Genetic and Environmental Influences on Individual Differences in Masculinity, Femininity, and Gender Diagnosticity: Analyzing Data From a Classic Twin Study

Richard Lippa
California State University, Fullerton
Scott Hershberger
California State University, Long Beach

ABSTRACT Analyzing data from Loehlin and Nichols's (1976) classic twin study, we computed measures of Masculine Instrumentality (M), Feminine Expressiveness (F), and Gender Diagnosticity (GD). Quantitative genetic modeling analyses of within-sex individual differences in M, F, and GD indicated that: (1) Additive genetic factors contribute significantly to individual differences in M, F, and GD. (2) The environmental effects on M, F, and GD tend to be nonshared. (3) The genetic and environmental components of individual differences in M, F, and GD tend not to show gender differences. Finally, (4) the estimated within-sex heritability of GD (.53) is significantly greater than the estimated within-sex heritabilities of either M (.36) or F (.38).

Psychologists have tried to measure within-sex gender-related individual differences for over half a century. From 1936 to the early 1970s, such individual differences were conceptualized as falling along a bipolar

Correspondence concerning this article should be addressed to Richard Lippa, Department of Psychology, California State University, Fullerton, CA 92834-6846. Electronic mail may be sent via the Internet to rlippa@fullerton.edu

Journal of Personality 67:1, February 1999.

dimension of masculinity–femininity (M–F), which was assessed with scales composed of self-report items that showed sex differences in normative populations. In the 1970s a two-dimensional conception of masculinity and femininity arose, with masculinity (M) defined in terms of instrumental personality traits and femininity (F) in terms of expressive traits (Bem, 1974; Spence, Helmreich, & Stapp, 1974). This two-dimensional approach has guided research for the past 20 years. Still other approaches to conceptualizing and measuring within-sex gender-related individual differences have emerged recently (Helgeson, 1995; Lippa & Connelly, 1990; Spence, 1993; Spence & Buckner, 1995). One of these, the method of gender diagnosticity (GD; Lippa, 1995a; Lippa & Connelly, 1990), proposes that masculinity and femininity are historically and culturally varying dispositions that can be assessed only via behaviors that show sex differences in *particular* populations of men and women. (Research on bipolar M–F, M and F, and gender diagnosticity will be reviewed in more detail later in this article.)

A fundamental question applies to each of these approaches to "masculinity" and "femininity": What are the sources of the individual differences assessed by each method? This question can be subdivided into several more specific questions: (1) Do genetic factors contribute to observed gender-related individual differences? (2) Do environmental factors also contribute to these individual differences, and if so are they environmental factors that are common to families, or are they environmental factors that are unique to individuals? (3) Are the relative contributions of genetic and environmental factors to gender-related individual differences the same for boys and girls, and for men and women? And, finally, (4) Do the relative contributions of environmental and genetic factors differ for various measures of gender-related individual differences (i.e., for M–F, M, F, and GD)?

Behavioral genetic analyses provide a means to answer the questions just posed. By analyzing patterns of trait covariance in populations of twins, parents and siblings, and adoptive families, behavioral geneticists partition trait variance into hereditary, shared environmental, and unique environmental components. In this article we present such behavior genetic analyses on data from a classic study of 839 same-sex pairs of twins (Loehlin & Nichols, 1976). The richness of the data set permitted us to assess twins on masculine instrumentality (M; based on Adjective Check List responses and the Dominance scale of the California Personality Inventory scale), feminine expressiveness (F; based on Adjective

Check List responses and the Good Impression and Socialization scales of the California Personality Inventory), Gender Diagnosticity (GD; based on twins' self-reported everyday activities, occupational preferences, and CPI item responses), and bipolar M–F as assessed by the femininity (Fe) scale of the California Personality Inventory (Gough, 1957). We could compute and compare intraclass correlations for monozygotic and dizygotic twins for the measures just described, and by applying behavior genetic models to the corresponding covariance data through the use of structural equation modeling (implemented via LISREL analyses), we could estimate the contributions of hereditary and environmental factors to observed individual differences in M–F, M, F, and GD, and thereby obtain quantitative answers to the questions just posed.

To place the current analyses in a broader context, it helps first to briefly review recent research on the behavioral genetics of personality, with a particular focus on gender-related individual differences. Second, it is useful to review the history of psychologists' attempts to understand and measure within-sex gender-related individual differences.

Review of Previous Research

Behavioral Genetic Analyses of Personality

The behavioral genetic analysis of personality has been an active area of research in recent years (for reviews, see Loehlin, 1992; Loehlin & Rowe, 1992; Plomin, Chipuer, & Loehlin, 1990; Rose, 1995; Saudino & Plomin, 1996). Progress has been fostered by the emergence of agreed-upon taxonomies of broad domains of personality, such as the five-factor model of personality (also referred to as the Big Five model; see Costa & McCrae, 1992, 1995; Goldberg, 1993; John, 1990).

In an excellent and comprehensive review, Loehlin (1992; see also Loehlin & Rowe, 1992) summarized evidence from a series of behavioral genetic studies of the Big Five traits of Surgency (also called Extraversion), Agreeableness, Conscientiousness, Emotional Stability (also called Neuroticism or Negative Affectivity), and Culture/Openness (see Loehlin, 1992, Table 3.20, p. 67). Using two simple behavior genetic models to analyze and pool the results from a number of studies (one model allowed for the possibility of nonadditive genetic effects, and the other admitted the possibility that the environments of monozygotic twins might be more similar than those of dizygotic twins), Loehlin reached the following broad conclusions: Each of the Big Five traits shows appreciable additive genetic effects, with heritability defined in its narrow sense ranging from .22 to .46. Additive genetic effects are greatest for Surgency (Extraversion) and Culture/Openness and lowest for Conscientiousness and Agreeableness. The effects of shared sibling environments (those aspects of the environment that are shared by siblings, such as effects of socioeconomic status and general parenting styles) were relatively small, ranging from virtually no effect for Surgency to about .09 for Agreeableness. Finally, for all Big Five traits, roughly 50% of trait variance was not accounted for by modeled factors. That is, about 50% of variance was consigned to a "leftover" category that included some combination of individual (unique) environments, gene–environment interactions, and unreliability.

Loehlin's review is consistent with much recent behavioral genetic research on personality in that it showed that: (1) shared environmental effects tend to be weak (see Rowe, 1994, for a comprehensive review); (2) most personality traits demonstrate significant and substantial degrees of heritability; and (3) heritability—both in its narrow sense (just additive genetic effects) and in its broader sense (additive and nonadditive effects combined)—typically accounts for less than 50% of the observed variance in broad personality dispositions such as the Big Five.

Loehlin's findings for Surgency and Agreeableness are particularly relevant to the current research because measures of masculine instrumentality (M) overlap most with Extraversion and measures of feminine expressiveness (F) overlap most with Agreeableness (Lippa, 1991, 1995b). Several studies have estimated the heritability of M and F directly. For example, one small study of 38 monozygotic and 32 dizygotic twin pairs estimated the heritability of M and F measures in children and found, consistent with Loehlin's review, the following: (1) heritability was greater for M than for F, and (2) shared environmental effects were small (Mitchell, Baker, & Jacklin, 1989). However, this study was not able to estimate heritability separately for boys and girls because of its small sample size. Another small study of high school– and college-age twins found evidence for significant heritability for M, but not for F (Rowe, 1982). This study also combined males and female twin pairs in its heritability analyses. Thus, the limited available findings point to the tentative conclusion that M shows stronger heritability than F.

