Journal of Personality and Social Psychology
1991, Vol. 61, No. 6, 1000-1011

0022-3514/91/$3.00

Some Psychometric Characteristics of Gender Diagnosticity Measures: Reliability, Validity, Consistency Across Domains, and Relationship to the Big Five

Richard Lippa
California State University, Fullerton

Preferences for various occupations, school subjects, everyday activities, and hobbies and amusements were rated by 119 male and 145 female Ss. Discriminant analyses were conducted to compute gender diagnostic probabilities. Ss also rated themselves on Big Five traits and completed the Bem Sex-Role Inventory (BSRI) and the Personal Attributes Questionnaire (PAQ). Results indicated that (a) gender diagnosticity measures showed high reliability, (b) gender diagnosticity predicted sex of S and self-ascribed masculinity (M) and femininity (F) better than contrasted-groups M-F scales, (c) gender diagnosticity measures displayed substantial consistency across domains both within and across the sexes, and (d) gender diagnosticity measures were independent of the Big Five and PAQ and BSRI scales both within and across the sexes, whereas PAQ and BSRI scales loaded highly on Big Five dimensions.

Psychologists have systematically studied gender-related individual differences for over half a century. Beginning with Terman and Miles's (1936) classic research on sex and personality, numbers of researchers have developed masculinity-femininity (M-F) scales on the basis of the assumption that M-F is a bipolar unidimensional trait that can be assessed using self-report questionnaires (Campbell, 1966; Gough, 1964; Hathaway & McKinley, 1943; Strong, 1943). In the 1970s researchers developed a new generation of scales that were based on the assumptions that M and F are separate and independent dimensions and that M can be defined in terms of self-ascribed instrumental personality traits and F can be described in terms of self-ascribed expressive traits (Bem, 1974; Cook, 1985; Heilbrun, 1976; Spence, Helmreich, & Stapp, 1974).[1]

Recently, Lippa and Connelly (1990) have argued for a new approach to assessing gender-related individual differences, a method they term *gender diagnosticity*. Gender diagnosticity, which is conceptually related to the diagnostic ratio approach to assessing stereotypes (McCauley & Stitt, 1978; McCauley, Stitt, & Segal, 1980), refers to the Bayesian probability that an individual is predicted to be male or female on the basis of some set of gender-related diagnostic indicators.

Consider the following simple example: Given that an individual wears pants, what is the probability that the individual is male or female? Assuming that 30% of women and 99% of men wear pants in a given population and further assuming for the sake of simplicity that 50% of the population is male and 50% is female, we can use Bayes' Theorem to compute the conditional probability than an individual is male given that he or she wears pants (p = .77) and the complementary probability that the individual is female given that he or she wears pants (p = 1 − .77 = .23).

Lippa and Connelly (1990) proposed that gender diagnostic probabilities can serve as reliable and valid measures of gender-related individual differences both within and across the sexes. Such gender diagnostic probabilities have a number of potential advantages over other methods of assessing gender-related individual differences. They do not reify M and F or freeze gender-related individual differences into specific constructs such as instrumental or expressive traits. Gender diagnostic probabilities can be computed from a wide variety of psychological data (e.g., from measures of occupational preferences, motor skills, cognitive abilities, temperament, items of intelligence tests, and items of personality and attitude scales) and thus can be used to study the degree of linkage between gender-related behaviors in different domains. Gender diagnostic probabilities assess gender-related individual differences in terms of behaviors that actually discriminate between men and women within a given population rather than in terms of gender stereotypes. Finally, gender diagnostic probabilities may be particularly useful in developmental and cross-cultural research, for they tailor the assessment of gender-related individual differences to particular populations and particular samples of behavior.

Lippa and Connelly (1990) demonstrated the reliability and validity of gender diagnostic probabilities in an empirical study of 117 male and 110 female college students. Using discriminant analyses (e.g., see Goldstein & Dillon, 1978; Hand, 1981; Lachenbruch, 1975), they computed gender diagnostic

I would like to thank the following assistants who helped conduct this research: Janet Heck, Stuart Kaplowitz, Joy Kobayashi, Michael Laboissonniere, and Jeanine Renne.

Correspondence concerning this article should be addressed to Richard Lippa, Department of Psychology, California State University, Fullerton, California 92634-9480.

[1] The terms *masculinity* and *femininity*, as operationally defined by instruments such as the PAQ and the BSRI, refer to self-ascribed instrumental and expressive personality traits. As other researchers have noted (e.g., Deaux, 1985; Paulhus, 1987; Spence, 1984), it may be more appropriate to label these constructs as *dominance* and *nurturance*.

probabilities from subjects' occupational preference ratings. More specifically, subjects rated their degree of preference for each of 70 occupations selected from the Strong–Campbell Interest Inventory (Campbell & Hansen, 1981). Seven discriminant analyses were conducted, each on 10 of the rated occupations. Sex of subject served as the grouping variable. Each analysis yielded the Bayesian probability, computed from a subject's discriminant function scores, that the subject was male (or, by subtracting this probability from 1, female).

The mean of the seven gender diagnostic probabilities computed from subjects' occupational preference ratings proved to be reliable both across and within the sexes (coefficient alpha was .87 for all subjects, .65 for men, and .66 for women). Factor analyses conducted for all subjects and for each sex separately indicated that gender diagnostic probabilities were factorially distinct from M, F, and M–F as assessed by the Personal Attributes Questionnaire (PAQ; Spence & Helmreich, 1978; Spence, Helmreich, & Stapp, 1974) and the Bem Sex-Role Inventory (BSRI; Bem, 1974, 1981).

Finally, gender diagnostic probabilities tended to predict a number of gender-related criteria (such as subjects' SAT math scores, mental rotation ability, smiling, and rated M–F of self-descriptive paragraphs, live appearance, photographed appearance, handwriting styles, and college majors) more strongly than did assessed M or F. For example, stepwise regression analyses indicated that over all subjects, gender diagnosticity accounted for 67% of the variance in a composite measure of eight gender-related criterion behaviors, F for only an additional 2%, and M for no additional variance (multiple $r = .83$). For men, only gender diagnosticity accounted for a significant amount of variance (29%) in the gender-related composite ($r = .54$). For women, both gender diagnosticity and F accounted for significant amounts of variance in the gender-related composite criterion (12% and 19%, respectively, multiple $r = .54$). Thus, for all subjects and for men, gender diagnosticity measures showed substantially greater predictive utility than did PAQ or BSRI scales, whereas for women, both gender diagnosticity and F served to significantly predict gender-related behaviors.

Although it provided strong preliminary evidence for the reliability and validity of gender diagnosticity measures, the Lippa and Connelly study left a number of psychometric questions unanswered. One of these was, Can the within-sex reliability of gender diagnosticity measures be increased? Lippa and Connelly (1990) conducted multiple discriminant analyses on subsets of occupational preference ratings to permit computation of the reliability of gender diagnostic probabilities. Whereas the reliabilities obtained indicated that gender diagnostic probabilities showed considerable within-sex consistency, within-sex reliabilities were still somewhat lower than those considered conventionally acceptable for psychological tests. In the current research, we attempted to compute more reliable gender diagnostic probabilities by increasing the number of gender diagnostic indicators used in the computations and by using 5-point rather than 3-point scales of preference.

