Psychological Masculinity-Femininity via the Gender Diagnosticity Approach: Heritability and Consistency Across Ages and Populations

John C. Loehlin
University of Texas at Austin
Erik G. Jönsson and J. Petter Gustavsson
Karolinska Institute, Stockholm, Sweden
Michael C. Stallings
University of Colorado, Boulder
Nathan A Gillespie, Margaret J. Wright, and Nicholas G. Martin
Queensland Institute of Medical Research, Brisbane, Australia

Jönsson and Gustavsson are from the Department of Clinical Neuroscience of the Karolinska Institute. Jönsson was supported by the HUBIN project. We thank our original collaborators in the studies from which this article derives: A. Spurdle, S. A. Treloar, S. E. Medland, and G. W. Montgomery in Australia; M. Schalling, C. von Gertten, Q.-P. Yuan, K. Lindblad-Toh, K. Forslund, G. Rylander, M. Mattila-Evenden, and M. Åsberg, in Sweden, and A. C. Heath, J. K. Hewitt, M. C. Neale, L. J. Eaves, T. Beresford, R. Cates, and J. Meyer in the United States. We are grateful to Richard Lippa for his comments on an earlier version of this article.

Correspondence should be addressed to J. C. Loehlin, The University of Texas at Austin, Psychology Department, 1 University Station A8000, Austin, TX 78712-0187. E-mail: loehlin@psy.utexas.edu.

Journal of Personality 73:5, October 2005

DOI: 10.1111/j.1467-6494.2005.00350.x

ABSTRACT Several aspects of the Gender Diagnosticity (GD) approach of Lippa (1995) to measuring the psychological trait of masculinity-femininity within sexes were explored in four samples ranging from 363 to 5,859 individuals, including Swedish and Australian adults, U.S. elderly, and Australian adolescents. Two ways of deriving GD scales yielded highly similar results. Moderate stability of individual differences was found across ages 12 to 16 among adolescents, but substantial shifts over age occurred in relationships with Eysenck scales. Considerable generality of GD scales was obtained across languages and populations. Substantial heritabilities (about 40%) and minimal effects of shared family environments suggest that within-sex masculinity-femininity behaves as a fairly typical personality trait. Cross-age continuity appeared mainly to reflect the influence of the genes.

There has been a longstanding interest among psychologists in how males may differ in attitudes and behavior from females and in how these differences may translate into a masculinity-femininity trait (or traits) within the two sexes (e.g., Bem, 1974; Spence & Helmreich, 1978; Terman & Miles, 1936). Sociologists, anthropologists, historians, and others have joined psychologists in asking questions about how these differences may vary across ages, eras, and cultures (e.g., Hofstede, 1998; Meade & Weisner-Hanks, 2004; Williams & Best, 1990).

There is a consensus among writers in this area that within any population, male-female differences on psychological traits are, at most, differences in averages: on any such trait there tends to be much variation within each sex and considerable overlap between the two sexes. This fact of within-sex variation in traits that (on average) distinguish the sexes has led to interest in masculinity-femininity (MF) as a psychological dimension within each sex, and, in turn, to questions about how differences on such a dimension might develop over age and differ across cultures and to what extent they reflect underlying biological factors such as level of sex hormones or sensitivity to them, and so forth.

A few years ago, some of us (Loehlin, Spurdle, Treloar, & Martin, 1999; Jönsson, et al., 2001) became interested in one such possible underlying biological factor, the length of a particular repeated sequence in the X-linked androgen receptor gene. This length varies over a range of 11 to 31 repeats in normal populations (Edwards, Hammond, Jin, Caskey, & Chakraborty, 1992), affects the transcription of testosterone (Chamberlain, Driver, & Miesfeld, 1994),

and thus might plausibly be related to within-sex MF. Given available data from samples in Australia and Sweden, a joint study required a measure of MF that would work across two countries in which different personality questionnaires in different languages were used. An initial effort to link the two via a U.S. sample of elderly twins who had been given versions of the questionnaires used in the Swedish and Australian studies was not notably successful (Loehlin, Jönsson, Gustavsson, Schalling, & Stallings, 2003).

One recent perspective on psychological MF, that of gender diagnosticity (Lippa, 1995; Lippa & Connelly, 1990), suggests that we might have been going about this in the wrong way. Instead of assuming that the trait MF should be measured by similar items across samples, Lippa says that we should assess MF within samples in terms of response to those items that *in that sample* discriminate males from females. Thus gender diagnosticity (GD) is not a trait with fixed content, but one that may vary with age, era, or population.

In addition to examining relationships with the androgen receptor gene, a study assessing GD with overlapping item sets in adult samples of differing ages in different countries permitted the evaluation of several aspects of the GD approach itself, such as the comparability of GD across languages, ages, and populations. Because the Australian and U.S. elderly samples consisted of twins, estimates of the heritability of GD in these populations could also be made.

Subsequent to the Australian and Swedish studies mentioned above and the U.S. study linking them, similar data became available for samples of Australian adolescent twins aged 12, 14, and 16 years, along with some of their siblings (Wright & Martin, 2004). Many of these adolescents repeated a personality questionnaire at two or more of these ages, so longitudinal data on GD across these years were available as well. Given that the period of adolescence is important in the development of adult masculinity and femininity, data spanning this age range are of considerable interest. These data provided the possibility of addressing the comparability of GD measures across ages in the short term from age 12 to 16, and, more broadly, between adolescents and adults. Because the adolescent samples contained both monozygotic (MZ) and dizygotic (DZ) twins, heritability estimates could be made for these ages and compared to estimates for the Australian adults and the U.S. elderly, as well as to the U.S. adolescent data of Lippa and Hershberger (1999) and Cleveland, Udry, and Chantala (2001) on GD measures derived

from different questionnaires. Finally, the presence of repeated measures for many of the adolescent twins permitted longitudinal behavior-genetic analysis, as well, to ascertain whether cross-age continuity was primarily genetic or environmental.