Support for differences in genetic and environmental contributions to M and F come also from research conducted under the auspices of the Minnesota Twin Study. In a study by Tellegen et al. (1988), 217 monozygotic and 114 dizygotic reared-together twin pairs, and 44 monozygotic and 27 reared-apart twin pairs, were assessed on the Multidimensional Personality Questionnaire, a broad-spectrum personality inventory that comprises scales for 11 primary traits and 3 higher-order factors (MPQ; Tellegen, 1982; Tellegen & Waller, in press). The MPQ primary scale that best approximates feminine expressiveness is Social Closeness, and the MPQ scale that best approximates masculine instrumentality is Social Potency.[1] Applying behavior genetic models to their data, Tellegen et al. (1982) estimated the additive genetic and the shared familial variance components for Social Closeness to be .40 and .19, respectively, and the corresponding components for Social Potency to be .54 and .10 (all significantly greater than zero, except for the last value). Thus, consistent with the findings reviewed earlier, both Social Closeness (the proxy F measure) and Social Potency (the proxy M measure) showed significant heritability, with Social Potency showing a tendency toward higher heritability than Social Closeness. Social Closeness showed stronger shared environmental effects than Social Potency; indeed, it was the only scale among the 11 MPQ primary personality scales to show a significant effect for shared familial environment. By implication, F would seem more likely than M to show significant shared familial effects.

Evidence for greater shared environmental effects for F than for M is also provided by a large-scale Swedish adoption and twin study (see Plomin, Chipuer, & Loehlin, 1990, pp. 230–233, for a summary). This study found, for both twins reared together and twins reared apart, strong evidence for significant genetic contributions to Extraversion. Intraclass correlations for identical twins were substantially more than twice the corresponding correlations for fraternal twins, suggesting nonadditive as well as additive genetic effects for Extraversion (and, by implication, perhaps for M). For Agreeableness (the Big Five proxy for F), the Swedish study found evidence for somewhat lower heritability than for

[1] According to Tellegen and Waller (in press), the individual who scores high on Social Closeness is "sociable, likes to be with people; takes pleasure in and values close interpersonal relationships; is warm and affectionate; turns to others for comfort and help," whereas the individual who scores high on Social Potency is "forceful and decisive; is persuasive and likes to influence others; enjoys or would enjoy leadership roles; enjoys being noticed, being the center of attention."

Extraversion; unlike Extraversion, Agreeableness showed shared environmental effects. This effect could be most directly observed by comparing the intraclass correlation on Agreeableness for identical twins reared together (.47) with the corresponding correlation for identical twins reared apart (.19).

Additional evidence for possible differences in the hereditary and environmental contributions to M, F, and M–F comes from a recent meta-analysis of twin studies conducted from 1967 to 1985 (McCartney, Harris, & Bernieri, 1990). Meta-analyses were conducted on eight dimensions of personality-temperament, including "dominance" (a clear M proxy) and "masculinity–femininity" (probably a proxy for bipolar M–F).[2] Unfortunately, no personality dimension in this review corresponded to F. Although McCartney, Harris, and Bernieri did not estimate components of variance directly, they did report mean intraclass correlations for identical and fraternal twins for the various personality dimensions. For the trait of dominance, mean correlations (averaged over 7 studies) were .51 for identical twins and .21 for fraternal twins. Falconer's (1960) method for computing an approximate estimate of heritability (h^2 = 2x [monozygotic correlation-dizygotic correlation]) yields h^2 = .60. The fact that the mean intraclass correlation for monozygotic twins was more than twice the value for dizygotic twins suggests the possibility of nonadditive genetic effects. For the trait of masculinity–femininity, the meta-analysis reported the mean (over 9 studies) intraclass correlations to be .52 for identical twins and .36 for fraternal twins, yielding an approximate heritability estimate of .32.

In sum, the research summarized above suggests that genetic effects may be stronger for M than for F. M–F also shows evidence for significant genetic effects, but heritability estimates for M–F seem to be lower than estimates for M, at least according to McCartney, Harris, and Bernieri's meta-analysis. However, this conclusion must be tentative, given uncertainty about how McCartney et al. classified personality scales as "M–F" measures. Although substantial shared environmental effects are typically not found for most personality traits, F (an Agreeableness proxy) may be an exception.

[2] McCartney, Harris, and Bernieri did not precisely describe how personality dimensions were classified as tapping "masculinity–femininity." The researchers simply noted that "dependent variables were classified via group consensus based on the knowledge of the researchers and on descriptions by authors" (p. 228).

Measuring Gender-Related Individual Differences: Masculinity, Femininity, and Gender Diagnosticity

Because measures of masculine instrumentality (M), feminine expressiveness (F), GD measures, and a bipolar M–F scale served as different measures of twins' gender-related individual differences in the current analyses, it is useful to describe in more detail the history and methodologies of these various approaches to assessing "masculinity" and "femininity."

Modern research on the measurement of within-sex gender-related individual differences began with the 1936 publication of Terman and Miles's *Sex and Personality*, which presented a bipolar conception of masculinity–femininity (M–F). In essence, this approach held that M–F is a single dimension, with masculinity and femininity as mutually exclusive end points. Terman and Miles and their many successors created M–F scales from items that showed reliable and strong sex differences in normative populations. One of the scales used in the research to be reported here, the Fe (Femininity) scale of the California Psychological Inventory (Gough, 1957), was developed in the Terman and Miles tradition of bipolar M–F scales. Indeed, in recent revisions of the CPI (Gough, 1987), this scale is referred to as a "Femininity/Masculinity" scale.

The bipolar approach to M–F waned by the early 1970s in the face of conceptual and empirical critiques (e.g., Block, 1973; Constantinople, 1973) that argued, in part, that supposedly unidimensional M–F scales were, in fact, multidimensional measures. The bipolar approach was supplanted in the early 1970s by a two-dimensional conception of masculinity and femininity, which has been dominant for the past 20 years. The two-dimensional approach holds that masculinity and femininity are separate dimensions, with Masculinity (M) defined in terms of instrumental personality traits (e.g., *aggressive*, *dominant*, *independent*) and Femininity (F) defined in terms of expressive traits (*warm, sensitive, nurturant*). During the 1970s a number of self-report inventories were developed to assess M and F as two separate dimensions. The best known of these are the Bem Sex-Role Inventory (BSRI; Bem, 1974, 1981a) and the Personal Attributes Questionnaire (PAQ; Spence, Helmreich, & Stapp, 1974; Spence & Helmreich, 1978). A large empirical literature now exists on the psychometric properties and correlates of M and F as assessed by these scales (see Ashmore, 1990; Cook, 1985; Lenney, 1991).

Although the PAQ and BSRI did not exist at the time when the data were collected for Loehlin and Nichols's (1976) classic twin study (in the early 1960s), these data contain a number of measures that can serve as proxy measures for M and F. For example, participants in Loehlin and Nichols's twin study completed a 159-item version of the Adjective Check List (ACL), and some ACL items directly assess instrumental traits (e.g., *assertive, dominant, independent*) while others directly assess expressive traits (e.g., *cooperative, helpful, kind*). In addition, Loehlin and Nichols's twins completed the California Psychological Inventory (CPI; Gough, 1957), and some CPI scales overlap with M (e.g., the CPI Dominance scale), and others overlap with F (e.g., the CPI Good Impression and Socialization scales). All of these measures provided a means of indirectly assessing Loehlin and Nichols's twins on M and F.

