resemblance between different sets of scores purporting to measure similar things, which they called convergent validity, but that equally important is the discrimination of a phenomenon from other phenomena which are in theory supposed to be unrelated. Using the same logic provides an interesting perspective on the structure of the SRBS-2. Since scores are combined vertically over the subtests with varying content within the three categories of M (male-valued), F (female-valued), and SS (sex-specific), one would expect that the subtests within each category would correlate rather highly. (E.g., male-valued interests would correlate with all other male-valued quantities to a high degree.) Furthermore, one would anticipate that the M and F scales would be uncorrelated, since there

is no theoretical rationale for assuming that they would be positively correlated, and the tripartite structure of the test implies a model which is not completely unidimensional and bipolar. As the test now stands, however, these assumptions would be wrong.

Subtests of the M and F scales, of course, are summed to create the overall M and F scores. The magnitude of these correlations between same-scale subtests can be contrasted directly to the correlations between M and F scores within each of the four content areas. When M and F scores are correlated for each of the four content areas, the coefficients range from .35 to .57. The male average is .49 and the female average is .36. These correlations are available in Table 33. If the magnitude of the correlations is used as a gauge of unidimensionality among the subtests, it would seem to make more sense to combine scales horizontally acheiving a total for each content area, rather than vertically along the M and F division which is the rationale for the test. Clearly however this would defeat the purpose of the test.

The data which have been described here elicit more doubt than confidence in the rationale and structure of the SRBS-2. The combination of subtest scores into overall M, F, and M-F scores is clearly unjustified by virtue of a lack of demonstrated unidimensionality, reflected in the low overall inter-item correlations and the highly variable inter-subscale correlations.

TABLE 33

Male-valued vs. Female-valued Scale Correlations

	Male Valued Interests	Male Valued Vocational Preference	Male Valued Soc/Dat Behaviors	Male Valued Marital Behaviors
Female-valued Recreational Interests				
Males	.38**	.54***	.22	-.17
Females	.36***	.29**	-.02	.01
Female-valued Vocational Preferences				
Males	.26	.57***	-.02	-.43**
Females	.23*	.37***	.13	-.02
Female-valued Social and Dating Behaviors				
Males	.12	.41**	.52***	-.07
Females	.32***	.20*	.35***	-.07
Female-valued Marital Behaviors				
Males	-.24	-.20	-.24	.50***
Females	-.19*	-.06	-.04	.37***

*p <.05

**p <.01

***p <.001

In addition to the derivation of total scores from subtests covering various domains, the SRBS is, in my view, subject to criticism on the basis of using a tripartite subscale structure rather than a simple dualistic structure. The inclusion of a third, bipolar M-F scale is a conceptual and theoretical error. Orlofsky chose the PAQ as the model for this test rather than the BSRI. The reasoning behind the decision to do so deserves at least cursory examination.

Orlofsky (1981) notes the failure of the item selection procedure for the Bem Sex Role Inventory to distinguish between items which are more socially desirable for one sex than the other and items which are simply less undesirable for one sex than the other. Consequently, he based the item selection procedure for the SRBS-1 on the methods used for the PAQ which (1) examined beliefs about how the typical young adult male and female differed vis a vis individual items, and (2) divided these sex-stereotyped items into subscales based on ratings of how desirable or appropriate the items were for each sex. This created three scales, the Male-valued (more typical of males, but desirable and appropriate for either sex), Female-valued (more typical of females, but desirable and appropriate for either sex), and Sex-specific (typically different and also different in terms of desirability depending on sex).

If the item selection procedure for Orlofsky's SRBS-2 is re-examined in light of the theoretical model presented

in Chapter 6, it can be argued that the items were effectively divided into four categories. Two of these categories were then combined to make up the bipolar M-F scale. The items which form the separate Male-valued and Female-valued scales are more typical of one sex than the other, but they are appropriate or desirable for either sex. It might be hypothesized therefore that these items would tend to fall toward the low to medium end of the Item Continuum, since items at the high end are those which only discriminate for individuals of one sex. The degree of correlation between each M and F score on the four subtests suggests that this might be the case. The theory suggests that scales composed of lower items tend to be correlated due to the large amount of method variance involved which pertains to the content area rather than M or F.

By contrast, the items of the M-F or Sex-specific scale are those which are both more typical of one sex than the other, and appropriate or desirable for the more typical sex. In the original report, Orlofsky (1981) gives the correlation between the SS(M) and SS(F) items as -.69, $p < .0001$, for the two sexes combined. The fact that this correlation is not broken down by sex leads one to suspect that it is inflated by sex differences. Nevertheless, it seems likely that by virtue of being appropriate for only one sex, these items fall at the high ends of the M and F item continua, which might account for their negative correlation.

Summary. The theoretical model would suggest that a re-examination of the structure of the SRBS-2 needs to be made to see if separate forms of the test are indicated for males and females, in place of the use of three subscales, one of which (MF or SS) is a compound measure of dubious validity. In any case a clearer division of the content areas needs to be made and the separate content domains can not justifiably combined for overall scores. As it stands now, the SRBS-2 is relatively uninformative and its validity can be questioned from a number of perspectives.

Appendix B

STEREOTYPE RATINGS

Suggestion Form

DIRECTIONS:

I need to collect a large number of potential items for a scale that I am creating. The purpose of the scale is to measure how uncomfortable or comfortable an individual feels in performing certain common everyday behaviors that are either "masculine", "feminine", or "Neutral".

Very uncomfortable or awkward					Very comfortable; Not awkward at all.	
/	/	/	/	/	/	/
1	2	3	4	5	6	7

Individuals will be asked to use the above scale to rate how they would feel performing each activity or behavior.

Your job is to help me think of more common, everyday behaviors like those listed below. "Masculine" behaviors should be those considered more appropriate in American society for males. "Feminine" behaviors should be those considered more appropriate in American society for females. females.

Please be creative. The more different behaviors I can collect, the easier my job will be. Thanks for helping.

Masculine Examples:

Using a screwdriver
Changing the air filter of your car
Painting a door
Taking out the garbage

Feminine Examples:

Ironing clothes
Using a hairdryer
Changing a diaper
Putting flowers in a vase

Neutral Examples:

Locking the door
Driving a car
Sharpening a knife
Reading the newspaper

'Stereotype Rating Form'

Name_____ Age_____
Year in School_____ Sex: M F
Major_____

DIRECTIONS:
Please estimate how comfortable or uncomfortable the typical American adult would feel in performing each of the behaviors on the following pages. Use this scale:

1 2 3 4 5 6 7

Very uncomfortable or very awkward — Very comfortable; Not awkward at all

Three answer sheets have been provided for this purpose. Start with Item 5 on Answer Sheet #1. When you have completed Items 5-100, go to Answer Sheet #2, and then go to #3.

Use a #2 pencil ONLY

Thanks for your cooperation.