The question of how gender diagnosticity is related to the androgen receptor gene is dealt with in a separate article (Loehlin et al., 2004). Suffice it to say here that matters turned out not to be simple—a weak but consistent relationship was observed for adults, but not for the adolescents. However, the present paper focuses not on relationships with this gene but on the properties of the GD measure itself, including alternative approaches to assessing it, its heritability, and its generalizability across ages and populations.

METHOD

Samples

The samples and data sets are briefly described in Table 1. The samples consisted of Swedish adults (mean age 43.1 years) from the Stockholm area tested in person, elderly twins from the United States (mean age 66.7 years) tested by mail, adult Australian twins (mean age 42.2 years), also tested by mail, and Australian adolescents. In the last named study, tests were administered in person to the twins as close to the target ages of 12, 14, and 16 as was feasible, as well as to a number of their siblings, who were included in the present analyses if they were within a year of the target age when tested—i.e., at the first, second, or third testings between 11 and 13, between 13 and 15, and between 15 and 17, respectively.

Questionnaires

The Swedish participants were administered the 135-item Karolinska Scales of Personality (KSP; Schalling, Åsberg, Edman, & Oreland, 1987). The adult Australian twins received a 56-item version of the Eysenck Personality Questionnaire (EPQ; Eysenck, Eysenck, & Barrett, 1985) and a 54-item version of Cloninger's Tridimensional Personality Questionnaire (TPQ; Cloninger, Przybeck, & Svrakic, 1991). The Australian adolescent twins received the 81-item Junior Eysenck Personality Scale (JEPQ; Eysenck & Eysenck, 1975). Each inventory contains scales purporting to measure several personality or temperament dimensions, but we treat them here as collections of items, some more often endorsed by males than females or vice versa, from which MF scales may be derived. The U.S. elderly sample had been administered a questionnaire that

Table 1
Summary of Data Sets in Four Samples

Attribute	Swedish adults	U.S. elderly	Australian adults	Australian adolescents
Sample size				
Male	203	1,029	2,022	1,048
Female	160	2,961	3,837	1,040
Same sex twin pairs				
MZ	—	735	1,320	408
DZ	—	338	748	308
Questionnaire items				
KSP	135	60	—	—
EPQ	—	54	56	—
TPQ	—	100	54	—
JEPQ	—	—	—	81

Note. MZ = monozygotic, DZ = dizygotic; KSP = Karolinska Scales of Personality, EPQ = Eysenck Personality Questionnaire, TPQ = Cloninger's Tridimensional Personality Questionnaire, JEPQ = Junior Eysenck Personality Questionnaire. Number of twin pairs is less than half overall sample size because total samples can include unpaired, nondiagnosed, or unlike-sex twins, and, for adolescents, non-twin siblings. The four samples are described in Jönsson, et al. (2001); Loehlin, et al. (1999); Stallings, et al. (1999); and Wright and Martin (2004).

included 60 of the KSP items (translated into English), 54 of the EPQ items, and a 100-item version of the TPQ that included the 54 items used in Australia. We were therefore able in each of the three adult samples to derive a full-length GD measure based on all the items in that sample, as well as one or two shorter GD measures based on just those items shared with another sample. Thus, each of the shorter GD measures can be scored in the sample in which it was derived and in another sample that responded to the same items (possibly in a different language).

In the Australian adolescent sample, separate GD scales were derived at each of the three ages. For comparisons across ages, a combined GD scale was used consisting of items that were common to all three of the separate age scales. Also, for control purposes, a non-GD scale was derived consisting of items that did *not* differentiate the sexes at any of the three ages.

Two Ways of Deriving GD Scales

Lippa has most often based his GD measures on occupational or avocational preferences, but he has also employed personality questionnaire

items and self-reports of specific behaviors (e.g., Lippa & Hershberger, 1999). In principle, any set of items for which males and females have differing endorsement frequencies can be used to derive a GD measure of within-sex MF—although, of course, the degree of relationship among GD scales based on different content domains remains an empirical question.

In deriving GD scales, Lippa has sometimes used the traditional approach, dating back to Terman and Miles (1936), of constructing a scale of items differentially endorsed by the two sexes and scoring an individual by the extent to which his or her endorsements concur with those favored by a particular sex. We will refer to this as the contrasted-groups method. However, Lippa's currently preferred method is to carry out two-group discriminant analyses on subsets of items, using males and females as the criterion groups, and for each such subset, to assign to each individual in the sample the probability (as calculated by the discriminant analysis program) that someone responding in this way is male or female. These probabilities are then averaged over all the item subsets to constitute the individual's GD score. Our first step will be to derive scales using the discriminant and contrasted-groups methods and compare them.

The procedure used for deriving discriminant-based GD scales followed Lippa's. Each item set was divided into 10- or 11-item subsets (separately for different questionnaires); thus, five discriminant analyses were carried out for the EPQ items in the Australian sample, and five for the TPQ items. The probability that an individual would be diagnosed as male or female was calculated for each of the 10 analyses by the program used (SPSS Discriminant, SPSS 1990). These 10 probabilities were then averaged to yield an overall GD score for the individual. A similar procedure was used with the KSP items in the Swedish sample, again with approximately 10- or 11-item subsets, and likewise for the items from the three scales in the U.S. elderly sample and the JEPQ items in the Australian adolescent subsamples.

For the contrasted-groups procedure, all items were correlated with sex, and those with correlations above a threshold were selected to form the scale. For the Swedish sample, the threshold used was an absolute correlation of .135, which corresponded to the .01 level of significance for a sample size of 363. For the other three samples, an arbitrary threshold correlation of .100 was used for item selection. With the large samples involved, smaller correlations than this would be nominally statistically significant, but not very useful for a scale. A threshold of .100 yielded a reasonable number of items in all samples.