Although M and F scales continue to be widely used in research on gender-related individual differences, these scales have been subject to both psychometric and conceptual critiques. Some researchers have argued that M and F scales do not really measure "masculinity" and "femininity" at all. Indeed, Spence and Helmreich (1980) argued soon after publishing the PAQ that M and F scales are, in fact, instrumentality and expressiveness scales, which show at best weak and inconsistent relationships to other kinds of gender-related behaviors and attitudes (see Spence & Buckner, 1995, for a recent theoretical discussion).

M and F scales and their associated constructs may suffer from additional problems as well (Lippa & Connelly, 1990; Lippa, 1991, 1995a, 1995b, 1998). As BSRI author Sandra Bem noted in her later work on gender schemas (Bem, 1981b, 1985), M and F scales may inappropriately reify gender-related individual differences and confuse psychologists' formal constructs of M and F with lay conceptions of masculinity and femininity. The reification implicit in M and F scales may restrict masculinity and femininity to broad, but overly limited, domains of behavior (Lippa, 1995b). For example, although M and F scales adequately describe the gender stereotypic personality traits that are components of most cultures' conceptions of masculinity and femininity (e.g., see Williams & Best, 1990), they fail to embrace a host of other characteristics that are highly relevant to everyday conceptions of masculinity and femininity—characteristics such as gender-related appearances; nonverbal behaviors; hobbies and interests; sexual behaviors;

ways of relating to friends, spouses, and lovers; and so on. And because of their fixed and limited item content, M and F scales fail to acknowledge that masculinity and femininity are fluid concepts that are, to some extent, culturally and historically relative.

The gender diagnosticity approach was developed to address some of these problems (Lippa, 1991, 1995a, 1995b; Lippa & Connelly, 1990). In brief, gender diagnosticity (GD) refers to the Bayesian probability that an individual is predicted to be male or female based on some set of gender-related indicators (such as occupational preference ratings). According to the GD perspective, a masculine person is an individual who shows “malelike” behaviors in comparison to a contemporaneous reference group of males and females, and a feminine person is an individual who shows “femalelike” behaviors.

For a single gender-related behavior, the computation of gender diagnostic probabilities is quite straightforward. For the sake of illustration, assume that in a given sample of men and women, 100% of the men and 50% of the women wear pants. GD refers to the computed probability that an individual is male or female given that he or she wears pants. Applying Bayes’s theorem to this example, we note that p (female|wears pants) = p (female) × p (wears pants|female)/p (wears pants). Assuming for the sake of simplicity that the base-rate probability of being female—p (female)—is .5, we compute that the probability that an individual is female given that he or she wears pants is .5 × .5/.75 = .33. The probability that an individual is male given the same diagnostic information is simply the complementary probability, or .67.

One virtue of the GD approach is that it acknowledges that a particular indicator of masculinity or femininity may vary over time and over different populations of men and women. For example, the behavior of “wearing pants” was more gender diagnostic 100 years ago than it is today in the United States, and it is currently more gender diagnostic in some countries than in others. By implication, an American woman wearing pants 100 years ago would have been judged more masculine as a result of her behavior than would an American woman wearing pants today, and a woman wearing pants in contemporary Saudi Arabia would likely be judged by members of her culture to be more masculine as a result of her behavior than would a woman wearing pants in contemporary America be judged by members of her culture.

GD is formally computed from sets of indicators through the application of discriminant analyses (see Lippa, 1991, 1995b; Lippa & Connelly, 1990; this process will be described more fully later in this article). Discriminant analysis identifies the linear combination of predictor variables—the discriminant function—that optimally discriminates membership in two categories or groups. To compute gender diagnostic probabilities of individuals in a particular population of males and females, a discriminant analysis is applied to a set of gender-related variables such as occupational preference ratings. This analysis generates a discriminant function, a weighted combination of predictor variables that optimally classifies individuals (based on some cutoff value) as male or female. Bayes's theorem is then applied to individuals' discriminant function scores to compute the probability that an individual is male or female. (The computation of such probabilities is a standard option in computerized statistical packages that perform discriminant analyses.)

Prior research on GD shows that it can be measured reliably within the sexes from self-report data such as occupational preference ratings and that GD measures are factorially distinct from M and F as assessed by the PAQ and BSRI (Lippa, 1995b, 1991; Lippa & Connelly, 1990). Furthermore, GD measures are largely independent of the Big Five personality superfactors, whereas M and F are not (Lippa, 1995b, 1991). Indeed, M and F correlate substantially with Big Five dimensions, with M loading highly on Extraversion and Neuroticism and F on Agreeableness. Finally, GD measures often predict varied gender-related behaviors and attitudes *within the sexes* (e.g., cognitive abilities, nonverbal masculinity–femininity, masculinity–femininity of chosen college major, self-ascribed masculinity–femininity, placement on fundamental dimensions of vocational interests, authoritarianism in men, sexual orientation, attitudes toward women's roles, and attitudes toward gay people) better than M and F do (Lippa, 1991, 1995b, 1997, 1998a, 1998b, 1998c; Lippa & Arad, 1997; Lippa & Connelly, 1990).

Loehlin and Nichols's (1976) twin data provide an ideal resource for the computation of GD measures, since the participating twins were assessed on their self-rated frequency of engaging in 324 everyday activities, their self-rated interest in 160 occupations listed in the Vocational Preference Inventory (Holland, 1958), and their responses to 480 items of the CPI (Gough, 1957). Thus, we were able to compute GD measures based on these three item sets.

An Overview of the Study

As noted previously, members of 839 twin pairs were assessed on the Adjective Check List (ACL), the California Psychological Inventory (CPI), their degree of participation in 324 everyday activities, and their preferences for 160 occupations. The Fe (Femininity) scale of the CPI served as a bipolar measure of M–F. ACL scales of expressive traits and the CPI Good Impression and Socialization scales served as proxy measures for F. ACL scales of instrumental traits and the CPI dominance scale served as proxy measures of M. Finally, GD measures were computed from three item sets: (1) self-reported participation in everyday activities, (2) occupational preferences, and (3) CPI items.

Because within-sex gender-related individual differences were operationalized in several different ways, we factor-analyzed our measures (separately for males and females) to ascertain that measures cohered as expected (e.g., that ACL expressiveness, CPI Good Impression, and CPI Socialization all tapped a feminine expressiveness dimension). As we shall show, scales generally did show the expected coherence: ACL instrumental traits and CPI dominance loaded highly on a masculine instrumentality factor; ACL expressive traits and CPI Good Impression and Socialization loaded on a feminine expressiveness factor; and GD based on occupational preferences, GD based on everyday activities, GD based on CPI items, and CPI Fe all loaded highly on a factor that could be labeled GD or bipolar M–F. Factor scores were created for twins on the three factors just described.

Thus, the groundwork was laid for the central analyses of this article: examination of intraclass correlations and corresponding covariance data for identical and fraternal twins on M, F, and GD. These correlations were computed separately for males and females, and indeed one great strength of the current data is that the sample size is large enough to permit meaningful comparisons of estimates for male and female twins. Using LISREL analyses, we tested specific behavior genetic models on the various measures of gender-related traits computed from Loehlin and Nichols's twin data. The goal of these analyses was generally to estimate three parameters for each assessed trait: *a* (additive genetic effects), *e* (nonshared environmental effects), and *c* (shared environmental effects). In the process of applying behavior genetic models to the data, we hoped to shed some light on the questions posed earlier: (1) Are there significant genetic components to individual differences in M, F, and GD? (2) Are there significant environmental components to individual differences in M, F, and GD, and if so, do they represent shared or nonshared effects? (3) Does the magnitude of genetic and environmental components differ for males and females? And (4) does the magnitude of genetic and environmental components differ for the three measures of within-sex gender-related individual differences (M, F, and GD), and also for the component measures comprising GD?