1. Reading science fiction
2. Combing your hair
3. Buying a gift for your mother
4. Cleaning the bathtub
5. Typing a letter
6. Rearranging the furniture
7. Lifting weights
8. Ironing a shirt
9. Knitting a scarf
10. Getting the mail
11. Doing crafts
12. Walking the dog
13. Setting the table
14. Using an electric drill
15. Replacing a washer in a leaky faucet
16. Mowing the lawn
17. Driving a sports car
18. Putting flowers in a vase
19. Visiting a friend in the hospital
20. Shopping for clothes
21. Barbecuing ribs
22. Crying in private
23. Replacing the plug on an electric cord
24. Peeling an orange
25. Comforting a child
26. Getting your hair curled
27. Eating a steak
28. Listening to music
29. Scrubbing a floor
30. Reading the sports page
31. Assembling a bicycle for a child
32. Watching television
33. Re-potting a plant
34. Getting rid of a dead mouse
35. Computing your income tax
36. Setting up appointments
37. Starting a fire in the fireplace
38. Going camping alone
39. Weeding a flower bed
40. Baking a cake from a mix
41. Packing for other family members
42. Packing your suitcase
43. Being a volunteer
44. Wrapping a present
45. Sending an anniversary card
46. Crying over a TV show

47. Pruning a tree limb
48. Laughing at a cartoon
49. Brushing your teeth
50. Cutting hair
51. Writing checks to cover bills
52. Changing a tire
53. Playing poker
54. Balancing a checkbook
55. Opening a door for someone else
56. Using a tiller to plow up a garden
57. Shaking hands
58. Babysitting for money
59. Raking leaves
60. Planting a vegetable garden
61. Shopping for children's clothes
62. Washing the car
63. Sweeping the driveway
64. Planning a menu
65. Changing a fuse
66. Writing a check
67. Going dancing
68. Looking for a new job
69. Carrying a handkerchief
70. Feeding the dog
71. Making coffee
72. Planning a party
73. Taking a bath
74. Crocheting
75. Reading a murder mystery
76. Using a screwdriver
77. Going to the laundromat
78. Playing softball
79. Making a bank deposit
80. Using a hammer
81. Using deodorant
82. Sharpening a knife
83. Taking a snapshot
84. Buying eye make-up
85. Dyeing your hair
86. Talking to a kid's teacher
87. Reading the business page
88. Jogging
89. Choosing a paint color
90. Watching the news on TV
91. Talking about sex
92. Building a model plane
93. Bathing a baby
94. Move a couch
95. Using cologne

96. Sewing
97. Watching basketball on TV
98. Picking up a child at school
99. Shampooing a child's hair
100. Playing Monopoly
101. Using a snowblower
102. Driving a pick-up truck
103. Carrying a packet of Kleenex
104. Getting your hair styled
105. Going fishing
106. Driving a car with a stick-shift
107. Folding clothes
108. Reading the newspaper
109. Repairing a toaster
110. Buying a wedding gift
111. Playing Bridge
112. Shaving your legs
113. Buying linens
114. Letting your spouse cook dinner for you
115. Riding a bicycle
116. Buying a used car
117. Playing pool
118. Going to a party
119. Taking a child to the dentist
120. Reading the society page
121. Going to a bar alone
122. Drinking a beer
123. Sorting laundry
124. Washing clothes
125. Buying a new car
126. Picking up a hitch-hiker
127. Loading car for a trip
128. Making ice-cubes
129. Going shopping for hours
130. Hugging a friend
131. Painting a door
132. Writing letters
133. Building shelves
134. Going to the movies
135. Changing sheets
136. Taking dictation
137. Using hairspray
138. Changing the car's oil
139. Picking flowers
140. Reading a clothing magazine
141. Cleaning house
142. Building a dog house
143. Playing with dominoes
144. Locking a door
145. Playing touch football

146. Making the bed
147. Having a shower party
148. Planting a flower garden
149. Replying to an invitation
150. Reading Playgirl
151. Reading Playboy
152. Swearing
153. Cleaning the rain gutters on a house
154. Taking prescription medication
155. Taking a shower
156. Climbing a tall ladder
157. Going on a trip alone
158. Going to the PTA
159. Buying a record
160. Sunbathing
161. Picking up the tab in a restaurant
162. Handling family finances
163. Getting up with a baby at night
164. Taking out the garbage
165. Sawing a board in half
166. Driving a boat
167. Mending socks
168. Cleaning out the garage
169. Helping a child get ready for school
170. Embroidering
171. Shovelling a sidewalk
172. Painting the house
173. Installing a window air conditioner
174. Dusting a table
175. Using a power saw
176. Checking the oil in a car
177. Oversleeping
178. Making dinner for company
179. Watching football on TV
180. Going to the dentist
181. Making a phone call
182. Cleaning a stove
183. Opening a tight jar-lid
184. Driving a car
185. Playing catch with a kid
186. Hanging pictures
187. Cleaning a fish tank
188. Washing the dishes
189. Changing a car's air filter
190. Changing a baby's diaper
191. Playing tennis
192. Buying car insurance

193. Using a blow dryer
194. Weeding a garden
195. Buying new dishes
196. Chopping firewood
197. Reading a gossip column
198. Reading to a child
199. Giving a bottle to a baby
200. Wearing high heeled shoes
201. Sharpening a pencil
202. Riding a motorcycle
203. Mopping the floor
204. Planting a tree
205. Defrosting the refrigerator
206. Asking someone for help
207. Trimming a hedge
208. Pumping your own gasoline
209. Building a simple table

TABLE 34

M items: Typical Male and Female Means

This table lists the mean ratings made for the three target groups by both males and females for the 70 preliminary Masculinity items. Items listed in descending order based on eta^2.

	Item	Typical M	Typical F	Typical A
7	Lifting weights	6.22	3.02	3.82
15	Replacing a washer in a leaky faucet	6.24	3.30	3.67
14	Using an electric drill	6.24	3.18	3.83
23	Replacing the plug on an electric cord	5.93	3.27	3.35
196	Chopping firewood	6.12	3.50	4.12
126	Picking up a hitchhiker	4.17	2.16	2.45
175	Using a power saw	5.88	3.20	3.98
138	Changing the car's oil	6.07	3.40	3.87
142	Building a dog house	5.75	3.45	3.73
189	Changing a car's air filter	6.22	3.82	4.10
52	Changing a tire	6.17	3.88	3.85
173	Installing a window air conditioner	5.82	3.36	3.88
109	Repairing a toaster	5.41	3.20	3.39
31	Assembling a bicycle for a child	6.03	3.80	4.36
101	Using a screwdriver	6.07	3.84	4.23
153	Cleaning the rain gutters on a house	5.54	3.48	3.62
151	Reading Playboy	6.02	3.70	4.56
30	Reading the sports page	6.41	4.45	4.85
209	Building a simple table	6.02	4.02	4.37
179	Watching football on TV	6.53	4.43	5.22
38	Going camping alone	4.68	2.53	3.10
34	Getting rid of a dead mouse	5.04	3.02	3.04
97	Watching basketball on TV	6.29	4.27	4.65
53	Playing poker	5.92	4.04	4.04
47	Pruning a tree limb	5.90	4.22	3.97
56	Using a tiller to plow up a garden	5.41	3.63	3.49
65	Changing a fuse	6.10	4.27	4.65
202	Riding a motorcycle	6.10	4.41	4.32
133	Building shelves	5.76	3.93	4.11
78	Playing softball	6.31	4.90	4.78