Appendix Table A1, included for archival purposes, contains the assignment of items to GD scales (and the non-GD scale) for all scales constructed by the contrasted-groups method. GD scores by the discriminant method do not involve item selection. They employ all the items in

Table 2
Within-Sex Correlation Between Discriminant and Contrasted-Groups Scales in Three Adult and Three Adolescent Samples

Sample	Males		Females	
Swedish adults	.86	(203)	.86	(159)
U.S. elderly	.77	(1,002)	.83	(2,785)
Australian adults	.94	(1,973)	.93	(3,699)
Australian age 12	.95	(819)	.93	(799)
Australian age 14	.93	(647)	.93	(621)
Australian age 16	.95	(599)	.95	(634)

Note. *N*s in parentheses.

each pool via weights assigned within item subsets to predict the respondent's sex.

Table 2 compares the two forms of scale derivation. The correlations of above .90 in the four Australian samples suggest that the methods are yielding essentially equivalent results in these samples. The correlations for the Swedish and U.S. samples are a little lower, in the range .77 to .86, but are large enough to suggest that we may regard the two versions as equivalent for many purposes. All these correlations are considerably higher than those originally reported by Lippa (1991). He obtained average correlations of .40 for men and .47 for women between scales derived by these two methods, using items dealing with occupational preferences, school subjects, activities, and amusements. In more recent publications, Lippa reports much higher correlations between the two types of scales—.93 to .96 for scales based on occupational preferences in one recent paper (Lippa, 2002).

Thus, we will consider the two versions as essentially interchangeable methods of measuring Gender Diagnosticity, using one or the other as convenient. For the basic heritability analyses, we will use the discriminant-derived version as that is most similar to the methods used in the studies to which we wish to compare our results. For the across-population analysis, we will use the contrasted-groups version, because of its portability across samples. For the adolescent age comparisons, we will use contrasted-groups versions, either age-specific or common as appropriate.

In retrospect, the near-equivalence of these two approaches is perhaps not too surprising. In effect, the contrasted-groups approach assigns weights of +1, 0, and −1 to items based on their ability to discriminate the sexes. The discriminant-derived version assigns continuously varying weights. It is well known to psychometrists that unit item weights give

about the same results in most situations as differential item weighting schemes. In a classic paper, Wainer (1976) noted that as long as you have the scoring direction right, "it don't make no nevermind" which item weighting scheme you use.

Reliabilities

Within-sex internal consistency reliabilities were calculated for both the contrasted-groups and discriminant-derived measures. The former were based on items, the latter on subscales. The reliabilities for adults were moderate, ranging from .68 to .88 in six samples (males and females in Sweden, Australia, and the United States). Those for the Australian adolescents ran a little lower, in the range .57 to .72. The reliabilities for adults are comparable to those reported for college students by Lippa (1991). For a discriminant-derived GD scale based on preferences for 131 occupations, he obtained alpha reliabilities of .76 and .78 for men and women, respectively, and for a contrasted-groups scale, reliabilities of .78 and .83. Both were based on 119 men and 145 women.

Because the above reliabilities were obtained in the same samples in which the scales were derived, one might wonder if they were thereby inflated. Actually, this is unlikely to be the case—indeed, there may well be a bias in the opposite direction. If an item were perfectly to discriminate between males and females, its contribution to within-groups reliability would necessarily be nil. This bias is, however, probably not a serious one. The absolute correlation of individual scale items with sex was rarely as high as .30, placing relatively little constraint on within-sex variation. (The highest item correlation with sex was .34 in the Swedish sample, .24 in the U.S. elderly sample, .25 in the Australian adult sample, and .25, .30, and .29 for the Australian 12-, 14-, and 16-year-olds, respectively.)

RESULTS

Generality Across Populations

Our data permit examining the extent to which GD scales generalize across populations. This is a central issue for GD. The philosophy underlying this approach is that GD scales are inherently population-specific: that which best defines individual differences in MF within each sex in a given population is that which best distinguishes the two sexes in that population.

Table 3 gives within-sex correlations separately for men and women for four cases in which two scales are compared, one derived

Table 3
Correlation in a Given Sample Between a GD Scale Derived in That Sample and a GD Scale Developed in Another Sample From the Same Item Set

Scale	Keyed items	Men		Women	
KSP					
In Swedish sample	22	.97	(203)	.96	(160)
In U.S. sample	18	.97	(1,032)	.97	(2,952)
EPQ/TPQ					
In U.S. sample	15	.81	(1,015)	.78	(2,953)
In Australian sample	37	.81	(2,020)	.78	(3,838)

Note. *N*s in parentheses. GD = Gender diagnosticity. KSP = Karolinska Scales of Personality, EPQ = Eysenck Personality Questionnaire, TPQ = Cloninger's Tridimensional Personality Questionnaire. KSP scales developed from 60 common items; EPQ/TPQ scales developed from 108 common items. All scales derived by contrasted groups method using same criteria as for full scales. Keyed items = number of items in scale derived in that sample from common items.

in the sample in question and one derived from the same set of items in another sample. Each individual is scored twice, once using the key developed in his or her own sample and once using the key developed in the other sample. The correlation reported is that between these two scores. The four correlations given in the table for each sex result from doing this in both directions for the two shared item sets—the 60 KSP items shared by the Swedish and U.S. samples, and the 108 EPQ/TPQ items shared by the U.S. and Australian samples. The correlations are all high, mostly in the .80s and .90s, suggesting that MF is not very sample-specific, at least for modern Western societies.

Age Changes from 12 to 16

For the Australian adolescent sample, we can address issues of across-age stability and change.