METHOD

The Data Set and Subjects

The data analyzed here are a subset of the data described by Loehlin and Nichols (1976) in their classic book *Heredity, Environment, and Personality*.[3] In this study, 839 pairs of twins who took the National Merit Qualifying Test in 1962 were assessed on a variety of self-report scales and inventories, including measures of their degree of participation in everyday activities, their occupational interests, and their personality. More specifically, self-report measures included self-rated frequency of engaging in 324 everyday activities (e.g., “Played checkers,” “Made minor repairs around the house,” “Said grace before meals,” “Rode a motorcycle,” “Drove a car over 80 M.P.H.”), self-rated interest in 160 occupations (e.g., “Aviator,” “Private investigator,” “YMCA secretary,” “Nursery school teacher,” “Lawyer”), and completion of 480 items of the California Psychological Inventory (CPI; Gough, 1957) and a 159-item version of the Adjective Check List (ACL). Personality questionnaires and self-report inventories were mailed to subjects in 1963, when subjects were approaching the end of their senior year in high school.

In soliciting their sample, Loehlin and Nichols (1976) contacted all same-sex twins (1,507 pairs) identified from the almost 600,000 United States high school juniors who took the National Merit Test in 1962. Because of the conscientiousness of their research effort, Loehlin and Nichols obtained an impressive response rate of 79%. The final sample included 216 pairs of male identical twins, 135 pairs of male fraternal twins, 293 pairs of female identical twins, and 195 pairs of female fraternal twins—yielding a grand total of 839 twin pairs comprising 1,678 individuals. Loehlin and Nichols reported that their twins’ mean scores on California Psychological Inventory scales were quite close to high school and college norms presented in the CPI manual. Thus the twins seemed quite typical and representative of their peers in terms of assessed personality.

[3] We wish to express great appreciation to John Loehlin for giving permission to analyze the raw data from Loehlin and Nichols’s (1976) study and for electronically transmitting the data file to the first author.

RESULTS

Measures of Gender-Related Individual Differences Computed in the Current Research

In the analyses that follow, Twin A will refer to the first member of a twin pair, and Twin B will refer to the second member of a pair. To ensure that the data in a given analysis were statistically independent, data for Twins A were at times analyzed separately from data for Twins B.

Subjects' gender diagnosticity (GD) scores were computed from three kinds of self-report data: ratings of their participation in everyday activities, ratings of occupational preferences, and self-ratings on items of the California Psychological Inventory (see Lippa & Connelly, 1990, and Lippa, 1991, 1995b, for additional details about the computation and psychometrics of gender diagnosticity measures).

To compute gender diagnostic probabilities from subjects' occupational preference ratings, 16 discriminant analyses were conducted on discrete sets of 10 occupations each. Thus, the 16 discriminant analyses included all 160 occupational preference items. Each discriminant analysis yielded the probability, computed from each subject's discriminant function score, that a given subject was male (or, by subtracting this probability from 1, female). Thus, on the basis of their occupational preference ratings, each subject had 16 separate gender diagnostic probabilities, each computed from a distinct subset of rated occupations. A subject's overall GD score was simply the average of the 16 probabilities. GD scores were computed separately for Twins A and Twins B.

The reason multiple gender diagnostic probabilities were computed for each subject was to permit the assessment of their reliability (see Lippa & Connelly, 1990; Lippa, 1991, 1995b). The reliability of GD based on occupational preferences was high for all subjects (alpha = .93 for Twins A and .92 for Twins B) as well as for men only (alpha = .85 for Twins A and .87 for Twins B) and for women only (alpha = .82 for Twins A and .77 for Twins B).

Similarly, to compute gender diagnostic probabilities (GD scores) from subjects' everyday activity ratings, 15 discriminant analyses were conducted on discrete sets of 21 or 22 activities each. Thus, the 15 discriminant analyses included all 324 everyday activity items. The reliability of GD based on everyday activity items was high for all subjects (alpha = .95 for Twins A and .95 for Twins B). Within-sex reliabilities were somewhat lower than for GD computed from occupational preferences (for men, alpha = .56 for Twins A and .65 for Twins B; for women, alpha = .61 for Twins A and .58 for Twins B).

Finally, to compute gender diagnostic probabilities (GD scores) from subjects' California Psychological Inventory (CPI) items, 15 discriminant analyses were conducted on discrete sets of 32 items each. Thus, the 15 discriminant analyses included all 480 personality items. The reliability of GD based on CPI items was high for all subjects (alpha = .85 for Twins A and .90 for Twins B). Within-sex reliabilities were again somewhat lower for GD based on CPI items than for GD computed from occupational preferences (for men, alpha = .47 for Twins A and .68 for Twins B; for women, alpha = .57 for Twins A and .67 for Twins B).

Subjects' scores on masculine instrumentality (M) and feminine expressiveness (F) were computed from their Adjective Check List responses. The following eight Adjective Check List items provided a relatively pure measure of masculine instrumentality: *Aggressive, Assertive, Confident, Dominant, Forceful, Outspoken, Self-Confident*, and *Independent*. Many of these items are quite similar to and even identical to items that appear on the PAQ and BSRI M scales. The following seven Adjective Check List items provided a relatively pure measure of feminine expressiveness: *Cooperative, Helpful, Kind, Sensitive, Tactful, Thoughtful*, and *Warm*. Again, many of these items are either quite similar to or even identical to items that appear on the PAQ and BSRI F scales.

A subject's M score was simply the total number of M items he or she endorsed as self-descriptive, and a subject's F score was the total number of F items endorsed as self-descriptive. The reliabilities of ACL M were .72 for male Twins A, .66 for female Twins A, .72 for male Twins B, and .67 for female Twins B. The reliabilities of ACL F were .70 for male Twins A, .66 for female Twins A, .65 for male Twins B, and .67 for female Twins B. Because ACL scales can be influenced by participants' overall tendency to endorse ACL items, normalized M and F scores were also computed by dividing a participant's raw ACL M and F score by the total number of ACL items endorsed by the participant.[4]

As noted before, the CPI Dominance scale served as another measure of M. In the 1957 manual to the CPI (Gough, 1957), individuals high on the Dominance scale are described as "aggressive, confident, persistent . . . ;

[4] Normalized scores were also computed using regression techniques: Residual M and F scores were computed after subtracting variance predicted by total number of ACL items endorsed. The two kinds of normalized M and F scores proved to be virtually identical, correlating .98 and above.

self-reliant and independent; . . . having leadership potential and initiative." The overlap with masculine instrumentality is apparent in this capsule description. CPI Good Impression and Socialization served as proxy measures of F. Gough (1957) describes individuals high on the Good Impression scale as "co-operative . . . warm, and helpful," and individuals high on the Socialization scale as "honest . . . modest, obliging, sincere." In the 1987 Administrator's Guide to the CPI, Gough (1987) presents evidence that the peers and spouses of individuals high on the Socialization scale tend to describe them as "reasonable," "kind," "cooperative," "tactful," "appreciative," and "considerate." Thus both the CPI Good Impression and Socialization scales show important areas of overlap with feminine expressiveness.