A second question unanswered by the Lippa and Connelly (1990) study was, What is the relationship between gender diagnosticity measures and M–F scales constructed according to classical contrasted-groups methodology? Clearly, the discriminant analyses conducted in the Lippa and Connelly study computed weighted combinations of items (i.e., preference ratings of various occupations) that best discriminated men from women. These weighted combinations (i.e., the discriminant functions) were then used to compute gender diagnostic probabilities. Traditional M–F scales also select items that best discriminate men from women and then sum subjects' responses over these items to form scales. Does the gender diagnosticity method yield measures that differ from such traditional M–F scales? Do gender diagnosticity measures and contrasted-groups M–F scales display differing degrees of validity?

A third question posed by Lippa and Connelly (1990) was, How much are gender diagnostic probabilities dependent on the domain of gender-related indicators used in computing the probabilities? Lippa and Connelly computed their most reliable measure of gender diagnostic probabilities from subjects' occupational preference ratings. A second, less reliable measure was computed from subjects' ratings of their participation in 22 hobbies. (Lippa and Connelly chose to assess work and leisure activity preferences because these two broad behavioral domains seem to be central to most people's lives and are likely to show average sex differences.) Factor analyses suggested that the two gender diagnosticity measures loaded highly on one factor that was distinct from PAQ and BSRI M and F, and this factor structure was obtained in separate factor analyses of data for men and women, as well as in a factor analysis of data for all subjects combined.

These results provided suggestive but not conclusive evidence that gender diagnosticity measures show consistency across different domains of gender-related behaviors. The current research extends this finding by examining gender diagnostic probabilities computed from subjects' ratings of preference for school subjects, everyday activities, and amusement and hobbies as well as from ratings of preference for occupations.

A final psychometric question unanswered by Lippa and Connelly's data concerns the relationship of gender diagnosticity to other commonly assessed dimensions of personality such as the Big Five (Digman, 1990; John, 1990; John, Angleitner, & Ostendorf, 1988; McCrae & Costa, 1987). Do gender diagnostic probabilities (and other measures of gender-related individual differences such as the M and F scales of the PAQ and BSRI) overlap with the Big Five? Are gender-related individual differences redundant with commonly assessed dimensions of personality such as the Big Five, or are they independent dimensions that account for unique behavioral variance?

The research to be reported here attempts to answer the four questions just posed and thus to clarify some of the psychometric characteristics of gender diagnosticity measures. To enable the computation of more reliable gender diagnostic probabilities, I asked a new sample of subjects to rate their preferences for 131 occupations (rather than the 70 occupations used in Lippa and Connelly's 1990 research). To assess the link between gender diagnosticity measures and traditional M–F scales, I computed gender diagnosticity measures from the preference data, and I also constructed traditional M–F scales according to classical contrasted-groups methodology. I then examined the convergent and discriminant validity of these two kinds of measures.

To assess the consistency of gender diagnosticity measures across domains, in addition to obtaining occupational preference ratings, I also asked subjects to rate their degree of preferences for 35 school subjects, 51 common activities, and 39 hobbies and amusements. This permitted the computation of gender diagnostic probabilities from preference ratings for four different behavioral domains (occupations, school subjects, everyday activities, and amusements and hobbies).

Subjects also completed the PAQ and the BSRI and rated themselves on 30 traits chosen as markers of the Big Five. Thus, the current research could assess the relationships among gender diagnosticity measures, M, F, and the Big Five.

Finally, in completing the BSRI, subjects rated themselves on the items "masculine" and "feminine." This allowed us to examine how well gender diagnosticity and contrasted-groups M–F scales predict subjects' self-ascribed M–F.

Method

Subjects

Subjects were 264 (119 male and 145 female) undergraduate introductory psychology students at California State University, Fullerton.

Materials

Subjects, who were solicited from four different introductory psychology classes, were administered in class a questionnaire packet that included the PAQ (as presented in Spence & Helmreich, 1978) and the BSRI (Bem, 1974, 1981). The packet also contained a section that asked subjects to rate their degree of preference for 131 occupations (e.g., art teacher, building contractor, and dietitian), 35 school subjects (e.g., algebra, art, and English composition), 51 activities (e.g., making a speech, cooking, and discussing politics), and 39 amusements and hobbies (e.g., fishing, jazz or rock concerts, and art galleries). Subjects rated their degree of preference on a 5-point scale: *strongly dislike,* (1) *slightly dislike* (2), *neutral or indifferent* (3), *slightly like* (4), and *strongly like* (5). The lists of occupations, school subjects, activities, and amusements were those appearing in Part I of the Strong–Campbell Interest Inventory, Form T325 (Campbell & Hansen, 1981).

Finally, subjects were asked to rate themselves on 30 personality traits chosen to tap the Big Five dimensions of personality. These traits were selected from adjectives presented in McCrae and Costa (1987, Table 3). The six traits tapping neuroticism were "calm," "worrying," "at ease," "nervous," "relaxed," and "high-strung." Those tapping extraversion were "retiring," "sociable," "sober," "fun loving," "quiet," and "talkative." Those tapping openness were "conventional," "original," "down to earth," "imaginative," "uncreative," and "creative." Those tapping agreeableness were "ruthless," "soft-hearted," "suspicious," "trusting," "vengeful," and "forgiving." Finally, those tapping conscientiousness were "careless," "careful," "undependable," "reliable," "negligent," and "conscientious." Subjects rated themselves on these traits using a 7-point scale ranging from *never or almost never true* (1) to *always or almost always true* (7) of oneself.

The packet's cover sheet asked subjects to report demographic information including their age, sex, whether they were born in the United States, and their ethnic group (Hispanic, Asian, Black, White, American Indian, or other). Subjects were also asked to list their major. If they had not yet declared a major, subjects were asked to list the two majors they thought they were most interested in pursuing. On the page following the cover sheet, subjects were asked to write a self-descriptive "personal narrative." The instructions printed on the top of this page were the following: "In the space provided below, in your own handwriting, please describe yourself, in two paragraphs or more, in terms of your personality and interests." The data collected about college majors and the personal narrative were used in a validity study, and most of the analyses conducted on these data are not included in the current report.

Results

Computing Gender Diagnosticity Measures

Gender diagnostic probabilities were computed from subjects' preference ratings for occupations, school subjects, activities, and amusements. To compute these probabilities, discriminant analyses were conducted using the discriminant analysis procedure of SPSS/PC + (Norusis, 1986, 1988) using default options. The grouping variable in all analyses was sex of subject.

As noted earlier, the computation of multiple diagnostic probabilities from subsets of items permitted the assessment of their reliability. To compute gender diagnostic probabilities from ratings of occupations, I conducted 13 discriminant analyses, 12 on sets of 10 occupations and 1 on a set of 11 occupations. Thus, the 13 discriminant analyses included all 131 rated occupations.

Similarly, to compute gender diagnostic probabilities from ratings of school subjects, I conducted six discriminant analyses, each on a distinct set of six school subjects. To compute gender diagnostic probabilities from ratings of activities, I conducted seven discriminant analyses, five analyses on sets of seven activities and two analyses on sets of eight activities. Finally, to compute gender diagnostic probabilities from ratings of amusements, I conducted six discriminant analyses, three analyses on sets of six amusements and three analyses on sets of seven amusements.

Each discriminant analysis yielded the Bayesian probability, computed from subjects' discriminant function scores, that a given subject was male (or, by subtracting this probability from 1, female). Thus, on the basis of occupational preference ratings, I assigned to each subject 13 gender diagnostic probabilities, each computed from a distinct set of rated occupations. Similarly, on the basis of ratings of school subjects, I assigned to each subject six gender diagnostic probabilities. On the basis of ratings of activities, I assigned to each subject seven gender diagnostic probabilities, and on the basis of ratings of amusements, I assigned to each subject six gender diagnostic probabilities.