Table 4 shows across-age correlations of contrasted-groups GD scores on the age-specific scales and the common scale (those items common to all three of the specific age scales). Individuals' positions on an MF dimension do appear to be changing across these ages, although not drastically so. The correlations are based on the group

Table 4
Across-Age Correlations of Gender Diagnosticity in Australian Adolescent Sample Tested at Three Ages

Group and scale	Ages 12–14	Ages 14–16	Ages 12–16	*N*
Age-specific scale				
Boys	.53	.53	.40	363
Girls	.42	.53	.32	395
Common scale				
Boys	.51	.55	.45	364
Girls	.43	.57	.38	395

Note. Based on individuals tested at all three ages. Age-specific scales: 26, 32, and 37 items; common scale: 20 items.

of individuals tested at all three ages. The 2-year, cross-age correlations on the age-specific scales, ranging from .42 to .53, average about .17 below the internal consistency reliabilities of .57 to .72 for these groups; the 4-year correlations of .32 and .40 are about .12 further down, although still representing some degree of individual stability. Compared to the boys, the girls show relatively more change between ages 12 and 14 than between ages 14 and 16; this may well reflect their earlier sexual maturation.

Table 5 shows means and standard deviations for the common-item GD scale at the three ages. The girls average higher than the boys, consistent with the derivation of the scales. A number of trends may be noted in the data. Both sexes decline in the endorsement of items in the feminine direction, with the drop occurring between ages 12 and 14 for both sexes. The boys drop relatively more than the girls do, .09 versus .04 points, so the difference between the sexes becomes larger between the ages of 12 and 14. As indicated by the standard deviations around the means, variability does not differ greatly between the sexes or across ages. The trends in mean differences were confirmed statistically by means of a nested repeated-measures analysis of variance (SPSS MANOVA: Repeated Measures, SPSS 1990). This analysis indicated significant effects for sex, age, and their interaction, as well as significant linear and quadratic effects for age.

To ascertain whether these age trends might merely reflect trends in some general aspect of response, such as item endorsement frequency, a scale was derived from items on the questionnaire for

Table 5
Age Trends on a GD and a Non-GD Scale for Australian Adolescents

	Age 12		Age 14		Age 16		
Group and scale	*M*	*SD*	*M*	*SD*	*M*	*SD*	*N*
GD, common scale (20 items)							
Boys	.53	.15	.44	.15	.44	.14	364
Girls	.65	.13	.61	.14	.61	.13	395
Non-GD scale (13 items)							
Boys	.62	.12	.62	.12	.63	.11	363
Girls	.62	.13	.62	.12	.64	.12	387

Note. Based on individuals having scores at all three ages. Means represent average frequencies of agreement with the items of the scale (keyed in the feminine direction, for GD scale).

which the sexes did *not* differ. Items were selected that did not exceed an absolute correlation of .05 with sex at any of the three ages. The means for this non-GD scale are also shown in Table 5. The means do not differ materially between the sexes (validating their construction); more relevantly, they show no change across age. Thus, mere changes in some overall response tendency seem unlikely as an explanation of the observed age trends.

GD Changes in Meaning Across a Broad Age Range

Does GD change in content over age? Because items from Eysenck scales were available for all but the Swedish sample, it becomes feasible to look at the correlation between GD and Eysenck's four personality dimensions at ages ranging from 12 years to the elderly. Table 6 shows correlations based on males and females in five samples: the 12-, 14-, and 16-year-olds in Australia, the Australian adults, and the U.S. elderly.

It is clear from looking down the columns of the Table 6 that the ways in which males differ from females shift systematically over age with respect to the Eysenck dimensions. At age 12, the main contrast is between Psychoticism's antisociality, hostility, and nonconformity and the super-good behavior that characterizes the Lie scale; i.e., between good little girls and somewhat-less-likely-to-be-good little boys. This contrast plays a decreasingly small role over age, until in

Table 6
Correlations of GD Scales with Eysenck Scales for Males and Females in Five Samples

	Eysenck Scale*				
Sample	Psych	Extr	Neur	Lie	*N*s
Australian boys age 12	−.76	−.31	.08	.64	819–823
girls age 12	−.60	−.36	.15	.60	801–809
boys age 14	−.61	−.29	.43	.56	648–650
girls age 14	−.49	−.34	.42	.51	622–625
boys age 16	−.50	−.21	.66	.41	599–602
girls age 16	−.37	−.23	.64	.38	634–639
Australian men	−.41	−.22	.51	.22	2,004–2,019
women	−.42	−.25	.51	.16	3,800–3,833
U.S. elderly men	−.12	−.24	.62	−.06	980–994
women	−.14	−.24	.68	−.04	2,879–2,906

Note. *Eysenck scales: Psych = Psychoticism (Psychopathy), Extr = Extraversion, Neur = Neuroticsm, Lie = Lie Scale (Social Conformity). GD scales: contrasted-groups method, full scales.

the elderly it has become a trivial factor. At the same time, Neuroticism's fears, worries, and complaints, which make a negligible contribution to the difference between 12-year-old boys and girls, increase in importance through age 16 and remain a major factor through adulthood. The fourth Eysenckian dimension, Extraversion, has a fairly consistent negative relationship with GD—femininity is modestly associated with introverted attitudes—a tendency that is somewhat stronger at ages 12 and 14 than for the 16-year-olds, adults, and elderly. At all ages, the correlations of GD with the Eysenck scales are similar for males and females. This last may well be at least partially artifactual: if an item is selected for the GD scale that is on a given Eysenck scale, the item overlap will contribute to a correlation between the two scales, and this will be equally true for both sexes.

Heritability of GD Scores

To what extent are masculinity and femininity within the sexes associated with genetic differences, in contrast to environmental ones?

The data on identical and fraternal twins in three of the samples permit heritability estimation by standard behavior-genetic, model-fitting methods. Table 7 provides same-sex twin pair correlations for the U.S. and the Australian twin samples and corresponding correlations reported by Lippa and Hershberger (1999) for a different twin sample, one based on twins who both took the National Merit Scholarship Qualifying Test as high school juniors in the United States in 1962 (Loehlin & Nichols, 1976). Twin correlations are also included from the recent nationwide Add Health sample of U.S. high school students (Cleveland et al., 2001).

In general, the correlations are low to moderate (in the range .14 to .49), suggesting that shared influences—genetic or environmental—are far from accounting for all the variance of MF. The MZ correlations tend to exceed the corresponding DZ correlations, suggesting that GD is appreciably influenced by the genes.