166	Driving a boat	6.17	4.45	4.81
105	Going fishing	6.46	4.81	5.20
121	Going to a bar alone	4.68	2.61	3.20
176	Checking the oil in a car	6.44	4.72	4.78
125	Buying a new car	6.35	4.98	5.07
172	Painting the house	6.00	4.36	4.54
157	Going on a trip alone	5.27	3.70	3.63
117	Playing pool	6.14	4.62	5.02
106	Driving a car with a stick shift	6.39	5.31	4.88
80	Using a hammer	6.54	5.09	5.34
207	Trimming a hedge	5.73	4.54	4.30
102	Driving a pick up truck	6.49	5.09	5.04
156	Climbing a tall ladder	5.46	4.25	3.80
17	Driving a sports car	6.65	5.68	5.39
76	Using a screwdriver	6.61	5.25	5.37
152	Swearing	5.24	3.69	4.15
92	Building a model plane	5.27	3.75	4.02
208	Pumping your own gasoline	6.50	5.40	5.29
116	Buying a used car	5.31	3.97	4.36
122	Drinking a beer	6.31	4.86	5.63
183	Opening a tight jar-lid	6.24	5.13	5.14
145	Playing touch football	5.97	4.50	4.78
16	Mowing the lawn	6.32	5.11	5.13
87	Reading the business page	5.48	4.61	4.00
161	Picking up the tab in a restaurant	6.07	5.09	4.90
192	Buying car insurance	5.73	4.61	4.35
171	Shovelling a sidewalk	6.12	4.93	5.00
37	Starting a fire in the fireplace	6.19	5.04	5.23
35	Computing your income tax	4.95	4.36	3.51
88	Jogging	5.85	5.11	4.66

The following items were disallowed by virtue of having an eta^2 value below .10:

82	Sharpening a knife	5.70	4.84	4.45
57	Shaking hands	6.19	5.36	5.39
185	Playing catch with a kid	6.29	5.32	5.88
184	Driving a car	6.75	6.27	6.00
162	Handling family finances	5.92	5.23	4.95
204	Planting a tree	5.72	4.79	4.83
66	Writing a check	6.54	6.11	5.64
114	Letting your spouse cook dinner for you	6.36	5.81	5.42
79	Making a bank deposit	6.51	6.00	5.72

TABLE 35

F items: Typical Male and Female Means

This table lists the mean ratings made for the three target groups by both males and females for the 79 preliminary Femininity items. Items listed in descending order based on the value of eta^2.

	Item	Typical M	Typical F	Typical A
84	Buying eye makeup	1.72	6.09	4.19
112	Shaving your legs	1.45	5.98	4.27
147	Having a shower party	2.63	5.89	4.38
74	Crocheting	1.82	5.07	3.43
46	Crying over a TV show	1.89	4.82	3.98
200	Wearing high-heeled shoes	2.18	5.48	4.00
129	Going shopping for hours	3.20	5.95	4.74
170	Embroidering	2.16	5.07	3.50
137	Using hairspray	2.15	5.14	3.98
148	Planting a flower garden	3.12	5.80	4.67
61	Shopping for children's clothing	3.41	6.18	4.60
113	Buying linens	3.27	6.00	4.67
26	Getting your hair curled	2.64	5.77	4.00
103	Carrying a packet of Kleenex	3.33	6.05	4.83
130	Hugging a friend	3.23	5.86	4.69
64	Planning a menu	3.22	5.64	4.30
195	Buying new dishes	3.51	5.97	5.12
9	Knitting a scarf	2.21	5.05	3.26
167	Mending socks	2.50	4.95	3.67
174	Dusting a table	3.68	6.05	5.29
190	Changing a baby's diaper	3.00	5.66	3.95
96	Sewing	2.74	5.45	4.16
18	Putting flowers in a vase	3.33	5.95	4.72
41	Packing for other family members	2.92	5.24	4.33
13	Setting the table	4.15	6.14	5.56
40	Baking a cake from a mix	3.78	6.07	5.07
85	Dyeing your hair	1.64	4.00	3.12
136	Taking dictation	2.37	4.41	2.73
141	Cleaning house	3.49	5.77	5.04
110	Buying a wedding gift	3.80	5.88	4.57
139	Picking flowers	3.47	5.95	4.98
33	Re-potting a plant	3.90	5.95	4.79
93	Bathing a baby	3.70	5.93	4.27

140	Reading a clothing magazine	3.43	5.79	4.83
58	Babysitting for money	3.70	5.80	4.67
44	Wrapping a present	3.80	5.91	5.14
8	Ironing a shirt	3.30	5.36	4.57
123	Sorting laundry	3.61	5.70	4.95
178	Making dinner for company	3.34	5.43	4.38
150	Reading Playgirl	2.18	4.47	3.69
169	Helping a child get ready for school	4.14	6.11	5.19
99	Shampooing a child's hair	4.29	6.04	4.86
135	Changing sheets	3.97	5.82	5.24
146	Making the bed	4.27	6.00	5.45
124	Washing clothes	3.90	5.80	5.31
6	Re-arranging the furniture	4.20	5.93	4.86
199	Giving a bottle to a baby	4.66	6.42	5.52
11	Doing crafts	4.03	5.84	4.74
72	Planning a party	4.05	5.75	4.77
29	Scrubbing a floor	3.12	4.91	4.05
22	Crying in private	4.02	5.91	4.88
107	Folding clothes	4.45	5.98	5.24
132	Writing letters	4.05	5.66	4.69
182	Cleaning a stove	3.29	5.07	3.88
198	Reading to a child	4.80	6.20	5.38
104	Getting your hair styled	4.12	5.60	4.36
39	Weeding a flower bed	3.95	5.52	4.39
188	Washing the dishes	4.02	5.72	5.05
4	Cleaning the bathtub	3.26	5.11	4.35
163	Getting up with a baby at night	4.40	5.80	4.48
25	Comforting a child	4.90	6.32	5.49
20	Shopping for clothes	4.90	6.16	5.09
205	Defrosting the refrigerator	3.80	5.41	4.60
203	Mopping the floor	3.93	5.40	4.83
3	Buying a gift for your mother	4.48	6.07	5.64
45	Sending an anniversary card	4.93	6.34	5.35
160	Sunbathing	4.12	5.51	5.48
149	Replying to an invitation	4.56	6.00	5.17
5	Typing a letter	4.22	5.63	4.53
197	Reading a gossip column	3.46	5.00	4.81
119	Taking a child to the dentist	4.63	5.73	4.69
158	Going to the PTA	3.78	5.00	3.90

The following items failed to meet the criterion of an eta^2 value exceeding .10:

98	Picking up a child at school	5.29	6.22	5.45
86	Talking to a kid's teacher	3.95	5.06	4.53
194	Weeding a garden	4.41	5.50	4.60
193	Using a blow-dryer	5.29	6.25	5.38
73	Taking a bath	5.61	6.61	6.09
120	Reading the Society page	4.00	5.21	4.79

71 Making coffee	5.05	5.98	5.48

TABLE 36

T-tests on 70 Masculinity variables

Results of t-tests to compare the stereotype ratings for the typical male and the typical female.

