Mutivariate versus Univariate conceptions of Sex Differences: Let the Contest Begin

Posted by RichardLippa on **05 Jan 2012** at **03:36 GMT**

In their paper, "The Distance Between Mars and Venus: Measuring global sex differences in personality," Del Giudice, Booth, and Irwing offer an interesting new perspective on sex differences and a useful critique of Hyde's gender similarities hypothesis [1]. At core, Del Giudice and his colleagues ask: What is the proper metric to use when assessing sex differences in multivariate domains? They nominate the Mahalanobis D statistic—the multivariate generalization of the d statistic—as the best metric to assess sex differences in multi-trait individual differences domains such as personality, cognitive abilities, and interests, and they show empirically that, while on-average sex differences in traits from a given domain (e.g., personality) may be relatively small, the multivariate effect size (D) can simultaneously be quite large.

By way of analogy, consider sex differences in body shape. The Hyde "gender similarities" approach would assess specific traits—e.g., shoulder-waist ratios, waist-hip ratios, torso-to-leg-length ratios, etc.—and then average the d values across these traits, to arrive at the likely conclusion that men and women are more similar than different in body shape. In contrast, the Del Giudice, Booth, and Irwing multivariate approach would more likely generate the conclusion that sex differences in human body shape are quite large, with men and women having distinct multivariate distributions that overlap very little.

Which conclusion is correct? Although there are no God-given prescriptions for proper metrics of effect size, my guess is that lay people would agree more with the Mahalonobis D than with the "mean d" result—i.e., if asked to classify actual human body outlines as "male" or "female," lay people would likely achieve extremely high levels of accuracy by intuitively aggregating across various body-shape dimensions and making "multivariate," configural judgments, despite the fact that ds for some individual body traits might be low.

In advocating the use of the Mahalanobis D statistic, Del Giudice, Booth, and Irwing seem, to me, to be advocating the notion that sex differences in various domains are often multivariate and configural in nature. Such a multivariate approach is especially important in research that explores how well sex differences in personality, cognitive abilities, and interests predict sex differences in real-life criteria, such as participation in STEM (science, technology, engineering, and math) fields, susceptibility to mental and physical illnesses, and the tendency to engage in antisocial behaviors.

For example, to adequately explain men's and women's different participation in STEM fields, researchers need to consider sex differences in a variety of cognitive ability domains: various visuospatial skills, math abilities, mechanical aptitudes, and so on. A still more complete account would focus on sex differences in interests and personality as well. Men's interests are, on average, considerably more thing-oriented and less people-oriented than women's interest are, and women exceed men some on personality traits (e.g., agreeableness, warmth) that may not always find satisfying expression in STEM fields [2, 3].

This discussion of predicting real-life criteria leads to the two additional methodological recommendations made by Del Giudice, Booth, and Irwing: When assessing sex differences in psychological traits, researchers should ensure that (1) trait measures are reliable, and (2) traits are measured at the proper level of specificity. Regarding point (1): Although many gender researchers may not have the statistical expertise or inclination to compute latent factor measures, they nonetheless need to recognize that unreliable trait measures can attenuate sex differences and

they must statistically correct for the unreliability of measures, when possible [4].

One nice feature of Del Guidice, Booth, and Irwing's recommendations is that they can be put to an empirical test. This can be illustrated by research on how well sex differences in personality account for sex differences in antisocial behavior [5]. Del Giudice, Booth, and Irwing suggest that, because of their finer resolution, Big Five facet scores will predict sex differences in antisocial behavior better than Big Five factor scores. This is a testable proposition. They also suggest that when researchers predict sex differences in antisocial behavior from personality measures, they need to employ a multivariate approach to personality. Research shows that sex differences in a number of personality traits—e.g., components of agreeableness, conscientiousness, and neuroticism—contribute to sex differences in antisocial behavior [5]. Thus, the large sex differences in antisocial behavior that are apparent in everyday life probably reflect large multivariate sex differences in personality (in keeping with Del Guidice, Booth, and Irwing's approach). Clearly, the power of the multivariate approach to predict sex differences in criteria such as antisocial behavior is open to empirical investigation.