Finally, the CPI Fe scale served as a measure of bipolar M–F. Gough (1957) reported that the CPI Fe scale showed a strong point biserial correlation with subject sex in populations of high school students, college students, and psychology graduate students, and that in an adult male population CPI Fe correlated moderately with the M–F scales of the Strong Vocational Interest Blank and the MMPI.

Descriptive statistics for the various measures of gender-related individual differences are shown in Table 1, computed separately for male Twins A, male Twins B, female Twins A, and female Twins B. T-tests indicated that all measures showed highly significant gender differences, except for the CPI Good Impression scale, which did not significantly differ for men and women.

Factor Analyses of Measures of Gender-Related Individual Differences

Principal component analyses with three-factor solutions subjected to orthogonal varimax rotation were conducted on the following measures of within-sex gender-related individual differences: GD based on occupational preferences, GD based on everyday activities, GD based on CPI items, CPI Femininity, ACL feminine expressiveness (raw), ACL feminine expressiveness (normalized), CPI Good Impression, CPI Socialization, ACL masculine instrumentality (raw), ACL masculine instrumentality (normalized), and CPI Dominance.[5] Four separate analyses were

[5] Both raw and normalized ACL scales were included because they were the purest measures of Masculine Instrumentality and Feminine Expressiveness in the current study, whereas CPI scales were imperfect proxies for these constructs. Including both raw and

Table 1
Descriptive Statistics for Measures of Gender-Related Individual Differences

Measure		Mean	Standard Deviation
GD Occupations	Males	.64	.12
(Twins A)	Females	.36	.12
GD Occupations	Males	.64	.13
(Twins B)	Females	.35	.11
GD Activities	Males	.76	.09
(Twins A)	Females	.23	.09
GD Activities	Males	.76	.10
(Twins B)	Females	.22	.09
GD CPI	Males	.65	.09
(Twins A)	Females	.34	.09
GD CPI	Males	.66	.10
(Twins B)	Females	.34	.10
CPI Fem	Males	16.75	3.64
(Twins A)	Females	23.86	3.36
CPI Fem	Males	16.71	3.86
(Twins B)	Females	24.13	3.00
Raw Feminine Expressiveness	Males	4.01	1.97
(Twins A)	Females	4.54	1.82
Raw Feminine Expressivenss	Males	3.91	1.87
(Twins B)	Females	4.54	1.87
Normalized Fem. Expressiveness	Males	.08	.04
(Twins A)	Females	.09	.03
Normalized Fem. Expressiveness	Males	.08	.04
(Twins B)	Females	.09	.03
CPI Good Impression	Males	17.31	6.11
(Twins A)	Females	16.73	5.78
CPI Good Impression	Males	16.82	6.18
(Twins B)	Females	16.72	5.72
CPI Socialization	Males	39.30	5.06
(Twins A)	Females	41.11	4.66
CPI Socialization	Males	39.26	5.23
(Twins B)	Females	41.22	4.86
Raw Masculine Instrumentality	Males	2.54	2.07
(Twins A)	Females	1.89	1.77

Table 1
Continued

Measure		Mean	Standard Deviation
Raw Masculine Instrumentality	Males	2.54	2.08
(Twins B)	Females	1.82	1.77
Normalized Masc. Instrumentality	Males	.04	.03
(Twins A)	Females	.03	.03
Normalized Masc. Instrumentality	Males	.05	.03
(Twins B)	Females	.03	.03
CPI Dominance	Males	27.95	5.91
(Twins A)	Females	26.92	5.69
CPI Dominance	Males	27.86	6.01
(Twins B)	Females	26.76	6.06

conducted, one for each of the following subject groups: male Twins A, male Twins B, female Twins A, and female Twins B. All factor analyses yielded quite similar results; for the sake of illustration, the rotated factor matrix for male Twins A is presented in Table 2.[6]

The three extracted factors shown in Table 2 accounted for 59% of the total variance in measures. Three distinct and highly interpretable factors emerged from the analysis. The three GD measures and CPI femininity

normalized scales had the effect of weighting these scales more in the final factor scores. Another reason for including both raw and normalized scores was the possibility that the normalization procedure may have overcorrected scores (in the sense that normalized M and F were negatively correlated, whereas raw M and F were positively correlated). Including both raw and normalized scores in the factor analysis helped compensate for any overcorrection that may have taken place.

As an empirical check, factor scores were also computed in factor analyses that excluded raw ACL scales as variables. In general, the same factors emerged from these analyses as in the factor analyses reported in the main article. The median correlation (over the four groups: male Twins A, male Twins B, female Twins A, and female Twins B) of the M factor scores created by the two methods was .92; the median correlation of the F factor scores was .88; and the median correlation of the GD factor scores was .99. In other words, the factor scores created by the two methods were very similar.

[6] Clearly, the current factor analyses indicate that GD has much more in common with bipolar M–F (e.g., as measured by CPI Fe) than with masculine instrumentality or feminine expressiveness. Still, it is important to note that GD measures are not identical to bipolar M–F scales. At least two studies (Lippa, 1991; Lippa, 1998a) have shown that GD measures at times show greater validity than traditional M–F scales based on the same item domain.

Table 2
Varimax Orthogonally Rotated Factor Matrix for Measures of Gender-Related Individual Differences—Twins A, Male

	Factor Labels		
	Gender Diagnosticity Factor	Masculine Instrumentality Factor	Feminine Expressiveness Factor
Variable			
GD Occupation	.69		
GD Activities	.65		
GD CPI items	.84		
CPI femininity	–.84		
ACL feminine expres (raw)			.76
ACL feminine expres (norml)			.80
CPI Good Impression			.48
CPI Socialization			.52
ACL masculine instrum (raw)		.92	
ACL masculine instrum (norml)		.89	
CPI Dominance		.68	

Note. Factor loadings < .35 have been omitted.

loaded highly on one factor labeled "Gender Diagnosticity." ACL measures of feminine expressiveness, CPI Good Impression, and CPI Socialization all loaded substantially on a second factor labeled "Feminine Expressiveness." Finally, ACL measures of masculine instrumentality and CPI Dominance all loaded substantially on a third factor labeled "Masculine Instrumentality." Factor scores were computed for all twins on these three factors.

To provide an estimate of the reliability of factor scores, the component measures of each factor were converted to *z*-scores, and coefficient alpha was computed for the sum of the component scales. For example, the four components of the Gender Diagnosticity factor (GD based on occupation, GD based on activities, GD based on CPI items, and CPI femininity) were converted to *z*-scores, and coefficient alpha was computed for the composite of the four. The reliabilities of the GD composite for male Twins A, male Twins B, female Twins A, and female Twins B were .74, .76, .70, and .68, respectively. Corresponding reliabilities for the Masculine Instrumentality composite were .80, .78, .79, and .79. Finally, corresponding reliabilities for the Feminine Expressiveness composite were .57, .50, .53, and .55. The lower reliabilities for Feminine Expressiveness suggest that the scales that made up this composite were less cohesive than those that made up the GD and Masculine Instrumentality composites. An examination of loadings on the Feminine Expressiveness factor in Table 2 shows that the CPI Good Impression and Socialization scales loaded more weakly on this factor (.48 and .52) than did the ACL expressiveness scales (.76 and .80). Thus, although these two CPI scales partially overlap with Feminine Expressiveness, they seem to assess other content as well. Still, we decided to maintain these two CPI scales as components of Feminine Expressiveness because their loadings on the factor were substantial and our reliability analyses showed that reliability was not increased by deleting either of these component scales from the composite.