	MALE RATERS				FEMALE RATERS			
#	M	F	t	p	M	F	t	p
7	6.05	2.55	7.47	.000	6.32	3.45	7.02	.000
14	6.06	2.66	7.38	.000	6.28	3.31	7.19	.000
15	6.18	2.94	6.50	.000	6.17	3.32	6.94	.000
16	6.25	4.33	3.98	.000	6.34	5.52	2.25	.028
17	6.81	5.17	3.59	.002	6.62	5.76	3.32	.002
23	5.76	2.89	6.48	.000	6.03	3.38	6.75	.000
30	5.94	3.44	5.02	.000	6.65	4.83	4.74	.000
31	6.13	3.44	5.51	.000	6.00	4.00	4.60	.000
34	4.94	2.72	4.84	.000	5.14	3.10	4.42	.000
35	4.70	4.06	1.12	.269	5.24	4.34	1.92	.061
37	6.17	4.50	4.07	.000	6.31	5.45	2.54	.014
38	4.24	2.83	2.46	.019	4.69	2.46	5.32	.000
47	5.69	3.94	3.37	.002	5.97	4.34	4.19	.000
52	6.24	2.94	7.10	.000	6.10	4.29	4.23	.000
53	5.82	3.83	4.33	.000	5.93	4.14	3.95	.000
56	5.41	3.06	4.89	.000	5.52	3.97	3.52	.001
57	6.06	5.17	2.20	.035	6.21	5.34	2.63	.011
65	6.12	3.89	7.21	.000	6.14	4.52	3.92	.000
66	6.52	5.78	2.21	.034	6.55	6.41	0.49	.626
76	6.53	4.35	5.81	.000	6.55	5.79	2.51	.015
78	6.24	4.83	3.53	.001	6.41	5.04	4.48	.000
79	6.76	5.50	4.43	.000	6.41	6.34	0.27	.789
80	6.53	4.61	5.55	.000	6.52	5.45	3.24	.002
82	5.82	4.50	2.98	.005	5.64	5.11	1.21	.230
87	5.29	4.11	2.20	.035	5.66	4.69	2.14	.036
88	5.59	5.11	1.00	.325	5.97	5.21	1.93	.059
92	5.29	3.22	4.60	.000	5.28	4.00	2.77	.008
97	5.82	3.94	3.73	.001	6.55	4.41	5.70	.000
101	5.94	3.35	6.24	.000	6.21	4.03	4.96	.000
102	6.41	4.71	4.26	.000	6.55	5.24	3.27	.002
105	6.12	3.82	5.45	.000	6.59	5.41	4.31	.000
106	6.71	4.65	5.87	.000	6.28	5.66	1.84	.071
109	5.18	2.82	4.74	.000	5.45	3.34	5.26	.000
114	6.18	5.18	1.96	.059	6.52	6.24	1.13	.264
116	5.29	3.53	3.59	.001	5.41	4.21	3.22	.002
117	6.29	4.06	5.00	.000	6.21	4.96	3.35	.001
121	4.12	3.06	1.88	.070	5.03	2.41	5.69	.000
122	5.65	4.76	1.66	.107	6.69	5.00	4.22	.000

#	M	F	t	p	M	F	t	p
125	6.58	4.47	5.64	.000	6.29	5.21	3.06	.003
126	3.12	2.29	2.43	.021	4.71	2.11	7.56	.000
133	5.65	3.65	4.50	.000	5.93	4.07	4.49	.000
138	6.00	2.94	6.36	.000	6.10	3.62	5.79	.000
142	5.76	3.17	5.19	.000	5.83	3.62	5.45	.000
145	6.35	4.53	4.87	.000	5.83	4.55	2.60	.012
151	5.88	3.53	4.29	.000	6.21	3.83	5.84	.000
152	4.94	3.58	2.39	.023	5.52	3.68	4.63	.000
153	5.35	3.35	4.34	.000	5.62	3.52	5.06	.000
156	5.23	4.00	2.41	.022	5.55	4.34	3.02	.004
157	5.29	3.64	3.86	.001	5.07	3.69	3.05	.003
161	5.94	4.06	4.08	.000	6.21	5.68	1.67	.102
162	6.06	4.82	4.62	.000	5.90	5.46	1.19	.240
165	6.12	4.18	4.61	.000	6.31	4.93	3.25	.002
166	5.94	4.29	4.38	.000	6.34	4.55	4.39	.000
171	5.82	4.17	3.81	.001	6.34	5.31	2.81	.007
172	5.71	3.53	4.99	.000	6.21	4.76	3.51	.001
173	5.64	2.94	5.72	.000	5.93	3.52	5.29	.000
175	5.94	2.88	6.58	.000	5.93	3.27	5.88	.000
176	6.47	3.82	5.70	.000	6.45	5.21	2.83	.006
179	6.12	4.00	4.12	.000	6.76	4.66	6.12	.000
183	6.18	4.59	3.55	.001	6.31	5.45	2.63	.011
184	6.65	5.88	2.59	.014	6.83	6.52	1.58	.120
185	6.41	4.29	5.69	.000	6.21	5.97	0.82	.414
189	6.47	3.59	7.30	.000	6.00	4.07	4.17	.000
192	5.53	4.53	2.16	.039	5.89	4.66	2.74	.008
196	6.12	2.76	8.65	.000	6.10	3.86	5.47	.000
202	5.88	4.06	3.74	.001	6.31	4.55	4.84	.000
204	5.76	4.29	3.63	.001	5.79	5.10	1.88	.066
207	5.71	4.41	3.48	.001	5.80	4.58	3.00	.004
208	6.59	4.65	6.28	.000	6.46	5.86	1.72	.092
209	6.18	3.24	7.35	.000	5.97	4.41	3.88	.000

Note:
The degrees of freedom for the male raters are usually 33.
The degrees of freedom for the female raters are usually 56.
Due to missing data, slight variations occurred in
some analyses with regard to the sample size.

TABLE 37

T-tests on 79 Femininity variables

	Male raters				Female raters			
#	M	F	t	p	M	F	t	p
3	5.18	5.61	- .76	.450	4.00	6.48	-5.75	.000
4	4.50	4.67	- .26	.790	2.86	5.45	-6.07	.000
5	4.12	5.17	-1.68	.102	4.55	5.67	-1.79	.007
6	4.41	5.50	-2.26	.030	4.29	6.17	-5.04	.000
8	3.24	5.06	-3.41	.002	3.54	5.68	-5.15	.000
9	1.56	4.89	-6.56	.000	2.44	4.97	-5.03	.000
11	4.41	5.44	-2.09	.045	3.63	6.10	-6.28	.000
13	3.88	5.82	-4.71	.000	4.24	6.31	-5.91	.000
18	3.94	6.11	-4.32	.000	2.93	5.90	-7.91	.000
20	4.94	6.11	-2.69	.011	4.83	6.24	-4.04	.000
22	4.35	5.39	-1.97	.058	3.93	6.34	-6.07	.000
25	5.06	5.56	-0.98	.332	4.93	6.72	-5.80	.000
26	2.24	5.72	-6.85	.000	2.81	5.83	-6.67	.000
29	3.65	4.44	-1.49	.145	2.90	5.28	-6.76	.000
33	4.12	5.50	-3.80	.001	3.79	6.28	-6.67	.000
39	4.18	5.00	-1.93	.062	3.59	5.90	-5.65	.000
40	4.06	5.72	-3.52	.001	3.48	6.34	-8.08	.000
41	3.18	4.39	-3.01	.005	2.81	5.74	-7.63	.000
44	4.12	5.28	-2.02	.052	3.72	6.28	-7.01	.000
45	5.41	5.89	-1.05	.303	4.59	6.66	-5.35	.000
46	2.06	4.50	-5.13	.000	1.81	5.07	-9.96	.000
58	3.59	5.78	-5.30	.000	3.60	5.82	-5.28	.000
61	3.35	5.78	-6.28	.000	3.40	6.34	-8.55	.000
64	4.12	4.94	-2.26	.031	2.90	6.00	-8.33	.000
71	5.29	5.61	- .80	.432	5.00	6.25	-4.33	.000
72	4.24	5.61	-2.53	.016	4.14	5.86	-4.54	.000
73	6.00	6.39	- .91	.368	5.48	6.79	-4.12	.000
74	2.06	4.83	-5.22	.000	1.62	5.10	-9.27	.000
84	1.64	6.27	-13.2	.000	1.67	6.03	-12.32	.000
85	1.35	4.61	-7.29	.000	1.70	3.48	-4.60	.000
86	4.18	4.22	- .09	.928	3.90	5.45	-3.97	.000
93	3.71	5.55	-4.20	.000	3.71	6.18	-5.86	.000
96	2.82	5.78	-5.77	.000	2.63	5.31	-6.13	.000
98	4.94	5.53	-1.28	.209	5.48	6.62	-4.41	.000
99	4.12	5.35	-3.51	.001	4.24	6.45	-6.76	.000
103	3.41	5.71	-4.98	.000	3.25	6.28	-7.51	.000
104	4.47	5.59	-2.06	.048	4.21	5.76	-3.79	.000
107	4.71	5.41	-1.76	.088	4.28	6.38	-6.53	.000
110	4.06	5.47	-3.08	.004	3.69	6.21	-7.14	.000
112	1.00	5.65	-11.08	.000	1.69	6.24	-11.15	.000
113	3.65	5.35	-3.00	.005	3.21	6.45	-10.89	.000
119	4.88	5.24	- .75	.461	4.45	6.00	-4.68	.000