Table 8 shows model-fitting results for within-sex MF in the Australian adult and adolescent and the U.S. elderly twin samples, along with estimates reported by Lippa and Hershberger (1999) for the National Merit sample and by Cleveland et al. (2001) for a recent nationwide adolescent sample (Add Health) containing sibling pairs of varying degrees of genetic similarity ranging from MZ twins

Table 7
Twin Correlations on GD Scale for Same-Sex Pairs from Five Studies

Sample	MZ males		MZ females		DZ males		DZ females	
Australian adults	.48	(383)	.44	(863)	.14	(215)	.23	(497)
U.S. elderly	.38	(143)	.46	(592)	.22	(61)	.27	(277)
Australian age 12	.46	(140)	.43	(144)	.16	(139)	.15	(123)
Australian age 14	.42	(116)	.49	(112)	.28	(112)	.27	(102)
Australian age 16	.47	(113)	.34	(128)	.37	(81)	.17	(85)
US high school students[a]	.45	(202)	.47	(288)	.34	(124)	.34	(193)
US high school students[b]	.30	(99)	.46	(108)	.20	(95)	.22	(95)

Note. MZ = monozygotic, DZ = dizygotic. Numbers of twin pairs in parentheses. Discriminant-based GD scales for all studies. U.S. high school correlations: [a]from Lippa and Hershberger (1999), based on 480 California Psychological Inventory items from National Merit twin sample (Loehlin & Nichols, 1976); [b]from Cleveland, Udry, & Chantala (2001), based on 16 sexually differentiating items from Add Health Wave II in-home sample.

(100% shared genes) to DZ twins and ordinary siblings (50%), to half-siblings (25%).

Standard methods of heritability estimation (Neale & Cardon, 1992, chap. 8.4) were used, which involved model fitting to the MZ and DZ covariance matrices and provided estimates of heritability (h^2), the effects of shared environment (c^2), and a residual (e^2). The models fit used h, c, and e as parameters, precluding negative variance estimates in Table 8, although c is sometimes estimated at its lower bound of zero. In these cases, it was fixed to zero to permit a determinate solution.

Two chi squares are reported in the table for each sample. The first is a test of whether the two sexes can be equated. The chi square represents the difference in fit between a model that fits parameters for the two sexes separately and one that constrains them to be the same for males and females. It has 2 or 3 *df*, depending on whether c was fixed to zero. Equality of the parameters for the two sexes is acceptable in all but three cases: the Australian adults, the Australian 12-year-olds, and the U.S. Add Health sample. For these, the estimates are shown separately for the two sexes. These tend not to differ much, suggesting that the sex differences may mainly involve the variances. The second chi square tests the overall fit of the model shown in the table—in no case does it represent a statistically significant misfit. The fitting program (LISREL 8, Jöreskog & Sörbom, 1993) provides t-tests for the significance of individual parameters. These tests are not shown separately in the table, but the results can easily be summarized: The parameters h and e were both always statistically significant ($p < .05$); c differed significantly from zero only in the case of the National Merit sample.

The model fitting confirms the general impression from inspection of the correlations that GD is moderately heritable (somewhere between 25% and 47% of the total variance). It further shows that in the majority of the samples the parameters for males and females can be assumed equal and that (except for the National Merit sample) the effects of shared environmental factors can be treated as negligible. This implies, by the way, that the substantial contribution of nongenetic factors to MF does not represent environmental influences shared by twins, such as their parents' attitudes and values or their socioeconomic status, but rather reflects such matters as individual experiences, differential genotype-environment interactions, random developmental fluctuations, and measurement error. The

Table 8
Estimates of Within-Sex Additive Genetic, Shared Environmental, and Residual Components of Gender Diagnosticity in Five Studies

Sample	h^2	c^2	e^2	Sexes equal		Goodness-of-fit	
				χ^2	*df*	χ^2	*df*
Australian adult				6.65	2	7.64	8
men	.47	.00	.53				
women	.44	.00	.56				
U.S. elderly	.36	.08	.56	4.03	3	8.42	9
Australian age 12				20.79	2	5.44	8
boys	.43	.00	.57				
girls	.42	.00	.58				
Australian age 14	.38	.08	.54	4.01	3	6.99	9
Australian age 16	.30	.12	.58	6.56	3	15.74	9
U.S. National Merit	.28	.18	.54	ns	3	0.51	5
U.S. Add Health				16.59	2	9.54	20
boys	.25	.00	.76				
girls	.38	.00	.62				

Note. Discriminant-based Gender Diagnosticity scales. h^2 = proportion of additive genetic variance, c^2 = proportion of shared environmental variance, e^2 = remaining variance, including unshared environment, interactions, developmental fluctuations, and errors of measurement. Sexes-equal χ^2 is difference in fit between sexes equated and not; if statistically significant ($p < .05$), parameters for both sexes shown. Goodness-of-fit χ^2 is for fit of model shown; acceptable in all cases ($p > .05$). All h and e parameters are significantly different from zero by t-test; c is only significant for National Merit sample. A c^2 of .00 is estimated at its lower bound, fixed to permit a determinate solution. Estimates for National Merit are from Lippa and Hershberger (1999); for Add Health. from Cleveland et al. (2001). All estimates based on MZ and DZ twins except for Add Health which also included siblings and half-siblings.

majority of the samples yielded fairly similar estimates of the heritability of MF, in the neighborhood of 40%. The high-school-age samples appeared to run a little lower. However, one should not take the exact numerical values for h^2 and c^2 in particular samples too seriously. Confidence intervals about such estimates tend to be wide, even with reasonably large samples such as these, and the assumptions of linearity and additivity, under which the estimates are derived, make them, at most, reasonable approximations. Consistency

across samples, which these estimates tend to show, is the most appropriate guarantee of their meaningfulness.

Genetic and Environmental Sources of Cross-Age Covariation

Model fitting to the longitudinal data from MZ and DZ twins permits a breakdown of the observed cross-age correlations into their genetic and environmental components.