Behavior Genetic Analysis

Intraclass correlations for masculine instrumentality, feminine expressiveness, and gender diagnostic components. Table 3 shows intraclass correlations for monozygotic and dizygotic twins for our three composite measures of gender-related individual differences, computed separately for male and female twins. The fact that all correlations for monozygotic twins in Table 3 exceed corresponding correlations for dizygotic twins suggests that there are significant genetic effects for each of the three measures. These correlations provide a descriptive account of our findings, but the application and testing of formal behavior genetic models was necessary to answer quantitatively the specific research questions posed earlier.

Behavior genetic model-fitting analyses. Univariate behavior genetic model-fitting analyses were conducted on the covariance and variance matrices associated with the twin intraclass correlations presented in Table 3. A model consisting of additive genetic (a), shared environmental (c), and nonshared environmental parameters was first fit to the data, equating the parameter estimates for men and women. The results are shown in Table 4. None of the models equating the sexes was rejected.

Table 3
Intraclass Correlations for Identical and Fraternal Twins for Masculine Instrumentality, Feminine Expressiveness, and Gender Diagnosticity Factors

	Correlations for Masculine Instrumentality	
	MZ Twins	DZ Twins
Males	.34	.11
	(*n* = 198)	(*n* = 122)
Females	.36	.19
	(*n* = 288)	(*n* = 192)
	Correlations for Feminine Expressiveness	
	MZ Twins	DZ Twins
Males	.35	.18
	(*n* = 198)	(*n* = 122)
Females	.39	.26
	(*n* = 288)	(*n* = 192)
	Correlations for Gender Diagnosticity	
	MZ Twins	DZ Twins
Males	.53	.35
	(*n* = 198)	(*n* = 122)
Females	.52	.32
	(*n* = 288)	(*n* = 192)

In addition, shared environmental effects were nonsignificant, as shown by chi-square difference tests for Masculine Instrumentality, Feminine Expressiveness, and Gender Diagnosticity (chi-square differences = 0.00, 0.47, and 1.77, respectively). Conversely, removing additive genetic effects resulted in a significant decrement in fit for all three variables: chi-square differences = 8.55, 5.35, 13.38, respectively. Therefore, all three variables may be described by a simple model comprising additive genetic and nonshared environmental effects.

To test the equality of additive and nonshared environmental effects across the three variables, a model was fit equating these parameters across the variables. This model was accepted; chi-square = 24.41, *df* = 22, *p* = .33. Nonetheless, examination of Table 3 indicates that the heritability for gender diagnosticity appears larger than the heritabilities for Masculine Instrumentality or Feminine Expressiveness. Revising the

Table 4
Standardized Maximum Likelihood Model-Fitting Results

Variable	a^2	t-value	e^2	t-value	χ^2	p
Masculine Instrumentality	.36	16.66	.64	34.11	1.35	.97
Feminine Expressiveness	.38	17.72	.62	33.96	1.61	.95
Gender Diagnosticity	.53	25.02	.47	33.00	2.09	.91

Note. All parameter estimates, $p < .001$. All models have $df = 6$.
a = additive genetic effects; e = nonshared environmental effects.

model to allow the heritability of Gender Diagnosticity not to equal the heritabilities of the other two variables resulted in a significant improvement in fit: chi-square difference = 18.98, $df = 2$, $p < .001$. Therefore, it may be concluded that the heritability of Gender Diagnosticity is significantly greater than the heritabilities of Masculine Instrumentality and Feminine Expressiveness.

The variables comprising the Gender Diagnosticity factor were also subjected to univariate behavior genetic analyses. The twin intraclass correlations for these variables are shown in Table 5, and the model-fitting results in Table 6. Sex differences were found for only one variable, CPI Femininity. For the other three variables, sex differences were not found. For all three of these variables, heritability was significant. For two of the variables not showing sex differences—GD activities and GD CPI items—shared environmental effects were also significant.

As just noted, the initial model equating the sexes on CPI femininity was rejected, chi-square = 20.02, $df = 5$, $p < .001$, whereas a model allowing the sexes to vary provided a good fit to the data: chi-square = 1.56, $df = 2$, $p = .46$. Further model exploration revealed that equating the sexes for additive genetic and nonshared environmental effects while allowing for shared environmental effects for men only provided the most parsimonious fit to the data: chi-square = 4.19, $df = 5$, $p = .52$. Although additive genetic and nonshared environmental effects were equated in the final model, the process of standardization makes the parameters appear to have different values across the sexes (e.g., $a^2 = .21$ for men, .27 for women). The reason for this lies with the retention of shared environmental effects for men only, and the significant difference in the variance of CPI femininity for men ($s^2 = 3.74$) and women ($s^2 = 3.19$): F (651, 959) = 1.38, $p < .001$.

Table 5
Twin Intraclass Correlations for Gender Diagnosticity (GD) Variables

Variable	MZM	*N*	DZM	*N*	MZF	*N*	DZF	*N*
GD Occupation	.44	210	.22	131	.40	293	.26	194
GD Activities	.59	215	.46	135	.59	293	.40	195
GD CPI items	.45	202	.34	124	.47	288	.34	193
CPI Femininity	.41	202	.26	124	.30	288	.14	192

Note. MZM = male monozygotic twins; DZM = male dizygotic twins; MZF = female monozygotic twins; DZF = female dizygotic twins.

Table 6
Standardized Maximum Likelihood Model-Fitting Results for Gender Diagnosticity (GD)

Variable	a^2	*t*-value	c^2	*t*-value	e^2	*t*-value	χ^2	*df*	*p*
GD Occupation	.44	20.36	—	—	.66	34.15	8.04	6	.24
GD Activities	.21	4.77	.32	7.69	.47	32.29	8.08	5	.15
GD CPI items	.28	4.42	.18	4.22	.54	32.02	.51	5	.99
CPI Femininity									
males	.21[a]	10.75	.23	7.35	.66[b]	34.61			
females	.27[a]	10.75	—	—	.63[b]	34.61			
							4.19	5	.52

Note. All parameter estimates, $p < .001$.
[a,b] Parameters with the same superscript have been equated during model-fitting.
a = additive genetic effects; *c* = shared environmental effects; *e* = nonshared environmental effects.

DISCUSSION

We began with four questions: (1) Do genetic factors contribute to individual differences in M, F, and GD? (2) Do environmental factors also contribute to these individual differences, and if so, do they represent shared or nonshared environmental effects? (3) Are the relative contributions of genetic and environmental factors to gender-related individual differences the same for males and females? And finally, (4) Do the relative contributions of environmental and genetic factors differ for various measures of gender-related individual differences (i.e., for M–F, M, F, and GD)?

Based on our behavior genetic modeling analyses of measures computed from Loehlin and Nichols's (1976) classic twin data, we offer the following answers: (1) Genetic factors contribute significantly to individual differences in each of the three measures studied: M, F, and GD. (2) The environmental effects on M, F, and GD tend to be nonshared. (3) The genetic and environmental components of individual differences in M, F, and GD generally do not show gender differences. (The only exception to this general conclusion was evidence for significant shared environmental effects in CPI Femininity for males, but not for females.) Finally, (4) the estimated heritability of GD proved to be significantly greater than the heritabilities of either M or F.

Conclusions (1) and (2) are consistent with the findings of much recent behavior genetic research on personality, which we reviewed earlier. That is, personality traits generally show significant heritability, and shared environmental effects for most personality traits tend to be small. Although occasional results have suggested gender differences in the genetic expression of some traits (see Rowe, 1994, Chapter 6, for a discussion), the balance of evidence does not favor gender differences.