#	M	F	t	p	M	F	t	p
120	3.71	5.35	-3.95	.000	4.14	5.17	-2.26	.027
123	4.47	5.29	-1.63	.114	3.41	5.97	-6.58	.000
124	4.41	5.41	-1.62	.116	3.86	6.03	-5.12	.000
129	3.71	5.82	-4.30	.000	3.10	6.00	-8.39	.000
130	3.41	5.29	-3.83	.001	3.26	6.21	-7.94	.000
132	4.24	5.71	-2.70	.011	4.17	5.62	-3.66	.001
135	4.35	5.18	-1.58	.125	3.97	6.21	-6.21	.000
136	2.53	4.47	-3.99	.000	2.24	4.34	-5.26	.000
137	2.18	5.59	-7.76	.000	2.32	4.86	-5.67	.000
139	3.76	5.59	-3.30	.002	3.38	6.21	-7.36	.000
140	3.36	5.35	-3.95	.000	3.46	6.17	-6.43	.000
141	4.00	5.41	-2.57	.015	3.48	6.03	-6.10	.000
146	4.41	5.53	-2.14	.040	4.24	6.31	-5.69	.000
147	2.52	5.65	-6.32	.000	2.69	6.07	-10.35	.000
148	3.47	5.17	-3.35	.002	2.97	6.14	-10.08	.000
149	4.65	5.35	-1.36	.185	4.62	6.38	-4.74	.000
150	1.53	4.69	-7.47	.000	2.50	4.41	-3.75	.000
158	3.94	4.76	-1.67	.104	3.66	5.10	-3.23	.002
160	4.53	5.41	-1.34	.190	4.17	5.66	-3.33	.002
163	4.47	5.18	-1.74	.091	4.57	6.14	-3.81	.000
167	2.76	4.59	-3.67	.001	2.32	5.17	-7.10	.000
169	4.17	5.52	-2.86	.007	4.31	6.45	-5.77	.000
170	2.35	4.59	-4.34	.000	2.00	5.34	-9.27	.000
174	4.53	5.71	-2.51	.017	3.41	6.24	-7.47	.000
178	3.53	5.35	-3.88	.000	3.14	5.45	-5.80	.000
182	3.71	4.41	-1.11	.275	3.00	5.38	-5.38	.000
188	4.47	5.31	-1.36	.183	3.90	6.00	-5.09	.000
190	2.94	5.35	-5.10	.000	3.15	5.83	-6.23	.000
193	5.29	5.94	-1.20	.237	5.45	6.45	-3.30	.002
194	4.59	5.12	-1.26	.216	4.28	5.72	-3.57	.001
195	3.82	5.24	-2.52	.017	3.30	6.41	-10.05	.000
197	4.00	5.35	-2.41	.022	3.34	4.72	-2.72	.009
198	5.35	5.71	- .90	.372	4.59	6.52	-6.07	.000
199	5.59	6.06	-1.11	.276	4.38	6.68	-5.80	.000
200	2.06	5.59	-6.72	.000	2.22	5.45	-7.61	.000
203	4.76	5.23	-1.02	.316	3.61	5.55	-4.85	.000
205	4.76	5.23	- .86	.398	3.29	5.55	-5.87	.000

Note:
The degrees of freedom for male raters are 33.
The degrees of freedom for the female raters are 56.
Due to missing data, slight variations occurred
with regard to the sample size.

Appendix C

STAT QUESTIONNAIRE RESPONSES

Original Sex Typed Activities Test Form

Your Social Security #

BELOW IS A LIST OF DAY TO DAY ACTIVITIES. PLEASE RATE YOURSELF FOR EACH OF THOSE ACTIVITIES IN TERMS OF THE FOLLOWING RATING SCALE:

1	2	3	4	5	6	7
Very uncomfortable or Very awkward					Very comfortable or Not awkward at all	

1. Playing poker
2. Getting your hair styled
3. Changing a fuse
4. Pruning a tree limb
5. Buying a wedding gift
6. Starting a fire in the fireplace
7. Taking a child to the dentist
8. Washing clothes
9. Computing your income tax
10. Writing letters
11. Reading the sports page
12. Driving a sports car
13. Changing sheets
14. Mowing the lawn
15. Replacing a washer in a leaky faucet
16. Replying to an invitation
17. Building a simple table
18. Going to the PTA
19. Trimming a hedge
20. Riding a motorcycle
21. Sunbathing
22. Buying car insurance
23. Getting up with a baby at night
24. Helping a child get ready for school
25. Opening a tight jar-lid
26. Washing the dishes
27. Watching football on TV
28. Checking the oil in a car
29. Mopping the floor
30. Painting the house
31. Typing a letter
32. Rearranging the furniture

33. Shovelling a sidewalk
34. Doing crafts
35. Driving a boat
36. Climbing a tall ladder
37. Setting the table
38. Playing touch football
39. Shopping for clothes
40. Crying in private
41. Buying a new car
42. Comforting a child
43. Drinking a beer
44. Playing pool
45. Re-potting a plant
46. Going fishing
47. Baking a cake from a mix
48. Wrapping a present
49. Driving a pick-up truck
50. Babysitting for money
51. Watching basketball on TV
52. Jogging
53. Planning a party
54. Reading the business page
55. Bathing a baby
56. Shampooing a child's hair
57. Using a hammer
58. Playing softball

Part II: Background Information:

Your age_____ Your sex_____

Which of the following best describes the community in which you live:

1. Farm or open country
Town or city of:
2. Less than 500 pop.
3. 500-1,999
4. 2,000-9,999
5. 10,000-49,999
6. 50,000-249,999
7. 250,000-499,999
8. 500,000-999,999
9. More than 1 million

Are you a twin? No___ Yes, identical___ Yes, fraternal___

How many brothers and sisters now living do you have?

Number of older brothers____
Number of younger brothers____
Number of older sisters____
Number of younger sisters___

Please answer the following as they apply to the time when you were growing up.

	All of the time	Most of the time	Occas-ionally	Never
My parents lived together.				
My father worked full time				
My mother worked full time				
My mother held a part-time job.				

PLEASE ANSWER THE FOLLOWING AS THEY APPLY TO YOUR OWN SEX (MALE OR FEMALE):

How well do you like being a male or a female:

1	2	3	4	5	6	7
I hate it			Indifferent			I love it

How much of the time do you feel dissatisfied with being male or female?