Computing M–F Scales Using Contrasted-Groups Methodology

As a first step in constructing traditional M–F scales, sex of subject was correlated with subjects' preference ratings for occupations, school subjects, activities, and amusements. Preference ratings for 59 of 131 occupations, 8 of 36 school subjects, 24 of 51 activities, and 20 of 39 amusements showed significant correlations with sex (i.e., sex differences). These items were combined into four scales, one for each domain of preference ratings. All items were keyed so that higher scale values represented preferences more typical of men and lower scale values represented preferences more typical of women.

Reliabilities of Gender Diagnostic Probabilities and Contrasted-Groups M–F Scales

An overall gender diagnostic probability that was based on occupational preference ratings was computed by averaging for each subject the 13 gender diagnostic probabilities computed from occupational preference ratings. Similarly, three overall gender diagnostic probabilities were computed using preference ratings of school subjects, activities, and amusements. Finally, a grand mean gender diagnostic probability was computed by computing the mean of the four overall gender diagnostic probabilities computed from ratings of occupations, school subjects, activities, and amusements.

The reliabilities of the four gender diagnostic probabilities computed from different domains (occupations, school subjects, activities, and amusements and hobbies) and the reliabilities of the grand mean are presented in Table 1. These reliabilities were computed separately for all subjects, for men, and for women. As the data in Table 1 indicate, the reliabilities computed over all subjects were quite high for all measures except gender diagnosticity that was computed from ratings of school subjects.

Particularly noteworthy was the finding that gender diagnostic probabilities that were based on occupational preference ratings were highly reliable within the sexes ($\alpha = .76$ for men and .78 for women) as well as over all subjects ($\alpha = .92$). The grand mean gender diagnostic probabilities also displayed substantial reliability within the sexes ($\alpha = .81$ for men and .80 for women) as well as over all subjects ($\alpha = .91$). The current data thus demonstrate more strongly than Lippa and Connelly's (1990) data that gender diagnostic probabilities can be computed at acceptably high levels of reliability both within and across the sexes.[2]

Table 1 also presents the reliabilities of the four contrasted-groups M–F scales and of the grand mean of these four scales. In general, the reliabilities of these scales were comparable to those obtained for gender diagnosticity measures. The one exception was the scale constructed from preference ratings for school subjects, which showed markedly lower reliabilities than the corresponding gender diagnosticity measure. The low reliability of this scale may have been due to the small number of items in the scale.

Table 1
Reliabilities (Coefficient Alpha) of Gender Diagnosticity (GD) Measures and Contrasted-Groups Scales (CG)

Behavioral domain	All subjects (N = 264)		Men (n = 119)		Women (n = 145)	
	GD	CG	GD	CG	GD	CG
Occupations	.92	.90	.76	.78	.78	.83
Subjects	.61	.37	.53	.08	.42	.13
Activities	.80	.72	.56	.50	.58	.51
Amusements	.82	.80	.49	.55	.68	.75
Mean	.91	.89	.81	.82	.80	.80

Relation Between Gender Diagnosticity Measures and Contrasted-Groups M–F Scales

The mean correlation between gender diagnosticity measures and contrasted-groups scales was .66 for all subjects, .40 for men, and .47 for women. Not surprisingly, both gender diagnosticity measures and contrasted-groups scales correlated significantly with sex of subject: Gender diagnosticity measures based on occupations, school subjects, everyday activities, and hobbies and amusements correlated .82, .58, .75, and .74, respectively, with sex (all $ps < .001$), whereas corresponding contrasted-groups scales correlated .60, .51, .61, and .60 (all $ps < .001$).

As suggested by the correlations just presented, sex of subject correlated with gender diagnosticity measures more strongly than with contrasted-groups M–F scales. These differences were statistically significant for measures computed for occupations, activities, and amusements ($p < .001$) and marginally significant for measures computed from school subjects ($p < .1$). (These significance levels were based on t tests for differences in correlations when two variables are correlated with the same third variable; see McNemar, 1962, p. 140).

Regression analyses were conducted to investigate further the utility of gender diagnosticity measures and contrasted-groups scales in predicting sex of subject. For each domain of measures (occupations, school subjects, everyday activities, and hobbies and amusements), I conducted regressions using gender diagnosticity measures and corresponding contrasted-groups scales to predict sex of subject. In four analyses, beta weights for gender diagnosticity measures were large and significant (βs = .86, .46, .74, and .78 for gender diagnosticity measures based on occupations, school subjects, everyday activities, and hobbies and amusements, respectively), whereas beta weights for contrasted groups scales were small and nonsignificant (βs = .06, .14, .00, and .05 for contrasted-groups scales based on occupations, school subjects, everyday activities, and hobbies and amusements, respectively). Thus, despite some degree of overlap between gender diagnosticity measures and contrasted-groups scales, gender diagnosticity measures proved to predict sex of subject better than contrasted-groups scales did.

Factor analyses (principal components, retaining factors with

[2] In the current research, mean discriminant functions scores were strongly correlated with corresponding mean diagnostic probabilities that were computed from various domains of preference ratings (correlations were in the range of .98–.99). The reliabilities for mean discriminant function scores (computed for discriminant functions all keyed in the same direction—e.g., men's mean positive and women's mean negative) were comparable with those presented in Table 1 for diagnostic probabilities. Diagnostic probabilities showed slightly higher within-sexes reliabilities than mean discriminant function scores for two domains of preference ratings: activities, and amusements and hobbies. We repeat here the recommendation made by Lippa and Connelly (1990) that researchers examine the predictive utility of both gender diagnostic probabilities and discriminant function scores.

eigenvalues greater than 1, varimax rotation) of the four gender diagnosticity measures and four contrasted-groups scales were conducted for all subjects, for men, and for women. The factor analysis conducted for all subjects yielded a one-factor solution that accounted for 73% of the variance. This one-factor solution reflected the fact that both gender diagnosticity measures and contrasted-groups scales correlated strongly with sex of subject over all subjects.

However, factor analyses conducted for male and female subjects only yielded two-factor solutions (the solution for men accounted for 67% of total variance, and the solution for women accounted for 69% of variance). The varimax rotated factor matrices for these analyses are presented in Table 2. For men, the two obtained factors were clearly marked by the two classes of measures. Gender diagnosticity measures based on occupations, school subjects, everyday activities, and hobbies and amusements defined one factor, whereas contrasted-groups scales based on occupations, school subjects, everyday activities, and hobbies and amusements defined the other factor. Thus for men, gender diagnosticity measures and contrasted-groups scales seemed not to measure the same construct, despite the fact that they were computed from the same data sets.

The factor analysis for female subjects yielded different results. One factor was marked most strongly by the gender diagnosticity measure and contrasted-groups scale based on hobbies and amusements, whereas the second factor was marked most strongly by the gender diagnosticity measure and contrasted-groups scale based on everyday activities. Thus for women, patterns of correlations among measures was determined more by domain of preference rating than by class of measure.