Perhaps the most intriguing result of the current research was that GD showed higher heritability than either M or F.[7] The narrow heritability

[7] As noted earlier in this article, CPI scales (Socialization, Good Impression, and Dominance) were imperfect proxies for F and M, and one of these (Good Impression) did not show gender differences in the studied population. Furthermore, reviewers of this article noted that some of the measures we factor-analyzed were redundant (e.g., GD based on CPI items and CPI Femininity, and normalized and raw ACL scales of M and F). To compute the purest possible factors scores for GD, M, and F, we conducted additional factor analyses on the three GD measures (based on occupational preferences, everyday activities, and CPI items), normalized ACL M, and normalized ACL F. As before, these analyses were performed separately for male Twins A, male Twins B, female Twins A, and female Twins B. Consistent with previous results, these analyses showed a pure three-factor structure, with the three GD measures loading highly on one factor, normalized M on the second, and normalized F on the third.

Twin variances and covariances were computed for these new factor scores, and univariate behavior genetic model-fitting analyses were conducted. A model consisting of additive genetic effects (*a*), shared environmental effects (*c*), and nonshared environmental effects (*e*) was first fit to the data, equating parameter estimates for men and women. None of the models equating the sexes was rejected. For each factor a simple model consisting of additive genetic and nonshared environmental effects provided the most parsimonious and best-fitting representation of the data (all chi-squares *ns*). Estimates of additive genetic effects (a^2) for GD, M, and F were respectively .59, .14, and .14, and estimates of nonshared environmental effects (e^2) were respectively .41, .86, and .86.

estimate for the GD factor was .53, which is high for the personality domain. As a basis of comparison, we can look to the estimated heritabilities of Big Five traits. Loehlin (1992) summarized evidence showing that additive genetic effects for the Big Five range from .28 (for Agreeableness and Conscientiousness) to .46 (for Openness). After analyzing just twin-family studies, Loehlin and Rowe (1992) estimated the additive genetic effects for the Big Five to be .22 for Openness, .27 for Emotional Stability (or Neuroticism), .29 for Agreeableness, .32 for Extraversion, and .43 for Openness.

Why did GD show higher heritability than did M or F in the current study (and higher heritability than is often demonstrated for Big Five traits)? The current data cannot answer this question, and our answer must perforce be speculative. GD measures—unlike M or F, or indeed, unlike any of the Big Five—are based on a biologically based grouping, that is, male versus female.[8] Extending the logic of GD measures (that sex differences serve to define gender-related individual differences within the sexes) to hypothetical genetic and biological processes that influence sexual differentiation, it is possible that the same biological processes that lead to sexual differentiation (or more precisely, variations in these processes) also influence gender-related individual differences *within the sexes*. Because the molecular genetic and biological mechanisms that lead to sexual differentiation may be easier to isolate and study than the mechanisms that influence Big Five traits (see Hoyenga & Hoyenga, 1993, for a broad review of the biology of sex and gender, and Zuckerman, 1995, for a recent discussion of the biology of personality), gender-related individual differences might provide an interesting "laboratory" in which to study genetic and biological influences on personality.

To test the equality of additive genetic effects across the three factors, a model was fit equating this parameter across the three factors. This model was rejected; chi-square = 106.19, *df* = 20, *p* = .00. However, a model equating additive genetic effects for M and F but permitting a different value for GD fit the data well; chi-square = 19.04, *df* = 19, *p* = .45. Once again, it may be concluded that the heritability of GD is greater than that of M and F (indeed, in this analysis more than four times greater). Thus, eliminating CPI scales from our analyses and using nonredundant M and F scales produced results consistent with, and even stronger than, the results reported in the main body of this article.

[8] It goes without saying that "male" and "female" are socially defined categories as well.

Regardless of the proper theoretical account for the high heritability of GD, it is important to emphasize the following point: The current findings do not speak at all to the issue of the heritability of gender differences, per se. Rather, they focus on the heritability of gender-related individual differences *within the sexes*. The current results indicate that the *within-sex* heritability of GD is higher than the *within-sex* heritability of M or F.

Much previous research has shown that GD is factorially distinct from M and F and that GD correlates with various criteria differently than do M and F (see Lippa, 1991, 1995b, 1996a, 1996b, 1997, 1998a, 1998b; Lippa & Arad, 1997; Lippa & Connelly, 1990). Interpreted in the broadest sense, then, the current findings indicate once again that GD is "different from" M and F—this time in relation to within-sex heritability.

GD proved to have higher heritability than most Big Five traits in the current analyses. In contrast, M and F had heritability estimates that were roughly comparable with heritability estimates for the Big Five traits in general, and for Extraversion and Agreeableness in particular. The current heritability estimate for F proved to be somewhat higher than previous estimates for Agreeableness, and furthermore, we did not find evidence for shared environmental effects for F (which might be predicted based on some of the research reviewed earlier). However, given that the F factor computed in the current study was less reliable than the M or GD factors, the behavior genetic parameter estimates for F should probably be regarded as the most tentative of our findings.

Our behavior genetic analyses of component GD measures (see Table 5) indicated that genetic effects were strongest for GD based on occupational preferences and weakest for GD based on activities.[9] At the same time, GD based on occupational preferences showed no evidence for shared environmental effects, whereas GD based on CPI items and on self-reported activities did show evidence for significant shared

[9] This difference in heritability may have been due in part to the fact that GD measures based on occupational preferences had greater reliability than the other two GD measures.

Similarly, it may seem paradoxical that our estimate of additive heritability for the GD factor was greater than the estimated heritabilities of GD factor components. However, this result is not unexpected, since factor scores are often more reliable than their components, and in general, more reliable variables yield higher estimates of heritability. Given that the M factor was the most reliable of the three assessed factors (M, F, and GD), it is all the more noteworthy that the heritability of the GD factor proved to be greater than that of the M factor.

environmental effects, with these effects greatest for GD based on activities.

One possible explanation for these varying environmental effects across GD measures is the following conjecture: GD based on everyday activities was based on self-reports of *actual behavior*, whereas GD based on occupations was based on self-reported behavioral *preferences*. It seems likely that the "press" of family environments on high-school-aged participants' actual behaviors (e.g., "say grace before meals," "go skiing," "attend athletic events," "fly in an airplane") might be greater than on their behavioral *preferences*. GD based on CPI items (and also CPI Femininity) showed shared environmental effects that were intermediate between those found for GD based on activities and GD based on occupations, and consistent with the previous hypothesis, CPI Femininity items tap both actual behaviors ("I become quite irritated when I see someone spit on the sidewalk," "At times I feel like picking a fist fight with someone") as well as behavioral preferences ("I think I would like the work of a building contractor," "I think I would like the work of a dress designer"). Loehlin and Nichols (1976, p. 91) noted in their original analyses of the National Merit twin data that activities showed evidence for substantial shared environmental effects, whereas vocational interests did not. The current findings replicate these results specifically for gender-related activities and gender-related vocational interests. In addition to viewing this finding as reflecting item domain differences (e.g., activities vs. vocational interests), we speculate that they may also reflect differences in the kind of self-report obtained (reports of actual behavior vs. behavioral preferences). At the very least, the current findings suggest that further behavior genetic research is warranted on this "actual behavior versus behavioral preference" distinction.