It is ironic that while the "gender similarities hypothesis" has gained currency among some psychologists, many biological and medical researchers appear to be moving in the opposite direction, increasingly emphasizing the importance of sex differences in various physiological and disease processes [6]. Would biological and medical researchers entertain the Hydean proposition that "males and females are similar on most, but not all, biological variables"? On some level, this assertion seems to be true but, as Del Giudice, Booth, and Irwing note, its truth value depends critically on the specific domain of sex differences under study and on the metric of similarity and difference that researchers use. In practical terms, Hyde's vague "gender similarities hypothesis" will probably provide cold comfort to men and women seeking sound and specific medical advice concerning their heart disease, autoimmune disorders, or medication levels. In biology and medicine, as in psychology, I believe it will prove useful to take a multivariate approach to sex-linked traits in various domains, to acknowledge that some sex differences are small while others are large, and to keep one's eye on the criteria that need to be predicted rather than on broad ideological statements.

Del Giudice, Booth, and Irwing's title employs the much-used "Mars and Venus" metaphor, suggesting a seemingly astronomical separation between the sexes. This is undoubtedly an exaggeration, reflecting a kind of poetic license. Hyde prefers to speak of the distance between North Dakota and South Dakota. However, her metaphor may, inadvertently, reflect a truth she is unwilling to acknowledge: that if you travel from the multivariate "centroid" of one state to the other, you'll still have a mighty long way to walk.

References

1. Hyde JS (2005). The gender similarities hypothesis. Amer Psychologist 60: 581-592.
2.Lippa RA (2005). Gender, nature, and nurture. Mahwah, NJ: Lawrence Erlbaum Associates.
3.Su R, Rounds J, Armstrong PI (2009). Men and things, women and people: A meta-analysis of sex differences in interests. Psych Bull, 135, 859-884.
4.Lippa RA (2006). The gender reality hypothesis. Amer. Psychologist 61: 639-640.
5.Moffit TE, Caspi A, Rutter M, Silva PA (2001). Sex differences in antisocial behavior. Cambridge, England: Cambridge University Press.
6.Blair ML (2007). Sex-based differences in physiology: What should we teach in the medical curriculum? Adv Physiol Educ, 31, 23-25.

The Distance Between North Dakota and South Dakota

Posted by jshyde on **05 Jan 2012** at **02:53 GMT**

In their article, The Distance Between Mars and Venus: Measuring Global Sex Differences in Personality, Del Giudice, Booth, and Irwing challenge my Gender Similarities Hypothesis in the case of personality. Below I show that their methods lead to uninterpretible findings that fly in the face of contemporary personality theory. The Gender Similarities Hypothesis is still accurate and supported by massive amounts of data.

The main innovation in the Del Giudice paper is to introduce the use of Mahalanobis D to the measurement of the magnitude of gender differences. A staple of multivariate statistics for decades, D in this application measures the distance between 2 centroids in multivariate space [1]. It is a multivariate generalization of the d statistic used in many meta-analyses. What is not apparent from the Del Giudice paper, however, is that D is computed by taking the linear combination of the original variables that maximizes the difference between groups. What they have shown is that, if one takes a large enough set of personality measures and then takes a linear combination to maximize gender differences, one can get a pretty big gender difference. That is all they have shown – no more, no less.

An assumption of multivariate normality is crucial to Mahalanobis D if it is to be accurate [2]. The authors provide no statistical verification that their variables are distributed multivariate normally. In other research, apparent findings of large gender differences have crumbled when appropriate statistical methods were used for the non-normal, skewed distributions [3].

The gender difference that Del Giudice and colleagues have found is along a dimension in multivariate space that is a linear combination of the original variables transformed into latent variables. A point that is not mentioned in the Del Giudice article is that this dimension is the first discriminant function. Aside from the fact that the linear combination introduces bias by maximizing differences, the resulting dimension here is uninterpretible. What does it mean to say that there are large gender differences on this undefined dimension in 15-dimensional space created from latent variables? The authors call it global personality, but what does that mean? They promise to measure personality with greater "resolution," yet in the end they have a single, undefined dimension of personality. They have blurred the question rather than offering higher resolution.

Reducing gender differences in personality to differences on a single dimension also flies in the face of contemporary personality theory, which emphasizes distinct dimensions of personality. There is abundant evidence for the Big Five factors of personality: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness [4]. Others would argue for 16 personality factors, captured by the 16PF used in the Del Giudice study [5]. No personality theorists, to my knowledge, are arguing for a single factor or dimension, yet Del Giudice and colleagues base their findings on just that.