1	2	3	4	5	6	7
Never						Always

Have you ever wished you could be the opposite sex instead of your own?

1	2	3	4	5	6	7
Never						Always

Compared to other members of your sex would you say you are more or less satisfied about being male or female?

1	2	3	4	5	6	7
Much less satisfied						Much more satisfied

Answer both of the following:

How do you think that you compare to other members of your sex in terms of masculinity?

1	2	3	4	5	6	7
Much less masculine						Much more masculine

How do you think that you compare to other members of your sex in terms of femininity?

1	2	3	4	5	6	7
Much less feminine						Much more feminine

How would you characterize your political views:

1	2	3	4	5
Far Left	Liberal	Middle of the road	Conser-vative	Far Right

Are you a religious person:

1	2	3	4
Not religious	Mildly religious	Moderately religious	Very religious

TABLE 38

Stepwise Regression Analysis: 58 STAT Items

Sex (0,1) predicted from all 58 STAT Items
Summary Table

Item	BETA	R	R^2	R^2 Change
48	-.11835	.61134	.37374	.37374
3	.12744	.73004	.53295	.15921
55	-.14915	.76512	.58451	.05246
5	-.12566	.78427	.61507	.02966
38	.07869	.79864	.63782	.02274
47	-.12886	.81015	.65635	.01853
40	-.09734	.81800	.66912	.01277
4	.07930	.82433	.67951	.01039
54	.05914	.82841	.68626	.00674
44	.06054	.83148	.69136	.00511
17	.04320	.83364	.69495	.00358
50	-.06278	.83562	.69826	.00331
8	-.06899	.83733	.70113	.00287
25	.05766	.83934	.70449	.00336
9	.04558	.84047	.70639	.00190
10	-.03653	.84164	.70835	.00196
35	.04347	.84252	.70984	.00149
45	-.04472	.84309	.71080	.00097
15	.04291	.84362	.71169	.00089
30	-.03327	.84425	.71275	.00106
39	-.02076	.84464	.71341	.00066
6	.02788	.84499	.71402	.00060
20	-.03002	.84532	.71457	.00055
31	-.02644	.84561	.71505	.00049
27	.03196	.84588	.71551	.00046
07	.02804	.84615	.71596	.00045
2	-.01735	.84632	.71626	.00030
49	.02953	.84652	.71659	.00033
24	-.05216	.84668	.71686	.00027
23	.02372	.84690	.71724	.00038
18	.02099	.84706	.71751	.00028
52	.02198	.84722	.71778	.00027
42	.02203	.84734	.71798	.00020
37	-.02875	.84743	.71813	.00015
12	-.01057	.84750	.71825	.00012
26	.01589	.84756	.71835	.00010
51	-.01091	.84760	.71843	.00008
21	-.01367	.84766	.71853	.00010
46	.00955	.84770	.71860	.00007

33	-.01191	.84774	.71867	.00007
36	.00997	.84778	.71874	.00007
28	-.01213	.84782	.71880	.00006
34	.00917	.84785	.71885	.00005
11	-.00921	.84788	.71890	.00005
41	-.01147	.84790	.71894	.00005
22	.01011	.84794	.71901	.00006
56	.00923	.84795	.71903	.00002
13	.00802	.84796	.71904	.00002
14	-.00619	.84797	.71906	.00002
57	.00507	.84798	.71907	.00002
1	.00229	.84798	.71908	.00000
58	.00255	.84799	.71908	.00000
16	-.00208	.84799	.71908	.00000

CONSTANT .84824

Note: The following items were not placed in the equation due to a failure to meet minimal inclusion criteria: Items 19, 29, 32, 43, 53.

TABLE 39

Stepwise Multiple Regression: Masculine items

Sex (0,1) predicted from 31 Masculine Items
Summary Table

Item	BETA	R	R^2	R^2 Change
03	.26597	.55977	.31334	.31334
27	.10541	.59796	.35756	.04422
30	-.15843	.61399	.37698	.01942
17	.09587	.63719	.40601	.02903
15	.15494	.64804	.41995	.01394
44	.09743	.65693	.43156	.01161
14	-.10558	.66350	.44023	.00867
4	.12176	.67044	.44949	.00927
38	.06185	.67308	.45304	.00355
1	.05942	.67507	.45573	.00268
54	-.05640	.67717	.45856	.00283
36	.04823	.67866	.46059	.00203
28	.06841	.68007	.46249	.00190
41	-.05806	.68118	.46401	.00152
35	.06602	.68250	.46622	.00220
46	-.03297	.68358	.46728	.00106
49	-.02622	.68417	.46808	.00081
33	-.04478	.68470	.46881	.00073
25	.03425	.68516	.46944	.00063
57	-.03770	.68569	.47018	.00073
58	.02691	.68606	.47068	.00050
11	.02049	.68623	.47091	.00023
9	.01100	.68632	.47103	.00012
52	.01380	.68639	.47113	.00010
12	-.01003	.68645	.47121	.00008
20	-.00821	.68648	.47126	.00005
19	-.00914	.68651	.47130	.00004
22	.00673	.68653	.47133	.00003
6	-.00315	.68654	.47133	.00001
CONSTANT	-.183557			

Note: The following items were not included in the final regression equation due to a failure to meet minimal inclusion criteria: Items 43,51.

TABLE 40

Stepwise Multiple Regression: 27 F Items

Sex (0,1) predicted from 27 Feminine Items
Summary Table

Item	BETA	R	R^2	R^2 Change
48	-.18687	.61134	.37374	.37374
55	-.20779	.66770	.44582	.07208
40	-.12687	.69989	.48984	.04403
05	-.19686	.71992	.51829	.02844
47	-.18191	.73373	.53835	.02007
18	.04826	.73927	.54651	.00816
10	-.09496	.74464	.55450	.00798
53	.08507	.75052	.56328	.00879
39	-.08669	.75571	.57110	.00781
7	.09482	.75861	.57549	.00439
50	-.10388	.76205	.58071	.00522
34	.05311	.76434	.58422	.00351
16	.05778	.76575	.58638	.00216
24	-.11776	.76736	.58885	.00247
23	.08593	.76931	.59183	.00298
2	-.04676	.77056	.59377	.00194
8	-.06392	.77148	.59518	.00141
45	.04270	.77251	.59678	.00160
29	.03766	.77305	.59761	.00083
21	.02739	.77342	.59818	.00057
37	-.04277	.77370	.59861	.00043
32	.02860	.77397	.59903	.00041
56	.02495	.77404	.59914	.02495
42	.01024	.77408	.59919	.00005
31	-.00901	.77410	.59924	.00004
26	.00899	.77413	.59928	.00005
13	.00863	.77415	.59931	.00003
CONSTANT	1.705379			

Note: All variables were entered into the equation.