In sum, the current findings suggest that although within-sex individual differences in M, F, and GD all show evidence for significant additive genetic effects, these genetic effects are larger for GD than for M or F. It may seem paradoxical that, although the overall GD factor showed evidence of high heritability, some component GD measures simultaneously showed evidence of significant shared environmental effects. Of course, these findings are not necessarily mutually exclusive. In both of these regards, GD measures showed differences from Big Five traits and their close relatives, Masculine Instrumentality and Feminine Expressiveness. Clearly, additional behavior genetic analyses of GD measures are warranted to replicate, extend, and further explain the current findings.

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 Consulting and Clinical Psychology*, **42**, 165–172.

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

Bem, S. L. (1981b). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, **88**, 354–364.

Bem, S. L. (1985). Androgyny and gender schema theory: A conceptual and empirical integration. In T. B. Sonderegger (Ed.), *Psychology and gender: Nebraska Symposium on Motivation, 1984* (pp. 179–226). Lincoln: University of Nebraska Press.

Bem, S. L. (1993). *The lenses of gender: Transforming the debate on sexual inequality*. New Haven, CT: Yale University Press.

Block, J. H. (1973). Conceptions of sex roles: Some cross-cultural and longitudinal perspectives. *American Psychologist*, **28**, 512–526.

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

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

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1995). Solid ground in the wetlands of personality: A reply to Block. *Psychological Bulletin*, **117**, 216–220.

Falconer, D. S. (1960). *Introduction to quantitative genetics*. New York: Ronald Press.

Goldberg, L. (1993). The structure of phenotypic personality traits. *American Psychologist*, **48**, 26–34.

Gough, H. B. (1957). *CPI manual*. Palo Alto, CA: Consulting Psychologists Press.

Gough, H. B. (1987). *CPI administrator's guide*. Palo Alto, CA: Consulting Psychologists Press.

Helgeson, V. S. (1995). Prototypes and dimensions of masculinity and femininity. *Sex Roles*, **31**, 653–683.

Holland, J. L. (1958). A personality inventory employing occupational titles. *Journal of Applied Psychology*, **42**, 336-342.

Hoyenga, K. B., & Hoyenga, K. T. (1993). *Gender-related difference: Origins and outcomes*. Boston: Allyn & Bacon.

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., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, **61**, 521–551.

Lenney, E. (1991). Sex roles: The measurement of masculinity, femininity, and androgyny. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 573–660). San Diego: Academic Press.

Lippa, R. (1991). Some psychometric characteristics of gender diagnosticity measures: Reliability, validity, consistency across domains, and relationship to the Big Five. *Journal of Personality and Social Psychology*, **61**, 1000–1011.

Lippa, R. (1995a). Do sex differences define gender-related individual differences within the sexes? Evidence from three studies. *Personality and Social Psychology Bulletin*, **21**, 349–355.

Lippa, R. (1995b). Gender-related individual differences and psychological adjustment in terms of the Big Five and circumplex models. *Journal of Personality and Social Psychology*, **69**, 1184–1202.

Lippa, R. (1997). The display of masculinity, femininity, and gender diagnosticity in self-descriptive photo essays. *Journal of Personality*, **65**, 137–169.

Lippa, R. (1998a). Gender-related individual differences and the structure of vocational interests: The important of the "People-Things" dimension. *Journal of Personality and Social Psychology*, **74**, 996–1009.

Lippa, R. (1998b). Gender-related individual differences and National Merit Test performance: Girls who are "masculine" and boys who are "feminine" tend to do better. In L. Ellis & L. Ebertz (Eds.), *Males, females, and behavior: Toward biological understanding*. New York: Praeger.

Lippa, R. (1998c). The nonverbal display and judgment of masculinity, femininity, and gender diagnosticity: A lens model analysis. *Journal of Research in Personality*, **32**, 80–107.

Lippa, R., & Arad, S. (1997). The structure of sexual orientation and its relation to masculinity, femininity, and gender diagnosticity: Different for men and women. *Sex Roles*, **37**, 187–208.

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

Loehlin, J. C. (1992). *Genes and environment in personality development*. Newbury Park, CA: Sage.

Loehlin, J. C., & Nichols, R. C. (1976). *Heredity, environment, and personality*. Austin: University of Texas Press.

Loehlin, J. C., & Rowe, D. C. (1992). Genes, environment, and personality. In G. Caprara & G. L. Van Heck (Eds.), *Modern personality psychology: Critical reviews and new directions* (pp. 352–370). New York: Harvester Wheatsheaf.

McCartney, K., Harris, M. J., & Bernieri, F. (1990). Growing up and growing apart: A developmental meta-analysis of twin studies. *Psychological Bulletin*, **107**, 226–237.

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

Mitchell, J. E., Baker, L. A., & Jacklin, C. N. (1989). Masculinity and femininity in twin children: Genetic and environmental factors. *Child Development*, **60**, 1475–1485.

Plomin, R., Chipuer, H. M., & Loehlin, J. C. (1990). Behavioral genetics and personality. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 225–243). New York: Guilford Press.

Rose, R. J. (1995). Genes and human behavior. *Annual Review of Psychology*, **46**, 625–654.

Rowe, D. C. (1982). Sources of variability in sex-linked personality attributes: A twin study. *Developmental Psychology*, **18**, 431–434.

Rowe, D. C. (1994). *The limits of family influence: Genes, experience, and behavior*. New York: Guilford Press.

Saudino, K. J., & Plomin, R. (1996). Personality and behavior genetics: Where have we been and where are we going? *Journal of Research in Personality,* **30**, 335–347.

Spence, J. T. (1985). Gender identity and its implications for the concepts of masculinity and femininity. In T. B. Sonderegger (Ed.), *Psychology and gender: Nebraska Symposium on Motivation, 1984* (pp. 59–95). Lincoln: University of Nebraska Press.

Spence, J. T. (1993). Gender-related traits and gender ideology: Evidence for a multifactorial theory. *Journal of Personality and Social Psychology*, **64**, 624–635.

Spence, J. T., & Buckner, C. (1995). Masculinity and femininity: Defining the undefinable. In P. J. Kalbfleisch & M. J. Cody (Eds.), *Gender, power, and communication in human relationships* (pp. 105–138). Hillsdale, NJ: Erlbaum.

Spence, J. T., & Helmreich, R. L. (1978). *Masculinity and Femininity: Their psychological dimensions, correlates, and antecedents*. Austin: 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. L., & Stapp, J. (1974). The Personal Attributes Questionnaire: A measure of sex role stereotypes and masculinity-femininity. *JSAS,* Catalog of Selected Documents in Psychology, *4*, 43-44 (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. X. Wallston (Eds.), *Women, gender, and social psychology* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Tellegen, A. (1982). *Brief manual for the Differential Personality Questionnaire.* Unpublished manuscript, University of Minnesota, Minneapolis.

Tellegen, A., Lykken, D. T., Bouchard, T. J. Jr., Wilcox, K. J., Segal, N. L., & Rich, S. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology*, **54**, 1031–1039.

Tellegen, A., & Waller, N. G. (In press). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1). Greenwich, CT: JAI Press.

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

Wiggins, J. S. (1991). Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. In W. M. Grove & D. Cicchetti (Eds.), *Thinking clearly about psychology: Vol. 2. Personality and psychopathology*. Minneapolis: University of Minnesota Press.

Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multination study*. Newbury Park, CA: Sage.

Zuckerman, M. (1995). Good and bad humors: Biochemical bases of personality and its disorders. *Psychological Science,* **6**, 325–332.