Another important point to note is that Del Giudice and colleagues' methods rely on subjective self-reports of personality. When Del Giudice and colleagues talk about "error-free" scores, they are using psychometric terms that may mask the fact that the data are still based on subjective self-reports. As an example, Feingold's meta-analysis [6] found gender differences in anxiety ranging in magnitude between d = -.15 and -.32. That is, the differences were small, with females scoring higher. All of the data were based on self-report personality inventories. In a meta-analysis of research on gender differences in temperament – some of it based on parent or other adult report, some of it based on behavioral measures – the effect size for the gender difference in fear was d = -0.12, i.e., a smaller difference than Feingold found for self-report measures [7]. Moreover, a behavioral study measuring children's distress to the insertion of an intravenous needle showed no significant gender difference [8]. That is, the boys were as anxious and fearful as the girls. Too much of the research on gender differences has relied on subjective self-reports, when objective, behavioral measures may show much different results.

Moreover, subjective self-reports are vulnerable to the effects of stereotyping, and many personality traits are gender stereotyped [9]. Research indicates that people engage in gender-based self-stereotyping when reporting personality traits [10]. The two personality factors that show the largest univariate gender differences in the Del Giudice study are Sensitivity (d = -2.29) and Warmth (d = -.89). It is no accident that warmth and sensitivity are highly female stereotyped traits. Both of them, for example, are items on the Femininity scale of the Bem Sex Role Inventory [11] and on Spence's Personal Attributes Questionnaire [12]. Most women would be reluctant to rate themselves as low on sensitivity and warmth. Many men, in contrast, would be loathe to describe themselves with the girlie attribute of sensitivity. In short, the larger gender differences found by Del Giudice and colleagues may represent no more than gender stereotypes, and may reveal little about actual behavior.

Finally, let us return to the original Gender Similarities Hypothesis [13]. It states that men and women are more similar than different on most (but not all) psychological variables. Evidence came from a review of 46 meta-analyses of psychological gender differences. Of the 124 effect sizes for gender differences extracted from the meta-analyses, 30% had d values falling between 0 and 0.10 (the trivial range) and an additional 48% fell between 0.11 and 0.35 (the small range). These effect sizes represented gender differences across a wide array of psychological variables, including mathematical problem solving, reading comprehension, attributions for success and failure, aggression, and leadership. Del Giudice and colleagues have focused on one particular topic, personality, and, using a technique to maximize differences between groups, they have found a large gender difference on a single, undefined dimension of personality. They argue that males and females are as different in personality as the distance between the planets Mars and Venus. Instead, the overwhelming evidence, across multiple psychological domains, is that men and women are more similar than different; the distance between them is more like the distance between North Dakota and South Dakota.

References

1.Stevens JP (2009). Applied multivariate statistics for the social sciences. 5th ed. New York: Routledge.
2.Overall JE & Klett CJ (1972). Applied multivariate analysis.New York: McGraw-Hill.
3.Pedersen WC, Mill LC, Putcha-Bhagavatula AD, Yang Y (2002). Evolved sex differences in the number of partners desired? The long and the short of it. Psych Sci, 13, 157-161.
4.McCrae RH, Costa PT (1997). Personality trait structure as a human universal. Amer. Psychologist 52: 509-516.
5.Conn SR, Rieke ML (1994) editors. The 16PF fifth edition technical manual. Champagne, IL: Institute for Personality and Ability Testing.
6.Feingold A (1994). Gender differences in personality: A meta-analysis. Psych Bull 116: 429-456.
7.Else-Quest NM, Hyde JS, Goldsmith, HH, Van Hulle, CA (2006). Gender differences in temperament: A meta-analysis. Psych Bull 132: 33-72.
8.McCarthy AM Kleiber C, Hanrahan K, Zimmerman MB, Westhus N, Allen S (2010). Factors explaining children's responses to intravenous needle insertions. Nurs Res 59: 407-416.
9.Gerber GL (2009). Status and the gender stereotyped personality traits: Toward an integration. Sex Roles 61: 297-316.
10.Oswald DL, Chapleau, KM (2010). Selective self-stereotyping and women's self-esteem maintenance. Pers Indiv Diff 49: 918-922.
11.Bem SL (1974). The measurement of psychological androgyny. J Consult Clin Psych 42: 155-162.
12.Spence JT, Buckner CE (2000). Instrumental and expressive traits, trait stereotypes, and sexist attitudes. Psych Women Quart 24: 44-62.
13.Hyde JS (2005). The gender similarities hypothesis. Amer Psychologist 60: 581-592.