Gender Diagnosticity Measures and Contrasted-Groups Scales as Predictors of Self-Ascribed M–F

As noted in the Method section, in completing the long form of the BSRI, subjects rated on a 7-point scale how "masculine" and how "feminine" they were. These items are of particular interest because they represent the most direct assessment of subjects' self-ascribed masculinity and femininity. The two-item scale defined by the BSRI items "masculine" minus "feminine" was correlated with gender diagnosticity measures and with contrasted-groups scales for all subjects, for men, and for women.[3]

Over all subjects, self-ascribed M–F correlated more strongly with gender diagnosticity measures (rs = .82, .59, .75, and .74 for measures based on occupations, school subjects, everyday activities, and hobbies and amusements, all significant at $p < .001$) than with contrasted-groups scales (corresponding rs = .63, .51, .59, and .59, all significant at $p < .001$). The differences between corresponding correlations were significant for all four domains, respective $ps < .001$, .02, .001, and .001.

Similarly for men, self-ascribed M–F correlated more strongly with gender diagnosticity measures (rs = .43, .25, .25, and .21, for measures based on occupations, subjects, everyday activities, and hobbies and amusements, all significant at $p < .05$) than with contrasted-groups scales (corresponding rs = .20, .12, .04, and .10, with only the first significant at $p < .05$). The differences between corresponding correlations were significant for measures based on occupations ($p < .001$) and everyday activities ($p < .02$), marginally significant for measures based on school subjects ($p < .1$), and nonsignificant for measures based on hobbies and amusements.

For women, self-ascribed M–F correlated significantly with both gender diagnosticity measures (rs = .34, .23, .31, and .34, all significant at $p < .01$) and with contrasted-groups scales (rs = .34, .27, .27, and .24, all significant at $p < .01$). For women, none of the differences between corresponding correlations were significant.

Intercorrelations Among Gender Diagnosticity Measures and M and F as Assessed by the PAQ and BSRI

Tables 3, 4, and 5 present the intercorrelations of gender diagnosticity measures, PAQ scales, and BSRI scales computed for all subjects (Table 3), for men (Table 4), and for women (Table 5). (Two male and 2 female subjects were not included in these analyses because of missing responses in their PAQ or BSRI questionnaires.)

These intercorrelations indicate substantial consistency among gender diagnostic probabilities computed from different domains of preference ratings. The mean intercorrelation of the four gender diagnosticity measures was .75 for correlations computed for all subjects and .52 for correlations computed both for men and for women separately.

The correlations presented in Tables 3, 4, and 5 also indicate that gender diagnostic probabilities were not strongly correlated with PAQ or BSRI scales. Over all subjects, correlations between gender diagnostic probabilities and PAQ and BSRI scales ranged from .19 to .39 for all subjects, from .03 to .31 for men, and from .03 to .31 for women. In contrast, the M scales of the PAQ and BSRI were substantially correlated with one another (.78 for all subjects, .80 for men, and .74 for women), as were the F scales of the PAQ and BSRI (.71 for all subjects, .68 for men, and .70 for women).

Table 2
Factor Analyses of Four Gender Diagnosticity (GD) Measures and Four Contrasted-Groups Masculinity–Femininity Scales (CG) for Men and Women (Varimax Orthogonally Rotated Factor Matrices)

	Men		Women	
Measure	Factor 1	Factor 2	Factor 1	Factor 2
GD occupation	−.17	.86	−.38	.74
GD school subject	−.41	.66	−.47	.43
GD everyday activities	−.17	.86	−.16	.91
GD hobbies and amusements	−.40	.64	−.83	.27
CG occupation	.83	−.26	.68	−.49
CG school subject	.74	−.24	.69	−.32
CG everyday activities	.72	−.27	.23	−.82
CG hobbies and amusements	.80	−.19	−.89	−.08

[3] "Feminine" was substracted from "masculine" because previous research has repeatedly shown that lay conceptions hold M and F to be negatively correlated (e.g., see Deaux, 1987). Indeed, in our data the correlations between subjects' self-ratings on the items "masculine" and "feminine" were −.84 over all subjects, −.48 for men, and −.41 for women (all significant at $p < .001$).

Table 3
Intercorrelations of Four Gender Diagnosticity (GD) Measures, Personal Attributes Questionnaire (PAQ) Scales, and Bem Sex-Role Inventory (BSRI) Scales (All Subjects, N = 260)

Measure	1	2	3	4	5	6	7	8	9
1. Sex	—								
2. GD occupation	−.82*	—							
3. GD school subject	−.57*	**.72***	—						
4. GD everyday activities	−.75*	**.85***	**.67***	—					
5. GD hobbies and amusements	−.74*	**.80***	**.68***	**.78***	—				
6. PAQ masculinity	−.25*	.31*	.28*	.30*	.19*	—			
7. PAQ femininity	.21*	−.26*	−.18*	−.31*	−.26*	.06	—		
8. PAQ masculinity–femininity	−.21*	.32*	.19*	.32*	.21*	.57*	−.26*	—	
9. BSRI masculinity	−.24*	.35*	.31*	.36*	.20*	**.78***	−.01	.54*	—
10. BSRI femininity	.32*	−.35*	−.24*	−.38*	−.34*	−.13*	**.71***	−.35*	−.04

Note. Correlations in boldface are between two different measures of the same construct.
* $p < .05$.

Factor Structure of Gender-Related Individual Difference Measures

Factor analyses (principal-components analysis, extraction of factors with eigenvalues greater than one, and orthogonal varimax rotation) were conducted on the following individual difference measures: gender diagnostic probabilities based on occupations, school subjects, everyday activities, and amusements and hobbies, PAQ M, PAQ F, PAQ M–F, BSRI M, and BSRI F. Three factor analyses were conducted, one for all subjects, one for men, and one for women. Rotated factor matrices from these three analyses are presented in Table 6.

As the data in Table 6 indicate, all three factor analyses showed a consistent three-factor solution. The four gender diagnosticity measures loaded highly on one factor labeled *Gender Diagnosticity,* the PAQ and BSRI masculinity scales loaded on another factor labeled *Masculinity,* and the PAQ and BSRI femininity scales loaded on a third factor labeled *Femininity.* The PAQ M–F scale was a hybrid scale that loaded on Masculinity and Femininity but not on Gender Diagnosticity. These findings replicate those reported by Lippa and Connelly (1990) and provide additional evidence that gender diagnosticity measures computed from different domains of behavior load on a unitary factor both across and within the sexes.

Gender Diagnosticity, M, F, and Self-Ascribed M–F

I noted earlier that gender diagnosticity measures generally predicted self-ascribed M–F better than contrasted-groups scales, particularly for all subjects and for men. To investigate how gender diagnosticity measures compared with M and F in predicting self-ascribed M–F, I conducted regression analyses using grand mean gender diagnosticity, PAQ M, and PAQ F to predict self-ascribed M–F. (PAQ M and F were used as predictors rather than BSRI M and F because the items "masculine" and "feminine" that comprised self-ascribed M–F came from the BSRI, and thus these items and the BSRI scales might share method variance that would inflate their intercorrelation.) The regression analyses were conducted for all subjects, for men, and for women.