TABLE 41

Males: Varimax Rotated Factor Matrix

	MASCULINE COMFORT ITEMS				
	I	II	III	IV	V
1	.26	-.13	.33	.04	.11
3	.73	.14	.01	.07	.03
4	.62	.31	-.09	.06	.08
6	.65	.15	.18	.06	.12
9	.26	.22	.02	.15	.16
11	.04	.12	.72	-.01	-.07
12	.52	-.05	.39	-.05	.15
14	.34	.65	.20	.04	-.04
15	.70	.34	.02	.15	-.03
17	.61	.12	.02	.08	.06
19	.52	.42	.00	.15	.05
20	.51	-.05	.29	.00	.13
22	.39	.12	.09	.21	.08
25	.44	.44	.24	.08	.08
27	.12	.12	.77	-.02	-.12
28	.57	.32	.35	.02	-.10
30	.39	.50	.13	.07	.11
33	.33	.64	.22	-.02	.02
35	.56	-.15	.27	.07	.20
36	.47	.16	.16	-.01	.09
38	.21	.09	.68	.05	.09
41	.33	-.23	.17	.11	.32
43	.25	-.05	.31	-.10	.17
44	.32	-.07	.51	-.08	.24
46	.38	.01	.33	.06	.08
49	.53	.20	.35	.05	.12
51	.00	.13	.67	.00	-.01
52	.15	.20	.46	.03	.23
54	.29	.14	.19	.27	.26
57	.62	.23	.28	.01	-.02
58	.22	.03	.69	.05	.03

FEMININE COMFORT ITEMS

	I	II	III	IV	V
2	.03	.01	.00	.11	.27
5	.10	.19	-.08	.21	.42
7	.25	.24	.02	.47	.07
8	.09	.49	.02	.22	.23
10	-.09	.26	.02	.02	.37
13	.12	.62	.03	.27	.24
16	.21	.30	.06	.18	.41
18	.05	.17	-.12	.45	.23
21	.17	.02	.20	.02	.48
23	.05	.17	.03	.75	.01
24	.08	.14	.01	.78	.11
26	.02	.70	.05	.24	.16
29	.19	.65	.04	.24	.20
31	.04	.34	.10	.12	.42
32	.15	.43	.05	.08	.39
34	.31	.14	-.12	.22	.35
37	.11	.59	.01	.28	.30
39	.01	.06	.12	.08	.56
40	-.03	.11	-.03	.19	.33
42	.10	.08	.00	.57	.25
45	.24	.25	-.09	.32	.42
47	.14	.28	.03	.31	.34
48	.13	.34	-.03	.22	.46
50	-.03	.23	.19	.40	.33
53	.28	-.05	.18	.10	.47
55	.01	.05	.01	.81	.23
56	.04	.08	.01	.80	.23

TABLE 42

Females: Varimax Rotated Factor Matrix

	MASCULINE COMFORT ITEMS				
	I	II	III	IV	V
1	-.16	.31	-.04	.13	.06
3	.01	.67	.03	-.05	-.04
4	.10	.60	.06	.00	-.07
6	.14	.42	.06	.08	.20
9	.11	.29	.07	.06	.08
11	.12	.06	-.02	.63	.05
12	-.07	.22	-.07	.19	.52
14	.54	.27	.05	.14	-.03
15	.05	.66	.05	.09	-.11
17	.02	.55	.09	.11	.03
19	.25	.58	.14	.04	-.04
20	-.11	.30	-.10	.30	.27
22	-.01	.44	.18	-.04	.22
25	.39	.31	.15	.18	-.10
27	.06	.04	-.02	.72	.02
28	.08	.53	-.13	.19	.04
30	.43	.42	.02	.22	.02
33	.56	.33	.03	.21	-.10
35	-.12	.35	.10	.25	.40
36	.15	.21	-.04	.22	.09
38	.04	.13	.02	.66	.17
41	-.22	.31	.09	.09	.52
43	-.03	.00	-.02	.17	.25
44	-.05	.27	.05	.37	.26
46	.14	.29	.15	.35	.03
49	.05	.31	-.03	.13	.36
51	.15	.04	.10	.57	-.04
52	.14	.09	.02	.38	.19
54	.11	.32	.19	.18	.19
57	.31	.40	.13	.26	.06
58	.09	.17	.04	.59	.19

FEMININE COMFORT ITEMS

	I	II	III	IV	V
2	.23	-.06	.13	-.11	.33
5	.22	.05	.15	-.10	.37
7	.31	.09	.39	.02	.14
8	.60	.10	.10	-.02	.03
10	.39	-.07	.11	.15	.16
13	.70	.06	.17	-.02	-.02
16	.34	.11	.18	.01	.25
18	.20	.29	.43	-.04	-.01
21	.21	-.06	-.03	.12	.42
23	.21	.17	.71	.01	.04
24	.26	.08	.73	.00	.02
26	.70	.02	.12	.07	-.10
29	.69	.24	.12	.06	-.09
31	.37	.03	.09	.16	.23
32	.49	.11	.17	.03	.30
34	.30	.22	.18	.04	.18
37	.70	-.02	.17	.04	.09
39	.24	-.25	.10	.11	.52
40	.16	-.01	.14	.06	.18
42	.12	-.05	.61	.07	.21
45	.37	.39	.18	.01	.12
47	.50	.00	.13	.05	.14
48	.57	.01	.09	.10	.24
50	.31	-.14	.26	.19	.24
53	.13	.06	.23	.14	.54
55	.13	.11	.82	.05	.10
56	.20	.10	.80	.04	.08

TABLE 43

58 STAT items: Stereotype Correlation Matrix

This table lists the intercorrelations between the average stereotype ratings of the 58 final STAT items for the three stereotype targest, separately by sex of rater.

Male Raters

	Female Target	Adult Target
Male Target	-.73***	.72***
Female Target		-.34**

Female Raters

	Female Target	Adult Target
Male Target	-.69***	-.41***
Female Target		.82***

*p < .05
**p < .01
***p < .001

TABLE 44

58 STAT items: Male v. Female Stereotypes

This table lists the intercorrelations of the average stereotype ratings for the final 58 STAT items made by males and females for the three targets.

Male Raters	Female Raters		
	Male Target	Female Target	Adult Target
Male Target	.94***	-.64***	-.37**
Female Target	-.77***	.86***	.71***
Adult Target	.68***	-.23*	.11

*p < .05
**p < .01
***p < .001

TABLE 45

Self vs. Stereotype Ratings: 58 Items

This table lists the intercorrelations between the average self-ratings made on each of the 58 final STAT items by males and females and the average stereotype ratings made by males and females for each of the three targets.

Stereotype Ratings	Avg. Male Self-ratings	Avg. Female Self-ratings
Male Raters		
Male Target	.84***	-.47***
Female Target	-.50***	.75***
Adult Target	.88***	.05
Female Raters		
Male Target	.78***	-.50***
Female Target	-.42***	.82***
Adult Target	-.06	.84***

*p < .05
**p < .01
***p < .001

REFERENCES

Bakan, D. The duality of human existence. Chicago: Rand-McNally, 1966.

Baucom, D. H. Independent masculinity and femininity scales on the California Psychological Inventory. Journal of Consulting and Clinical Psychology, 1976, 44, 876.

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 the Pedhazur-Tetenbaum and Locksley-Colten critiques. Journal of Personality and Social Psychology, 1979, 37, 1947-1954.

Bem, S. L., & Lenney, E. Sex typing and the avoidance of cross-sex behavior. Journal of Personality and Social Psychology, 1976, 33, 48-54.

Bem, S. L., Martyna, W., & Watson, C. Sex typing and androgyny: Further explorations of the expressive domain. Journal of Personality and Social Psychology, 1976, 34, 1016-1023.

Berzins, J. I., Welling, M. A., & Wetter, R. E. A new measure of psychological androgyny based on the Personality Research Form. Journal of Consulting and Clinical Psychology, 1978, 46, 126-138.

Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosenkrantz, P., & Vogel, S. R. Sex-role stereotypes and clinical judgments of mental health. Journal of Consulting and Clinical Psychology, 1970, 34, 1-7.