ACKNOLEDGEMENTS

The author thanks Kelley Kidwell and Michael Newton for their very helpful comments on an earlier draft of this paper.

Authors' Reply

Posted by marco_dg on **07 Jan 2012** at **14:40 GMT**

First of all, we wish to thank Janet Hyde and Richard Lippa for taking the time to comment on our paper. Here we briefly respond to Hyde's critical remarks.

A. We are not advocating a one-dimensional model of personality. It is a mistake to conflate the dimensionality of the personality model with the fact that D is a single number. The weakness of this criticism can be appreciated by directing it to the measurement of physical distances: if one were to say that the distance between the Earth and the Sun is ~150 millions Km, would this amount to advocating a one-dimensional model of the universe?

B. Hyde wrote, "A point that is not mentioned in the Del Giudice article is that this dimension is the first discriminant function. Aside from the fact that the linear combination introduces bias by maximizing differences, the resulting dimension here is uninterpretable. What does it mean to say that there are large gender differences on this undefined dimension in 15-dimensional space created from latent variables? The authors call it global personality, but what does that mean?" This paragraph contains a number of inaccuracies. First of all, we did mention that D "represents the standardized difference between two groups along the discriminant axis" (p. 3). Second, we never talked about "global personality" (an admittedly vague concept), but rather about "global sex differences" (operationalized with D). Third, neither the D statistic nor the discriminant axis are uninterpretable. The discriminant function can be interpreted as a blend of different traits, but that is not the main point of using D. The most informative way to interpret D is by converting it to an overlap coefficient, as we did in the paper. The same logic (i.e., supporting the interpretation of Cohen's d with overlap coefficients) was followed by Hyde in her 2005 paper [1]; it is unclear to us why distribution overlap should be a valid criterion for univariate comparisons, but not for multivariate ones. Fourth, it is true that estimates of D are affected by sampling error (which may bias the results as more variables are included), but we were careful to compute and report confidence intervals for all the D's in the paper. Unsurprisingly given the large sample size, they turned out to be quite narrow (e.g., 2.66-2.76).

C. Hyde's assertion that "an assumption of multivariate normality is crucial to Mahalanobis D if it is to be accurate" is potentially misleading. On the one hand, we agree that multivariate normality must be assumed in order to accurately convert D to an overlap coefficient. In our analysis using parceled data, the absolute values of the largest skewness and kurtosis statistics were 1.009 and 1.238 respectively, indicating no substantial departure from normality. On the other hand, one may ask whether the univariate estimates and correlations used as input to calculate D have been generated by appropriate methods. In our analysis, explained in detail in a previous paper [2], we used robust maximum likelihood in order to estimate the correlation matrix and the vector of univariate standardized mean differences required to estimate D. Given the low values of skewness and kurtosis and the large sample size, our parameter estimates using robust maximum likelihood can be regarded as highly accurate [3,4].

D. The validity of self-reports is, of course, a long-standing issue in personality research. We acknowledge that self-reports have limitations, and in the paper we call for further research employing multiple assessment methods. In the paper, we also cite research on sex differences in aggression [5] showing that effect sizes on a sex-typed trait can be similar regardless of the assessment method; but the real problem is, does a method exist that is immune to the effects of stereotyping? We worry that the same criticism could easily be raised against other-reports, as well as against behavioral observations (unless one is willing to assume that stereotypes do not influence behavior). Unfortunately, this will make it impossible to satisfy the critics.

Finally, Hyde wrote: “the two personality factors that show the largest univariate gender differences in the Del Giudice study are Sensitivity (d = -2.29) and Warmth (d = -.89). It is no accident that warmth and sensitivity are highly female stereotyped traits.” We agree, but of course this is a chicken-and-egg problem; do these traits show large differences because they are stereotyped, or are they stereotyped because they reliably differ between the sexes? Cross-cultural studies may help resolve this issue, and in the paper we noted that cross-cultural evidence supports the existence of robust sex differences on a personality dimension that closely matches the Sensitivity factor [6]. Whatever the ultimate reason for the large sex differences in sensitivity and warmth, the fact that these traits are regarded as stereotypically feminine does not, by itself, invalidate our findings.