Table 9 shows genetic and environmental contributions to the continuity of MF for males and females over ages from 12 to 16, based on the subgroup of same-sex Australian adolescent twins who were tested at all three ages. It results from fitting Cholesky models (Neale & Cardon, 1992) to the data. A Cholesky is a simple orthogonal factor model that loads all three ages on its first factor, only the second and third on the second factor, and only the third on the third factor. This economical but arbitrary factor solution is not necessarily meaningful in its own right, but it may be multiplied by itself to yield an implied matrix of correlations or covariances among the variables. The first step was to fit two such Cholesky models simultaneously to the covariances of GD across the time periods, one model reflecting genetic contributions to the covariance and the other nonshared environmental contributions. (A nonshared environmental contribution to cross-age covariation may be present because events that occur to only one twin may have effects that persist over time.) Models that incorporated shared environmental covariation C were also considered, but, as in the univariate case, they failed to represent a statistical improvement over the A+E model. The table shows model fitting separately for males and females—the improvement in chi square from fitting the two sexes separately was statistically significant (chi square difference = 27.01 for 12 *df*, $p < .01$). The overall fit of the sexes-different model was excellent (chi square = 53.66, 60 *df*, $p = .70$). The two sides of the table agree in showing high genetic correlations across this age range, correlations approaching unity for the boys. The environment-based, cross-age correlations are much lower, especially for the boys. This contrast is exaggerated by the presence of errors of measurement in the diagonal of the E matrix, which would tend to deflate correlations. Nonetheless, we may conclude that most of the continuity in MF across this age range reflects the persisting effects of genes. The major exception is the girls from age 14 to 16, for whom there appears

Table 9
Multivariate Behavior-Genetic Analysis of Gender Diagnosticity for Males and Females in Australian Adolescent Sample, Ages 12 to 16

		Males			Females		
Source	Age	12	14	16	12	14	16
A	12	.34	*.92*	*1.00*	.36	*.91*	*.86*
	14	.41	.58	*.95*	.31	.32	*.60*
	16	.44	.54	.56	.18	.12	.12
E	12	.66	*.15*	*.09*	.64	*.16*	*.28*
	14	.08	.42	*.14*	.10	.68	*.58*
	16	.05	.06	.44	.21	.45	.88
A+E	12	1.00			1.00		
	14	.49	1.00		.41	1.00	
	16	.49	.60	1.00	.39	.57	1.00

Note. Based on common GD scale, individuals tested at all three ages. Top two matrices: on and below diagonal—variances and covariances calculated via Cholesky models for A (additive genetic variance) and E (residual variance), fitted simultaneously to covariance matrices for male and female MZ and same-sex DZ pairs and standardized; above diagonal, in italics—corresponding genetic and environmental correlations. Bottom matrices: sum of A and E variance-covariance matrices (equivalent to phenotypic correlation matrices).

to be an appreciable environmental contribution to continuity. At the bottom of the table are shown the sum of the A and E covariance matrices, which approximates the ordinary phenotypic correlations for individuals between GD scores at these ages. They agree more or less with the empirical correlations in Table 4, as expected, although they are not identical to them. The Table 9 correlations are inferred from a fitted model based on simplifying assumptions (including the absence of a C component) and are based on smaller samples which do not include unlike-sex twins and non-twin siblings.

DISCUSSION

The analyses in this study have shed various kinds of light on the gender diagnosticity construct and its measurement. Two methods of measurement, the traditional approach via selection of items differentially endorsed by males and females and Lippa's approach via

discriminant-analysis derived probabilities of group membership, were highly correlated in our samples (mostly in the .80s and .90s).

It is a central theoretical feature of gender diagnosticity that it is population-specific. But how specific? We were able to compare measures derived in one population and applied in another, using a U.S. sample of the elderly along with adult samples in Sweden and Australia who had responded to common sets of questionnaire items. About half the correlations between the two scorings were above .95; the lowest were slightly below .80. Thus, among adults in Western societies, gender diagnosticity scales appear to be highly generalizable. This suggests that (within this range) the gender constructs themselves do not vary much across populations.

Williams and Best (1990) did a 25-nation cross-cultural study of MF stereotypes using a contrasted-groups procedure with personality-descriptive adjectives. They found a "considerable degree of cross-cultural generality in the personality characteristics differentially associated with women and men" (p. 226). Adjectives more strongly associated with males in all 25 countries included "adventurous," "forceful," and "independent." Adjectives more strongly associated with females included "sentimental," "submissive," and "superstitious." However, the authors found differences among countries, as well, in such matters as the degree of differentiation of men and women, how similar the country's stereotype was to the cross-nation consensus, and between countries belonging to different religious traditions, such as Catholic versus Protestant, or Hinduism versus Islam. Differences in MF stereotypes were for the most part *not* significantly related to general demographic indicators of social and economic development or of the economic and educational status of women.

One might speculate that in a wider variety of cultures that included primitive societies with very different sex roles, or across long periods of historical time, the distinctions between the sexes, and hence definitions of masculinity and femininity within them, might differ still more widely, making some such approach as gender diagnosticity essential. In our own study, GD provided a convenient way of assessing relationships with a genetic characteristic in samples that had received different personality questionnaires in different languages in different countries.

A somewhat contentious issue in the MF literature is whether MF should be considered as a single dimension or as more than one. Bem

(1974) argued for replacing a single bipolar MF dimension by two orthogonal dimensions of Masculinity and Femininity (allowing for so-called androgynous individuals who scored high on both). Spence and Helmreich (1978) incorporated both approaches. They included separate Masculinity and Femininity dimensions, using items that were positively valued for both sexes but judged more characteristic of one sex than the other. They also retained a bipolar MF dimension consisting of items judged as favorable for one sex and unfavorable for the other. The fact that the present article concentrates on a single MF dimension, albeit one that may differ in content in different samples, does not prejudge the issue of the potential multidimensionality of masculinity and femininity. Williams and Best (1990), for example, began with a single MF dimension, but went on to consider it from several multidimensional perspectives, including Osgood's affective meaning (male-describing adjectives ranked higher on Activity and Potency), and Murray's needs (male-describing adjectives were more strongly associated with needs for Dominance, Autonomy, Aggression, and Exhibition; female-describing adjectives with needs for Nurturance, Succorance, Deference, and Abasement).