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F.E., & Rosenkrantz, P. S. Sex role stereotypes: A current appraisal. Journal of Social Issues, 1972, 28, 59-78.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation of the multi-trait multi-method matrix. Psychological Bulletin, 1959, 56, 81-105.

Carlson, R. Sex differences in ego functioning. Journal of Consulting and Clinical Psychology, 1971, 37, 267-277.

Comrey, A. Common methodological problems in factor analytic studies. Journal of Consulting and Clinical Psychology, 1978, 46, 648-659.

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

Cosentino, F., & Heilbrun, A. B. Anxiety correlates of sex role identity in college students. Psychological Reports, 1964, 14, 729-730.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Crowne, D. P., & Marlowe, O. The Approval Motive. New York: Wiley, 1964.

Dahlstrom, W. G., & Welsh, G. S. An MMPI handbook. Minneapolis: University of Minnesota Press, 1960.

Edwards, A. L., & Ashworth, C. D. A replication study of item selection for the Bem Sex Role Inventory. Applied Psychological Measurement, 1977, 1, 501-507.

Feirstein, B. Real Men Don't Eat Quiche. New York: Pocket Books, 1982.

Franck, K., & Rosen, E. A projective test of masculinity-femininity. Journal of Consulting Psychology, 1949, 13, 247-256.

Gough, H. G. Identifying psychological femininity. Educational and Psychological Measurement, 1952, 12, 427-439.

Gough, H. G. A cross cultural analysis of the CPI Femininity Scale. Journal of Consulting Psychology, 1966, 30, 136-141.

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

Helmreich, R. L., Spence, J. T., & Holahan, C. K. Psychological androgyny and sex role flexibility: A test of two hypotheses. Journal of Personality and Social Psychology, 1979, 37, 1631-1644.

Helmreich, R. L., Spence, J. T., & Wilhelm, J. A. A psychometric analysis of the Personal Attributes Questionnaire. Sex Roles, 1981, 7, 1097-1108.

Helmreich, R. L., & Stapp, J. Short forms of the Texas Social Behavior Inventory (TSBI), an objective measure of self-esteem. Bulletin of the Psychonomic Society, 1974, 4, 183-190.

Helmreich, R. L., Stapp, J., & Ervin, C. The Texas Social Behavior Inventory: An objective measure of self-esteem or social competence. JSAS Catalog of Selected Documents in Psychology, 1974, 4, 79. (Ms. No. 681)

Hogan, R., DeSoto, C. B., & Solano, C. Traits, tests, and personality research. American Psychologist, 1977, 32, 255-264.

Jackson, D. N. Personality Research Form Manual. Goshen, N.Y.: Research Psychologists Press, 1967.

Kelly, G. A. A Theory of Personality: The Psychology of Personal Constructs. New York: W. W. Norton, 1963.

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

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

Kim, J. O., & Kohout, F. J. Multiple Regression Analysis: Subprogram Regression. In Nie, N. H., Hull, H., Jenkins, J., Steinbrenner, K., & Bent, P. (Eds.), SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill, 1975.

Locksley, A., & Colten, M. E. Psychological androgyny: A case of mistaken identity. Journal of Personality and Social Psychology, 1979, 37, 1017-1031.

Lunneborg, P. W. Stereotypic aspects in masculinity-femininity measurement. Journal of Consulting and Clinical Psychology, 1970, 34, 113-118.

Lunneborg, P. W. Dimensionality of MF. Journal of Clinical Psychology, 1972, 28, 313-317.

Lunneborg, P. W., & Lunneborg, C. E. Factor structure of MF scales and items. Journal of Clinical Psychology, 1970, 26, 360-366.

Maccoby, E. E., & Jacklin, C. N. The Psychology of Sex Differences. Stanford, California: Stanford University Press, 1974.

Markus, H. Self-schemata and processing information about the self. Journal of Personality and Social Psychology, 1977, 35, 63-78.

Mischel, W. On the future of personality measurement. American Psychologist, 1977, 32, 246-254.

Myers, A. M., & Gonda, G. Empirical validation of the Bem Sex Role Inventory. Journal of Personality and Social Psychology, 1982a, 43, 304-318.

Myers, A. M., & Gonda, G. Utility of the masculinity-femininity construct: Comparison of the traditional and androgyny approaches. Journal of Personality and Social Psychology, 1982b, 43, 514-522.

Nie, N., Hull, H., Jenkins, J., Steinbrenner, K., & Bent, P. SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill, 1975.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1978.

Orlofsky, J. L. Relationship between sex role attitudes and personality traits and the Sex Role Behavior Scale-1: A new measure of masculine and feminine role behaviors and interests. Journal of Personality and Social Psychology, 1981, 40, 927-940.

Orlofsky, J. L., Ramsden, M. W., & Cohen, R. S. Development of the revised Sex Role Behavior Scale. Journal of Personality Assessment, in press.

Orlofsky, J. L., Ramsden, M. W., & Cohen, R. S. Appendix: Item analysis and construction of the SRBS-2. Unpublished manuscript, 1981. (Available from J. Orlofsky, Department of Psychology, University of Missouri--St. Louis, St. Louis, Missouri 63121.)

Parsons, T., & Bales, R. F. Family, socialization and interaction process. New York: Free Press of Glencoe, 1955.

Pedhazur, E. J., & Tetenbaum, T. J. Bem Sex Role Inventory: A theoretical and methodological critique. Journal of Personality and Social Psychology, 1979, 37, 996-1017.

Pleck, J. H. Masculinity-femininity: Current and alternative paradigms. Sex Roles, 1975, 1, 161-178.

Rosenkrantz, P., Vogel, S., Bee, H., Broverman, I., & Broverman, D. M. Sex-role stereotypes and self concepts in college students. Journal of Consulting and Clinical Psychology, 1968, 32, 287-295.

Spence, J. T., & Helmreich, R. L. Masculinity and Femininity: Their psychological dimensions, correlates, and antecedents. Austin, Texas: University of Texas Press, 1978.

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, 1032-1046.

Spence, J. T., & Helmreich, R. L. Masculine instrumentality and feminine expressiveness: Their relationships with sex role attitudes and behaviors. Psychology of Women Quarterly, 1980, 5, 147-163.

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

Spence, J. T., Helmreich, R. L., & Sawin, L. L. The Male-Female Relations Questionnaire: A self-report inventory of sex-role behaviors and preferences and its relationships to masculine and feminine personality traits, sex-role attitudes and other measures. Unpublished manuscript, 1981. (Available from J. Spence, Psychology Department, University of Texas at Austin, Austin, Texas 78712)

Spence, J. T., Helmreich, R. L., & Stapp, J. 37, 1779-1789. their relation to self-esteem and conceptions of masculinity and femininity. Journal of Personality and Social Psychology, 1975, 32, 29-39.

Storms, M. D. Sex role identity and its relationships to sex role attributes and sex role stereotypes. Journal of Personality and Social Psychology, 1979, 37, 1779-1789.

Taylor, M. C., & Hall, J. A. Psychological androgyny: Theories, methods, and conclusions. Psychological Bulletin, 1982, 92, 347-366.

Terman, L., & Miles, C. C. Sex and Personality. New York: McGraw-Hill, 1936.

Walkup, H., & Abbott, R. D. Cross-validation of item selection on the Bem Sex Role Inventory. Applied Psychological Measurement, 1978, 2, 63-71.