E. Hyde cites two studies in which small sex differences were found [7,8], in support of her argument that sex differences in self-reports are inflated by stereotypical responding. One study is a meta-analysis of children aged 3 months to 13 years [7]; the other is a study of children aged 4 to 10 years [8]. Clearly, some sex differences can be expected to increase with age – especially after puberty. Findings of smaller effect sizes in young children do not invalidate findings of larger effect sizes in adults. Furthermore, none of these studies attempted to estimate effect sizes on latent variables, and the amount of error contained in the measures is unknown. If unreliable or otherwise “noisy” measures are employed, the apparent size of sex differences can be substantially attenuated. This is why, in our paper, we stressed that methods for correcting measurement error should be employed whenever possible.

F. Finally, Hyde returns to her own review of meta-analyses of sex differences [1]. While we acknowledge the very valuable contribution that meta-analyses have made to our understanding of sex differences, we must note that meta-analytic studies often suffer from the same methodological limitations we highlighted in our paper. First, many meta-analyses are not psychometric, that is, they do not correct for artifacts such as measurement error or restriction of range. For this reason, they often provide systematic underestimates of effect sizes. Second, meta-analytic results are dependent on the level of analysis. For example, Hyde and Linn [9] report an overall female advantage on verbal ability of 0.11. However, there is evidence that large sex differences may exist in the specific components of verbal ability. For example, Lynn and Irwing [10] report an overall male advantage on general knowledge of 0.51. In contrast, women evidence advantages on reading (-0.18 to -0.30) and writing (-0.53 to -0.61) [11]. Thus, sex differences at the component level may be averaged out in the broader construct of verbal ability, much as we have argued for personality. Finally, most meta-analyses are conducted on observed scores, and the resulting estimates of effect sizes are therefore distorted by various sources of error.

These points aside, we hope that readers will not miss the major point of our paper, which is eloquently discussed by Lippa in his comment. Our argument is not about how many large or small sex differences there are, it is rather about the methodology for measuring them as accurately as possible. It is also worth stressing that we do not wish to replace univariate effect sizes with their multivariate counterparts; they answer different kinds of questions, and indeed, in our analysis we reported and discussed sex differences on individual factors before turning to multivariate effects. Exactly what the most appropriate methods to measure sex differences are is an important topic of scientific debate. The main goal of our article was to add to this debate. We encourage other researchers to replicate our study with different samples and measures, and of course we look forward to future refinements and extensions of our methodological proposal.

References

1. Hyde JS (2005) The gender similarities hypothesis. Amer Psychologist 60: 581-592.
2. Booth T, Irwing P (2011) Sex differences in the 16PF5, test of measurement

invariance and mean differences in the US standardisation sample. Pers Indiv Diff 50: 553–558.
3. Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In Cudeck, R., Du Toit, S. and Sörbom, D. (Eds), Structural Equation Modeling: Present and future. Lincolnwood, IL: Scientific Software International.
4. Browne, M. W. (1987). Robustness of statistical inference in factor analysis and related models. Biometrica, 74, 375-384.
5. Archer J (2009) Does sexual selection explain human sex differences in aggression? Behav Brain Sci 32: 249–266.
6. Costa PTJ, Terracciano A, McCrae RR (2001) Gender differences in personality traits across cultures: Robust and surprising findings. J Pers Soc Psychol 81: 322–331.
7. Else-Quest NM, Hyde JS, Goldsmith, HH, Van Hulle, CA (2006) Gender differences in temperament: A meta-analysis. Psych Bull 132: 33-72.
8. McCarthy AM Kleiber C, Hanrahan K, Zimmerman MB, Westhus N, Allen S (2010) Factors explaining children's responses to intravenous needle insertions. Nurs Res 59: 407-416.
9. Hyde JS, Linn MC (1988) Gender differences in verbal ability: A meta-analysis. Psychol Bull 104: 53-69.
10. Lynn R, Irwing P, & Cammock, T. (2001) Sex differences in general knowledge. Intelligence 30: 27-39.
11. Hedges LV, Nowell A (1995) Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science 269: 41-45.