It will indeed be interesting in the long run to see if different MF components turn out to have different degrees of cross-cultural generality, different contributions of genes and environments to their variation, and different associations with biological variables. Indeed, we have done a little of this ourselves (Loehlin, et al., 1999). Nonetheless, there are advantages in beginning with a single overall MF dimension, and Lippa's GD approach provides a flexible approach to assessing such a dimension.

In the present article, a number of aspects of GD have been examined. For example, the study provides some information about consistency and changes on such a masculine-feminine dimension in the adolescent years. A 4-year correlation of about .40 between individuals at age 12 and age 16 suggests some degree of stability through the adolescent period. Compared to single-occasion, internal-consistency reliabilities in the mid-.60s in this sample and 2-year correlations in the .50s, it also suggests some degree of change. Changes in group means were greater between the ages of 12 and 14 than between 14 and 16; the average difference between boys and girls also increased during the age 12 to 14 interval. Multivariate behavior-genetic analysis suggested that much of the stability across this age range reflected the genes, particularly for the boys.

It is also apparent that, despite moderate stability of individual differences in MF through the adolescent ages, the content of GD scales does change with age. Social conformity, as indexed by the Eysenck scales Psychoticism (−) and Lie (+), is a major factor at age 12 but recedes to triviality among the elderly. Neuroticism is a negligible factor in femininity for the youngest group but is the largest component after age 16. Introversion is a modest component of femininity across the age range, a little more for the 12- and 14-year-olds than at age 16 or later.

Despite such changes, the heritability of GD appears to be moderately consistent at around 25%–40% across the adolescent ages, and perhaps a little higher for adults and the elderly, suggesting that genetic influences may account for a considerable share of stable individual differences in MF. The trait of MF appears to resemble questionnaire-measured personality traits in general, for which heritabilities in the 40%–50% range and minimal effects of shared family environment are typical (Bouchard & Loehlin, 2001). Note that this is not an automatic consequence of the fact that the GD scales in this study are made up of personality questionnaire items. It is in the covariances among its items that the meaning of a scale primarily resides. In principle, a given set of items could be divided in one way into scales whose heritabilities were high and in another way into scales dominated by environment by grouping items with shared genetic or shared environmental components.

We remain, of course, a long way from understanding the details of genetic and environmental influences on degrees of masculinity and femininity of individuals. Both genes and environment seem to be important, although one part of the environment, that shared by family members, seems not to be a major contributor. Further studies involving particular genes and particular environmental factors will be necessary if we are to understand why individuals within a given sex differ in overall MF, as well as in the specific behaviors and attitudes that constitute it. Much more can also be learned about the nature of and reasons for changes in the components of MF across age.

In summary, in this article we have examined several aspects of Lippa's Gender Diagnosticity measure, a scale that arranges individuals within a sex along a dimension of how much their endorsements of statements about themselves resemble those of the opposite sex, or to what extent they would be predictive of the respondent's

gender. These two different approaches to assessing GD agreed well, and each was moderately reliable in distinguishing among individuals within sexes. Considerable generality across adult populations in Sweden, the United States, and Australia was suggested by high correlations between GD scales derived from the same items in two different populations. Correlations across adolescent ages suggested some, but less than perfect, consistency across age; this was further indicated by trends in means suggesting that gender discriminability increased between ages 12 and 14. Substantial changes in the content of the items diagnostic of gender was also observed over the age span from age 12 to the elderly, in terms of correlations with the four Eysenck dimensions. Correlations suggestive of social conformity (negative with Psychoticism, positive with Lie) predominated early; correlations suggestive of emotional maladjustment and instability (positive with Neuroticism) predominated late. Finally, behavior-genetic model-fitting to the data from MZ and DZ twins suggested that at all ages GD behaved as a fairly typical personality trait, with about 40% of its within-sex variability associated with the genes and little or none with shared environment. Most of the cross-age correlation, especially for the males, appeared to be genetic in origin.

REFERENCES

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

Bouchard, T. J. Jr., & Loehlin, J. C. (2001). Genes, evolution, and personality. *Behavior Genetics*, **31**, 243–273.

Chamberlain, N. L., Driver, E. D., & Miesfeld, R. L. (1994). The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. *Nucleic Acids Research*, **22**, 3181–3186.

Cleveland, H. H., Udry, J. R., & Chantala, K. (2001). Environmental and genetic influences on sex-typed behaviors and attitudes of male and female adolescents. *Personality and Social Psychology Bulletin*, **27**, 1587–1598.

Cloninger, C. R., Przybeck, T. R., & Svrakic, D. M. (1991). The Tridimensional Personality Questionnaire: U. S. normative data. *Psychological Reports*, **69**, 1047–1057.

Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T., & Chakraborty, R. (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, **12**, 241–253.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual for the Eysenck Personality Questionnaire*. London: Hodder & Stoughton.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.

Hofstede, G. (1998). *Masculinity and femininity: The taboo dimension of national cultures*. Thousand Oaks, CA: Sage.

Jönsson, E. G., von Gertten, C., Gustavsson, J. P., Yuan, Q.-P., Lindblad-Toh, K., Forslund, K., et al. (2001). Androgen receptor trinucleotide repeat polymorphism and personality traits. *Psychiatric Genetics*, **11**, 19–23.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.

Lippa, R. A. (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. A. (1995). Do sex differences define gender-related individual differences within the sexes? Evidence from three studies. *Personality and Social Psychology Bulletin*, **21**, 349–366.

Lippa, R. A. (2002). Gender-related traits of heterosexual and homosexual men and women. *Archives of Sexual Behavior*, **31**, 83–98.