The regression for all subjects showed gender diagnosticity to be the only significant predictor of self-ascribed M–F (βs = .79, .02, and −.05 for gender diagnosticity, PAQ M, and PAQ F multiple $r = .82$, $p < .001$). The regression for men showed that gender diagnosticity and PAQ M but not PAQ F significantly predicted self-ascribed M and F (βs = .30, .22, and −.09 for gender diagnosticity, PAQ M, and PAQ F, multiple $r = .42$, $p < .001$). Finally, the regression for women showed that gender diagnosticity and PAQ F but not PAQ M significantly predicted

Table 4
Intercorrelations of Four Gender Diagnosticity (GD) Measures, Personal Attributes Questionnaire (PAQ) Scales, and Bem Sex-Role Inventory (BSRI) Scales (Men, n = 117)

Measure	1	2	3	4	5	6	7	8
1. GD occupation	—							
2. GD school subject	**.56***	—						
3. GD everyday activities	**.60***	**.48***	—					
4. GD hobbies and amusements	**.48***	**.47***	**.54***	—				
5. PAQ masculinity	.20*	.25*	.21*	.08	—			
6. PAQ femininity	−.07	−.02	−.17	−.08	.10	—		
7. PAQ masculinity–femininity	.24*	.13	.20*	.13	.55*	−.23*	—	
8. BSRI masculinity	.31*	.30*	.26*	.09	**.80***	−.01	.54*	—
9. BSRI femininity	−.14	−.03	−.18	−.16	−.02	**.68***	−.32*	−.04

Note. Correlations in boldface are between two different measures of the same construct.
* $p < .05$.

Table 5
Intercorrelations of Four Gender Diagnosticity (GD) Measures, Personal Attributes Questionnaire (PAQ) Scales, and Bem Sex-Role Inventory (BSRI) Scales (Women, n = 143)

Measure	1	2	3	4	5	6	7	8
1. GD occupation	—							
2. GD school subject	**.53***	—						
3. GD everyday activities	**.69***	**.43***	—					
4. GD hobbies and amusements	**.52***	**.45***	**.49***	—				
5. PAQ masculinity	.19*	.11	.16	−.03	—			
6. PAQ femininity	−.25*	−.16*	−.30*	−.23*	.12	—		
7. PAQ masculinity–femininity	.28*	.05	.31*	.05	.54*	−.22*	—	
8. BSRI masculinity	.25*	.16	.28*	−.02	**.74***	.05	.52*	—
9. BSRI femininity	−.18*	−.10	−.24*	−.14	−.08	**.70***	−.29*	−.05

Note. Correlations in boldface are between two different measures of the same construct.
* $p < .05$.

self-ascribed M–F (βs = .33, −.11, and −.24 for gender diagnosticity, PAQ M, and PAQ F, multiple $r = .48$, $p < .001$). Thus, gender diagnosticity was the only variable to significantly contribute to the prediction of self-ascribed M–F for all subjects, for men, and for women. Furthermore, beta weights in each of the three regression analyses suggested that gender diagnosticity accounted for more unique variance in self-ascribed M–F than did either assessed M or F.

Gender-Related Individual Differences and the Big Five

As noted earlier, subjects rated themselves on 30 traits chosen to tap the Big Five: neuroticism, extraversion, openness, agreeableness, and conscientiousness. Six traits were used to assess each of the Big Five dimensions.

A preliminary analysis was conducted to determine the reliability of these five scales. Alphas were, respectively, .68, .58, .56, .67, and .73. To increase the reliability of these scales, one or two items with the lowest item-total correlations were dropped from each scale. "High-strung" was dropped from the neuroticism scale, increasing its reliability to .70. "Sober" was dropped from the extraversion scale, increasing its reliability to .67. "Conventional" and "down to earth" were dropped from the openness scale, increasing its reliability to .82. "Suspicious" was dropped from the agreeableness scale, with alpha remaining at .67. And "conscientious" was dropped from the conscientiousness scale, increasing its reliability to .76.

A factor analysis of the 24 items retained in the scales (principal-components analysis, five-factor solution, orthogonal varimax rotation) showed the expected five-factor structure. That is, items from a given scale loaded highly on a single factor and tended not to load on the other four factors. The five factors extracted accounted for 56% of the variance in the 24 items.

The main focus of the current research was to assess the relationship of the Big Five to measures of gender-related individual differences. Table 7 presents correlations between measures of gender-related individual differences (the five gender diagnosticity measures and the PAQ and BSRI scales) and the Big Five. These correlations were computed for all subjects, for men, and for women. In general, these correlations show that PAQ and BSRI scales were strongly correlated with Big Five factors, whereas gender diagnostic probabilities were not. PAQ and BSRI F showed relatively large correlations with agreeableness, and they showed moderate correlations with conscientiousness, particularly for women. PAQ and BSRI M showed moderate positive correlations with openness and extraversion and moderate negative correlations with neuroticism.

To investigate further the dimensionality of individual difference measures, I conducted factor analyses on the following variables: gender diagnostic probabilities based on occupations, school subjects, everyday activities, and hobbies and amusements, PAQ scales, BSRI scales, and the scales assessing each of the Big Five. Three factor analyses (principal-components analysis, extraction of factors with eigenvalues greater than 1, and orthogonal varimax rotation) were conducted, one for all subjects, one for men, and one for women. The rotated factor matrices from each of these analyses are presented in Table 8.

All three factor analyses indicated that gender diagnosticity measures were factorially distinct from the Big Five, whereas M and F as assessed by the PAQ and BSRI showed considerable overlap with Big Five dimensions. Consistent with the correlational data, these factor analyses indicated that neuroticism, extraversion, and openness were related to M, with M being negatively related to neuroticism and positively related to extraversion and openness. F was related primarily to agreeableness and conscientiousness. For all subjects and for women, three-factor solutions were obtained (comprising gender diagnosticity, M, and F factors). For men, a four-factor solution was obtained—the M factor obtained in the other two factor analyses split into two factors, one tapping extraversion and openness, the other tapping nonneuroticism and conscientiousness.

Because gender diagnosticity measures proved to be independent of the Big Five in our factor analyses and because of our a priori knowledge that the Big Five represent distinct factors of personality, factor analyses were also conducted that forced a six-factor solution (again using principal-components factor analysis, orthogonal varimax rotation). The results of these analyses (for all subjects, for men, and for women) generally showed the expected six factor solutions: Each of the Big Five defined one factor, and gender diagnosticity measures defined the sixth factor. Consistent with the previous analyses, PAQ and BSRI

Table 6
Factor Analyses of Four Gender Diagnosticity (GD) Measures, Personal Attributes Questionnaire (PAQ) Scales, and Bem Sex-Role Inventory (BSRI) Scales (Varimax Orthogonally Rotated Factor Matrices)

	All subjects (N = 260)			Men (n = 117)			Women (n = 143)		
Measure	Masculinity	Femininity	Gender diagnosticity	Masculinity	Femininity	Gender diagnosticity	Masculinity	Femininity	Gender diagnosticity
GD occupation	.21	−.17	**.90**	.18	−.06	**.81**	.22	−.13	**.83**
GD school subject	.15	−.03	**.84**	.18	.11	**.77**	.04	.02	**.77**
GD everyday activities	.21	−.22	**.87**	.14	−.16	**.80**	.24	−.23	**.76**
GD hobbies and amusements	.04	−.17	**.89**	−.05	−.10	**.79**	−.13	−.10	**.79**
PAQ masculinity	**.90**	.06	.17	**.91**	.09	.11	**.90**	.07	.04
PAQ femininity	.03	**.90**	−.14	.03	**.90**	−.04	.06	**.90**	−.20
PAQ masculinity–femininity	**.76**	**−.38**	.08	**.75**	**−.37**	.08	**.76**	**−.35**	.07
BSRI masculinity	**.88**	.09	.22	**.89**	.10	.19	**.88**	.13	.14
BSRI femininity	−.10	**.88**	−.22	−.05	**.90**	−.10	−.07	**.91**	−.07

Note. Factor loadings in boldface are greater than .35.

scales loaded highly on a number of Big Five factors, but not on the gender diagnosticity factor.