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

Lippa, R. A., & Hershberger, S. (1999). Genetic and environmental influences on individual differences in masculinity, femininity, and gender diagnosticity: Analyzing data from a classic twin study. *Journal of Personality*, **67**, 127–155.

Loehlin, J. C., Jönsson, E. G., Gustavsson, J. P., Schalling, M., Medland, S. E., Montgomery, G. W., et al. (2004). Gender diagnosticity and androgen receptor gene CAG repeat sequence. *Twin Research*, **7**, 456–461.

Loehlin, J. C., Jönsson, E. G., Gustavsson, J. P., Schalling, M., & Stallings, M. C. (2003). The androgen receptor gene and psychological traits: Are results consistent in Sweden and Australia? *Twin Research*, **6**, 201–208.

Loehlin, J. C., & Nichols, R. C. (1976). *Heredity, environment and personality: A study of 850 sets of twins*. Austin, TX: University of Texas Press.

Loehlin, J. C., Spurdle, A., Treloar, S. A., & Martin, N. G. (1999). Number of X-linked androgen receptor gene CAG repeats and femininity in women. *Personality and Individual Differences*, **27**, 887–899.

Meade, T. A., & Wiesner-Hanks, M. E. (Eds.) (2004). *A companion to gender history*. Malden, MA: Blackwell.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. Dordrecht, The Netherlands: Kluwer.

Schalling, D., Åsberg, M., Edman, G., & Oreland, L. (1987). Markers for vulnerability to psychopathology: Temperament traits associated with platelet MAO activity. *Acta Psychiatrica Scandinavica*, **76**, 172–182.

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

SPSS (1990). *SPSS Reference Guide*. Chicago: SPSS Inc.

Stallings, M. C., Hewitt, J. K., Beresford, T., Heath, A. C., & Eaves, L. J. (1999). A twin study of drinking and smoking onset and latencies from first use to regular use. *Behavior Genetics*, **29**, 409–421.

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

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, **73**, 213–217.

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

Wright, M. J., & Martin, N. G. (2004). The Brisbane Adolescent Twin Study: Outline of study methods and research projects. *Australian Journal of Psychology*, **56**, 65–78.

Appendix Table A1
Items Scored on Contrasted-Groups GD Scales

Source	Country	Scale	Item pool	Items Scored	Positive items/Negative Items
KSP	Sweden	Adult Full	135	40	k4 k5 k8 k9 k13 k14 k17 k23 k25 k29 k33 k34 k35 k36 k42 k57 k59 k61 k63 k65 k67 k74 k76 k81 k83 k86 k87 k88 k91 k93 k97 k112 k116 k124 k127 k128 k133 k135 / k92 k120
KSP	Sweden	Adult Sw/U.S.	60	22	k4 k8 k9 k14 k17 k29 k33 k34 k36 k42 k57 k61 k74 k76 k81 k86 k88 k97 k116 k124 k127 k128 /
	U.S.	Elderly U.S./Sw	60	18	k4 k9 k17 k21 k33 k34 k36 k42 k57 k61 k74 k76 k81 k86 k97 k108 k116 k121 /
KSP/ EPQ/ TPQ	U.S.	Elderly full	214	41	k4 k9 k17 k21 k33 k34 k36 k42 k57 k61 k74 k76 k81 k86 k97 k108 k116 k121 t18 t27 t31 t41 e6 e28 e38 e47 / t8 t17 t26 t28 t29 t44 t47 t53 t62 t76 t90 t91 e22 e26 e44
EPQ/ TPQ	U.S.	Elderly U.S./Au	108	15	t18 t31 t41 e6 e28 e38 e47 / t8 t17 t26 t62 t90 e22 e26 e44
EPQ/ TPQ	Australia	Adult Au/U.S.	108	37	t3 t18 t19 t23 t31 t34 t83 e4 e5 e6 e14 e28 e38 / t1 t8 t12 t17 t24 t25 t26 t29 t32 t36 t42 t58 t62 t63 t75 t86 t90 e8 e15 e22 e41 e43 e44 e45
EPQ/ TPQ	Australia	Adult full	110	37	t3 t18 t19 t23 t31 t34 t83 e4 e5 e6 e14 e28 e38 / t1 t8 t12 t17 t24 t25 t26 t29 t32 t36 t42 t58 t62 t63 t75 t86 t90 e8 e15 e22 e41 e43 e44 e45

Appendix (Continued)

Source	Country	Scale	Item pool	Items Scored	Positive items/Negative Items
JEPQ	Australia	Age 12	81	26	j26 j27 j29 j30 j31 j43 j45 j63 j73 j77 / j7 j11 j12 j15 j16 j19 j21 j23 j25 j32 j42 j46 j54 j57 j58 j81
JEPQ	Australia	Age 14	81	32	j18 j26 j27 j29 j30 j31 j38 j43 j45 j55 j63 j68 j72 j73 j77 j80 / j7 j11 j12 j15 j16 j19 j25 j35 j42 j46 j50 j52 j54 j58 j75 j81
JEPQ	Australia	Age 16	81	37	j2 j10 j18 j26 j27 j29 j30 j31 j34 j38 j43 j45 j55 j59 j62 j63 j68 j70 j72 j73 j77 j80 / j7 j11 j12 j15 j19 j25 j32 j35 j46 j54 j58 j60 j75 j78 j81
JEPQ	Australia	12 to16 common	81	20	j26 j27 j29 j30 j31 j43 j45 j63 j73 j77 / j7 j11 j12 j15 j19 j25 j46 j54 j58 j81
JEPQ	Australia	12 to16 non-GD	81	13	j1 j5 j8 j9 j14 j33 j36 j47 j53 j61 j64 j66 j74

Note. Item numbers beginning with k correspond to KSP item numbering in the Swedish sample; those beginning with t and e correspond to TPQ and EPQ in the U.S sample; those beginning with j are for JEPQ in the Australian sample. Positive items (those that precede the slash) are items endorsed more frequently by females than by males.