Discussion

My data provide clear answers to the four questions posed at the start of this paper. First, gender diagnostic probabilities can be assessed with conventionally acceptable levels of reliability both within and across the sexes. Second, gender diagnosticity measures are not identical to scales constructed using contrasted-groups methodology, and the differences between the two kinds of measures seem particularly pronounced for men. Third, gender diagnostic probabilities computed from different interest domains show substantial consistency across domains both within and across the sexes. Finally, gender diagnostic probabilities prove to be independent of the Big Five, whereas M and F as assessed by the PAQ and BSRI show strong overlap with Big Five dimensions.

Although the primary focus of the current research was methodological, the results reported here may in addition hold some broader implications for research on gender and gender-related individual differences. The following sections address some of these broader implications.

Reliability

The current research shows that it is possible to compute gender diagnostic probabilities with a level of reliability comparable to that displayed by most personality scales. It is intriguing that gender diagnosticity—an individual difference measure defined by probabilistically predicting membership in two social groups—also displays substantial within-sexes reliability. Lippa and Connelly (1990) speculated that within the sexes, gender diagnosticity may assess how conventionally sex-typed individuals are.

Lippa and Connelly (1990) noted that gender diagnosticity is not a test per se but rather a method for assessing how "male-like" or "female-like" an individual's set of diagnostic indicators is in the context of a particular population of men and women. That is, gender diagnosticity is not linked to a normative sample in the way that most personality tests are and is thus computed anew for each sample of subjects and behavioral indicators.[4] Because gender diagnosticity represents a novel ap-

[4] An anonymous reviewer of this article noted that although the gender diagnosticity approach "does provide flexibility, it also limits—actually eliminates—generalizability. Each gender diagnosticity measure is ad hoc, applicable only to the sample on which it is generated." This criticism is valid, but it is important to note that it applies to all personality measures to some extent. The fact that a questionnaire scale maintains constancy in its wording and physical form does not guarantee that the items are interpreted or responded to equivalently in different populations or in different historical eras. Many famous personality, vocational interest, and intelligence tests undergo radical, albeit infrequent, alterations when they are renormed and "brought up to date." The method of gender diagnosticity opts for a more gradualist approach whereby assessments of gender-related individual differences are fine-tuned to a particular population. The assessment of gender diagnosticity may change over time if different items are weighted differently over time by the discriminant analyses. It is important to note, however, that the item domain itself may remain constant over assessments. Furthermore, it is possible to assess

Table 7
Correlations of Gender Diagnosticity (GD) Measures, Personal Attributes Questionnaire (PAQ) Scales, and Bem Sex-Role Inventory (BSRI) Scales With Big Five Scales for All Subjects, for Men, and for Women

Measure	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
GD occupation					
All subjects	−.22***	−.11	.20**	−.24***	−.08
Men	−.28**	.14	.10	−.04	.23*
Women	−.22**	−.01	.07	−.13	−.03
GD school subject					
All subjects	−.16*	−.06	.19**	−.14*	.04
Men	−.22*	.14	.12	.08	.24*
Women	−.04	−.02	.09	−.10	.10
GD everyday activities					
All subjects	−.21**	−.06	.15*	−.27***	−.11
Men	−.27**	.15	.07	−.15	.06
Women	−.15	.08	−.01	−.17*	−.02
GD hobbies and amusements					
All subjects	−.13*	−.16**	.08	−.18**	−.12
Men	−.20*	.02	−.10	−.06	.03
Women	.02	−.12	−.07	.01	−.04
PAQ masculinity					
All subjects	−.39***	.37***	.46***	−.10	.21**
Men	−.38***	.39***	.45***	−.02	.28**
Women	−.38***	.47***	.43***	−.07	.25**
PAQ femininity					
All subjects	.02	.24***	.17**	.57***	.33***
Men	−.03	.17	.22*	.59***	.21*
Women	.02	.25**	.22*	.51***	.41***
PAQ masculinity–femininity					
All subjects	−.37***	.27***	.23***	−.23***	.05
Men	−.30**	.23*	.06	−.16	.14
Women	−.41***	.41***	.32***	−.23**	.04
BSRI masculinity					
All subjects	−.36***	.40***	.52***	−.11	.25***
Men	−.26**	.43***	.49***	−.07	.30**
Women	−.40***	.51***	.50***	−.05	.30***
BSRI femininity					
All subjects	.04	.17**	.17**	.59***	.38***
Men	.02	.19*	.31**	.59***	.22*
Women	−.01	.06	.20*	.55***	.47***

* $p < .05$, two-tailed. ** $p < .01$, two-tailed. *** $p < .001$, two-tailed.

the utility of a weighting scheme developed in one population in a second population. Indeed, such research can provide useful information about the cross-population and cross-cultural consistency or variability of gender diagnosticity measures (and more generally of gender-related individual differences). The approach typified by the traditional standardized test has the advantage that it facilitates cumulative research and provides a "steady target" for researchers. The approach typified by the method of gender diagnosticity has the advantage that it acknowledges that some individual difference dimensions (like masculinity and femininity) are to a significant degree cultural constructs and thus must be flexibly assessed in different groups and in different historical eras. If the goal of research is to optimally predict behavior within a given population at a given time, then the gender diagnosticity approach may yield greater predictive utility than traditional standardized M–F scales. The gender diagnosticity approach may also be particularly appropriate for long-term longitudinal studies that must deal with the thorny question of how to best extract dispositional consistencies from samples of behavior that change over the lifespan and that possess meaning only in the context of the developmental epochs and historical eras in which they occur (see Caspi and Bem, 1990, for a broad discussion of this issue).

proach to assessing individual differences, it is particularly important to demonstrate its reliability.

Lippa and Connelly (1990) found that gender diagnostic probabilities tended to predict gender-related criteria more strongly within and across the sexes than did PAQ or BSRI scales. These results were particularly noteworthy given that the PAQ and BSRI scales had higher reliabilities than did within-sexes gender diagnosticity measures. The increased reliability of gender diagnosticity measures obtained in the current research holds the promise that gender diagnosticity measures may show even greater predictive validity in future research.

Indeed, although I have yet to complete analyses of validity data collected in the current study, preliminary evidence suggests that consistent with Lippa and Connelly's results, gender diagnostic probabilities predicted gender-related criteria better than PAQ and BSRI scales both within and across the sexes. Furthermore, preliminary evidence suggests that gender diagnostic probabilities also predicted criteria over all subjects better than did sex of subject.

For example, as in Lippa and Connelly's study, subjects' col-

Table 8
Factor Analyses of Four Gender Diagnosticity (GD) Measures, Personal Attributes Questionnaire (PAQ) Scales, Bem Sex-Role Inventory (BSRI) Scales, and Big Five Scales (Varimax Orthogonally Rotated Factor Matrices)

	All subjects (N = 260)			Men (n = 117)				Women (n = 143)		
Measure	Masculinity	Femininity	Gender diagnosticity	Masculinity-1	Masculinity-2	Femininity	Gender diagnosticity	Masculinity	Femininity	Gender diagnosticity
GD occupation	.19	.17	**.90**	.10	.22	−.04	**.79**	.18	−.13	**.83**
GD school subject	.15	−.02	**.84**	.12	.17	.11	**.76**	.04	.02	**.74**
GD everyday activities	.19	−.22	**.86**	.15	.05	−.18	**.80**	.19	−.21	**.77**
GD hobbies and amusements	.02	−.15	**.89**	−.09	.03	−.08	**.79**	−.14	−.02	**.78**
PAQ masculinity	**.84**	−.04	.19	**.67**	**.60**	−.07	.08	**.83**	−.01	.06
PAQ femininity	.06	**.82**	−.18	.17	−.03	**.82**	−.06	.11	**.79**	−.27
PAQ masculinity–femininity	**.70**	**−.38**	.12	.28	**.67**	**−.39**	.05	**.73**	**−.35**	.13
BSRI masculinity	**.86**	−.01	.21	**.75**	**.48**	−.08	.16	**.85**	.04	.14
BSRI femininity	−.04	**.85**	−.25	.22	−.15	**.84**	−.11	−.02	**.87**	−.14
Neuroticism	**−.52**	−.07	−.18	−.05	**−.61**	−.03	−.25	**−.58**	−.12	−.15
Extraversion	**.63**	.16	−.25	**.74**	−.04	.10	.10	**.70**	.08	−.13
Openness	**.63**	.25	.13	**.73**	.05	.25	.00	**.67**	.24	−.03
Agreeableness	−.07	**.80**	−.09	−.07	.15	**.84**	−.02	−.07	**.77**	−.02
Conscientiousness	.31	**.58**	.00	.01	.61	**.44**	.11	.31	**.67**	.10

Note. Factor loadings in boldface are greater than .35.

lege majors were rated on a 7-point scale of stereotypic M and F. Over all subjects, gender diagnosticity (grand mean) correlated more strongly with rated M and F of subjects' majors (r = .37, p < .001) than did sex of subject (r = .23, p < .001) or any PAQ or BSRI scale (highest correlation was with BSRI M, r = .16, p < .01). Thus, gender diagnosticity accounted for well over twice as much variance in the M and F of subjects' majors as did sex of subject or any PAQ or BSRI scale. The same pattern held within the sexes: Gender diagnosticity correlated more strongly with M and F of majors (.37 for men, p < .001, and .29 for women, p < .001) than did any PAQ or BSRI scales (highest correlation was .16 and mean correlation was .08). These preliminary analyses imply that the more highly reliable gender diagnosticity measures obtained in the current study displayed higher validity as well.

Discriminant Validity Compared With Contrasted-Groups Scales

Our data suggested some overlap between gender diagnosticity measures and contrasted-groups scales, particularly when male and female subjects were pooled together. This finding is not surprising given that both gender diagnosticity measures and contrasted-groups scales correlated substantially with sex of subject. Indeed, they were constructed to display such correlations. Although both methods generally yielded reliable scales, I uncovered two noteworthy differences between the two approaches: (a) Gender diagnosticity measures correlated more strongly with sex of subjects than did contrasted-groups scales and (b) gender diagnosticity measures tended to be more predictive of subjects' self-ascribed M and F than were contrasted-groups scales, particularly over all subjects and for men. Thus, these data suggest that gender diagnosticity measures are not identical to contrasted-groups scales.

The findings were complicated by the fact that the pattern in the data was not the same for men as for women. For example, gender diagnosticity measures and contrasted-groups scales seemed to assess separate constructs for men but not for women. Furthermore, gender diagnosticity measures predicted self-ascribed M and F better than contrasted-groups scales for men, but not for women. The current data serve to underscore Lippa and Connelly's (1990) finding that gender-related variables seem to be patterned differently for men than for women.

The differing findings for men and women in the current study should be explored further. For now, they provide gender researchers with the following cautionary message: Gender-related individual differences may sometimes need to be assessed and conceptualized differently for men than for women. Both research on unidimensional M and F (Gough, 1964; Hathaway & McKinley, 1943; Strong, 1943; Terman & Miles, 1936) and more recent work on M and F as separate dimensions (Bem, 1974; Heilbrun, 1976; Spence et al., 1974) have implicitly assumed that their assessment instruments applied with equal validity to men and to women. This assumption should be scrutinized more carefully. For example, recent research indicates that PAQ and BSRI items that best predict gender-related behaviors for men may not be those that best predict the same behaviors for women (Lippa, 1991).

Consistency Across Domains

In the current research, I computed gender diagnostic probabilities from ratings of occupations, school subjects, everyday activities, and amusements and hobbies. In general, these gender diagnosticity measures proved to be substantially intercorrelated, both within the sexes and over all subjects. In addition to providing evidence for the psychometric stability and reliability of gender diagnosticity measures, these data provided some support for the more substantive point that gender-related behaviors may show some degree of coherence and consistency within individuals when they are translated into appropriate individual difference measures.

A number of recent reviewers have suggested that gender-related behaviors are inconsistent within individuals and situationally variable (Ashmore, 1990; Deaux, 1985, 1987; Deaux & Major, 1987; Spence, Deaux, & Helmreich, 1985; Spence & Helmreich, 1978, 1980; Spence & Sawin, 1985). Drawing on the earlier work of Huston (1983), Ashmore (1990) has recently proposed five general domains of gender-related behaviors: personal and social attributes, social relationships, interests and abilities, symbolic and stylistic behaviors, and biological, physical, and material attributes. Under this conceptual scheme, it is clear that the gender diagnosticity measures computed in the current study focus on the "interests and abilities" domain.

Over the past 15 years, research on gender-related individual differences has focused almost exclusively on personal attributes (i.e., self-ascribed personality traits, such as those assessed by the PAQ and BSRI). The method of gender diagnosticity encourages researchers to focus attention on other important domains of gender-related behaviors. One fruitful direction for future research will be to determine the degree of coherence that exists among gender diagnosticity measures assessed across the five broad behavioral domains proposed by Ashmore.

Regardless of the outcome of such research, the method of gender diagnosticity provides a useful method for assessing gender-related individual differences in various domains of gender-related behaviors. Ashmore (1990) argues for a "loose glue" model of gender—that gender-related behaviors, thoughts, and feelings are only loosely interrelated. The method of gender diagnosticity provides a uniform metric of gender-related individual differences across various domains and thus provides a new way of empirically studying how "loose the glue" actually is.

The current study found evidence for a substantial degree of coherence among gender-related individual differences that were assessed from four different domains of interests. One reason why the current study showed such coherence may be that the method of gender diagnosticity focused on actual rather than stereotypic sex differences. In this regard it is useful to contrast the current study with one by Orlofsky (1981), in which subjects were asked to rate how typical various behaviors are of men and women and how valued the same behaviors are for men and women. Orlofsky used these ratings to develop scales of masculine and feminine behaviors in different domains. When subjects rated themselves on these scales, Orlofsky found only modest correlations between gender-related behaviors in difference domains. Unlike subjects' scores on Orlofsky's scales, gender diagnostic probabilities are computed from behavioral indices that actually (rather than stereotypically) differentiate men from women.[5]

The method of gender diagnosticity may also prove to be more successful than previous research methods in demonstrating coherent gender-related individual differences, because it leads to an optimal aggregation of gender-related behaviors. For example, in the current study, subjects' ratings of their degree of preference for 131 occupations provided a large set of ratings available for aggregation. However, the method of gender diagnosticity does not include aggregation dictated by a priori considerations (e.g., combining items that are stereotypically masculine or feminine), nor does it indiscriminately aggregate all items. Rather, it uses an optimal, statistically driven process of aggregation: Items that are strongly predictive of gender are weighted highly in the discriminant analyses, and they thus most influence the discriminant function scores that are used to compute gender diagnostic probabilities. In a sense, the discriminant analyses used to compute gender diagnostic probabilities optimally "squeeze out" of the ratings the variance tapping gender-related individual differences in that population. The evidence presented here suggests that the method of gender diagnosticity may be more successful in "squeezing out" such variance than are traditional contrasted-groups scaling techniques.

There is yet a third reason why gender diagnosticity measures may show gender-related individual differences to be more consistent and coherent than would be suggested by other research methods: Diagnostic probability measures of individual differences may possess desirable statistical and metric properties. For example, most personality measures are not ratio scales (see Buss and Craik, 1984, for a discussion of this issue in relation to the act-frequency approach to personality assessment). Gender diagnostic probabilities constitute a true ratio scale of individual differences, and they define an individual difference measure with a fixed range (0–1) and rich interpretations as probabilities. Diagnostic probability measures thus constitute a new domain of inquiry for the field of psychometrics.

Gender-Related Individual Differences and the Big Five

The current data demonstrated that M and F as assessed by the PAQ and BSRI are redundant with Big Five dimensions, whereas gender diagnosticity measures are not. These findings are interesting at face value, for they suggest that gender diagnosticity measures assess a unique domain of gender-related individual differences, whereas existing M and F scales do not. Other researchers have noted that M and F scales tap broader personality domains (Paulhus, 1987; Wiggins & Holzmuller, 1978), and this may help explain why they relate to gender-related behaviors only indirectly (Spence & Helmreich, 1980). Gender diagnosticity measures, unlike traditional bipolar M–F scales, seem to be unrelated to the Big Five, yet at the same time

[5] Items that are chosen to assess stereotypical sex differences do not necessarily show actual sex differences in specific populations of subjects. Indeed, recent research (Lippa, 1991) shows that many of the items that make up the M and F scales of the PAQ and the BSRI do not show strong sex differences in contemporary college student populations.

related to a host of substantive psychological variables, including cognitive abilities, self-reported aggressiveness, gender-related nonverbal behaviors, gender-related aspects of self-concept, and self-ascribed M and F (see Lippa & Connelly, 1990).

In summary, the results presented here suggest that gender diagnostic probabilities can be measured with high reliability, that they display discriminant validity in comparison with corresponding contrasted-groups scales, that they are relatively consistent across four interest domains, and that they are largely independent of the Big Five and of M and F as assessed by the PAQ and BSRI. These results provide additional evidence that gender diagnosticity measures are psychometrically sound and worthy of continued research attention.

References

Ashmore, R. D. (1990). Sex, gender, and the individual. In L. A. Pervin (Ed.), *Handbook of personality theory and research* (pp. 486–526). New York: Guilford Press.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Counseling and Clinical Psychology, 42,* 155–162.

Bem, S. L. (1981) *Bem Sex-Role Inventory: Professional manual.* Palo Alto, CA: Consulting Psychologists Press.

Buss, D. M., & Craik, K. H. (1984). Acts, dispositions, and personality. In B. A. Maher & W. B. Maher (Eds.), *Progress in experimental personality research* (pp. 241–301). San Diego, CA: Academic Press.

Campbell, D. F. (1966). *Revised manual for Strong Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB-SCII, Strong-Campbell Interest Inventory, Form T325 of the Strong Vocational Interest Blank* (3rd ed.). Stanford, CA: Stanford University Press.

Caspi, A., & Bem, D. J. (1990). Personality continuity and change across the life course. In L. A. Pervin (Ed.), *Handbook of personality theory and research* (pp. 549–575). New York: Guilford Press.

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

Deaux, K. (1985). Sex and gender. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 36, pp. 49–81). Palo Alto, CA: Annual Reviews.

Deaux, K. (1987). Psychological constructions of masculinity and femininity. In J. M. Reinisch, L. A. Rosenblum, & S. A. Sanders (Eds.), *Masculinity/femininity: Basic perspectives,* (pp. 289–303). New York: Oxford University Press.

Deaux, K., & Major, B. (1987). Putting gender in context: An interactive model of gender-related behavior. *Psychological Review, 94,* 369–389.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. In M. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 41, pp. 417–440). Palo Alto, CA: Annual Reviews.

Goldstein, M., & Dillon, W. R. (1978). *Discrete discriminant analysis.* New York: Wiley.

Gough, H. G. (1964). *California psychological inventory: Manual.* Palo Alto, CA: Consulting Psychologists Press.

Hand, D. J. (1981). *Discrimination and classification.* New York: Wiley.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory.* New York: Psychological Corporation.

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

Huston, A. C. (1983). Sex-typing. In E. M. Hetherington (Ed.), *Handbook of child psychology (4th ed.): Vol. 4. Socialization, personality, and social development* (pp. 387–467). New York: Wiley.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality theory and research* (pp. 66–100). New York: Guilford Press.

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality, 2,* 171–203.

Lachenbruch, P. A. (1975). *Discriminant analysis.* New York: Hafner Press.

Lippa, R. (1991). Sex differences define gender-related individual differences within the sexes: Evidence from three studies. Unpublished manuscript.

Lippa, R., & Connelly, S. (1990). Gender diagnosticity: A new Bayesian approach to gender-related individual differences. *Journal of Personality and Social Psychology, 59,* 1051–1065.

McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology, 36,* 929–940.

McCauley, C., Stitt, C. L., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin, 87,* 195–208.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality. *Journal of Personality and Social Psychology, 52,* 81–90.

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). New York: Wiley.

Norusis, M. J. (1986). *Advanced statistics: SPSS/PC+ for the IBM PC/XT/AT.* Chicago: SPSS.

Norusis, M. J. (1988). *SPSS/PC+ V2.0 Base Manual for the IBM PC/XT/AT AND PS/2.* Chicago: SPSS.

Orlofsky, J. L. (1981). Relationships 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, 40,* 927–940.

Paulhus, D. L. (1987). Effects of group selection on correlations and factor patterns in sex role research. *Journal of Personality and Social Psychology, 53,* 314–327.

Spence, J. T. (1984). Masculinity, femininity, and gender-related traits: A conceptual analysis and critique of current research. *Progress in Experimental Personality Research, 13,* 1–97.

Spence, J. T., Deaux, K., & Helmreich, R. L. (1985). Sex roles in contemporary American society. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, pp. 149–178). New York: Random House.

Spence, J. T., & Helmreich, R. L. (1978). *Masculinity and femininity: Their psychological dimensions, correlates, and antecedents.* Austin, TX: University of Texas Press.

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

Spence, J. T., Helmreich, R., & Stapp, J. (1974). The Personal Attributes Questionnaire: A measure of sex role stereotypes and masculinity–femininity. *JSAS Catalog of Selected Documents in Psychology, 4,* 43 (ms. No. 617).

Spence, J. T., & Sawin, L. L. (1985). Images of masculinity and femininity: A reconceptualization. In V. E. O'Leary, R. K. Unger, & B. S. Wallston (Eds.), *Women, gender, and social psychology* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Strong, E. K. (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.

Terman, L. M., & Miles, C. C. (1936). *Sex and personality: Studies in masculinity and femininity.* New York: Russell & Russell.

Wiggins, J. S., & Holzmuller, A. (1978). Psychological androgyny and interpersonal behavior. *Journal of Consulting and Clinical Psychology, 46,* 40–52.

Received May 2, 1990
Revision received January 22, 1991
Accepted February 14, 1991 ■