Chapter 1
Measuring Sex Differences and Similarities

Marco Del Giudice

Abstract This chapter offers a concise, systematic introduction to quantification in sex differences research. The chapter reviews the main methods used to measure sex differences and similarities, including standardized distances (Cohen's *d* and Mahalanobis' *D*), indices of overlap, variance ratios, and tail ratios. Some less common approaches (e.g., relative distribution methods, taxometrics) are also reviewed and discussed. The chapter examines the strengths and limitations of each method, considers various statistical and methodological factors that may either inflate or deflate the size of sex differences, and discusses the available options to minimize their influence. Other topics addressed include the effective visualization of sex differences/similarities, and the rationale for treating sex as a binary variable despite the complexities of sex-related identity and behavior.

Keywords Cohen's *d* · Gender differences · Gender similarities · Effect size · Mahalanobis' *D* · Measurement · Meta-analysis · Sex differences · Sex similarities

Few topics in psychology can rival sex differences in their power to stir controversy and captivate both scientists and the public. Debates in this area revolve around two types of questions: *explanatory* questions about the role of social learning and biological factors in determining patterns of sex-related behavior, and *descriptive* questions about the size and variability of those effects. These questions are logically distinct and can be addressed independently; however, throughout the history of the discipline the answers have tended to cluster together (see Eagly & Wood, 2013; Lippa, 2005). More often than not, researchers who emphasize socio-cognitive factors typically view sex differences as small, outweighed by similarities, and highly context-dependent. They also tend to worry that exaggerated beliefs about the extent of sex differences and their stability may have pernicious influences on individuals and society (e.g., Hyde, 2005; Hyde et al., 2019; Rippon et al., 2014;

M. Del Giudice (✉)
Department of Psychology, University of New Mexico, Albuquerque, NM, USA
e-mail: marcodg@unm.edu

Unger, 1979). Conversely, most biologically oriented scholars argue that—at least in regard to certain traits—differences between the sexes can be large, pervasive, and potentially universal (e.g., Buss, 1995; Davies & Shackelford, 2008; Ellis, 2011; Geary, 2010; Schmitt, 2015). While not all scholars can be neatly placed in one of these two "camps," the long-standing divide contributes to explain why measurement and quantification are so often at the center of disputes in the field (Eagly & Wood, 2013).

Regardless of one's theoretical background, it is clear that future progress will depend on our ability to quantify differences and similarities as accurately and meaningfully as possible. Doing so requires not only the proper statistical tools, but also awareness of the many factors that may distort empirical findings and make them less interpretable, or even potentially misleading. Despite the importance of these issues, the relevant literature is fragmented; as far as I know, there have been no attempts to organize it in an accessible form. This chapter aims to fill this gap with a concise but systematic introduction to quantification in sex differences research. I begin with a meta-methodological note about the meaning of "sex" and "gender," and the rationale for treating sex as a binary variable despite the complexities of sex-related identity and behavior (a point that necessitates a brief detour into evolutionary biology). In the following section, I review the main approaches to quantification, examine their strengths and limitations, and offer suggestions for visualization. Finally, I discuss various statistical and methodological factors that may inflate or deflate the apparent size of sex differences, and consider the available options to minimize their influence.

1.1 Sex or Gender?

While many authors in psychology and other disciplines treat "sex" and "gender" as synonyms (Haig, 2004), these terms have different histories and implications. The contemporary usage of "gender" as the social and/or psychological counterpart of biological sex was introduced in psychology by Money (1955), though Bentley (1945) had drawn the same distinction 10 years before. Popularized by Stoller (1968), the term was rapidly adopted by feminist scholars in the 1970s (Haig, 2004; Janssen, 2018). The motivation was to distinguish the biological characteristics of males and females from the social roles, behaviors, and aspects of identity associated with male/female labels; usually with the assumption that sociocultural factors are more powerful and consequential than biological ones, and that psychological differences are largely or exclusively determined by socialization (e.g., Oakley, 1972; Unger, 1979). As many have noted over the years, the sex-gender distinction is problematic and ultimately unworkable, which is probably why few authors actually follow it in their writing. Not only does it suggest a clear-cut separation between social and biological explanations; it also presupposes that one already *knows* whether a certain aspect of behavior is biological or socially

constructed in order to pick the appropriate term (Deaux, 1985; Ellis et al., 2008; Haig, 2004).

Having grown uneasy with the sex-gender distinction, some feminist scholars have started to promote the use of the hybrid term "sex/gender" (or "gender/sex") as a way to recognize that biological and social factors are inseparable, encourage critical examination of the processes that lead to observable male-female differences, and underscore the potential for plasticity (Fausto-Sterling, 2012; Hyde et al., 2019; Jordan-Young & Rumiati, 2012; Rippon et al., 2014). Of course this is a legitimate stance; but the new terminology has its own problems, and I suspect that the cure would be worse than the disease. Sex/gender is often described by its proponents as a continuum, or even a multidimensional collection of semi-independent features; from this perspective, a person's sex/gender may be regarded as hybrid, fluid, or otherwise nonbinary (see, e.g., Hyde et al., 2019). Yet, the same term is also used in the context of the distinction between males and females as groups (Jordan-Young & Rumiati, 2012). Some authors have carried this tension to its logical conclusion and suggested that researchers should stop using sex as a binary variable (Joel & Fausto-Sterling, 2016). On this view, "male" and "female" should be replaced with multiple overlapping categories, or even (multi)dimensional scores of gendered self-concepts and attitudes (Hyde et al., 2019; Joel & Fausto-Sterling, 2016). This radical methodological change is justified with the need to overcome the "gender binary." However, the binary nature of sex is not an illusion to dispel but a biological reality, as I now briefly discuss.

1.1.1 The Sex Binary

In the social sciences, sex is usually defined as a collection of traits—X/Y chromosomes, gonads, hormones, and genitals—that cluster together in most people but may also occur in atypical combinations (e.g., Blakemore et al., 2009; Fausto-Sterling, 2012; Helgeson, 2016; Joel, 2012). This definition is the basis for the widely repeated claim that up to 2% of live births are intersex (Blackless et al., 2000). Few researchers and commenters seem aware that the 2% figure is a gross overestimate. To begin, correcting for inaccuracies and counting errors in the original report brings the total frequency down to less than 0.5% (Hull, 2003). More importantly, Blackless et al. (2000) defined intersex very broadly as individuals who deviate from the "Platonic ideal" of sex dimorphism; accordingly, they included several conditions (e.g., Klinefelter syndrome, vaginal agenesis, congenital adrenal hyperplasia) that affect the development of sexual characters but can be classified as "intersex" only in a loose sense (Sax, 2002). If one restricts the term to conditions that involve a discordance between chromosomal and phenotypic sex, or a phenotype that cannot be classified unambiguously as either male or female, the frequency of intersex is much lower—almost certainly less than 0.02% (Sax, 2002; see also Hull, 2003).

A deeper issue with the "patchwork" definition of sex used in the social sciences is the lack of a functional rationale, in stark contrast with how the sexes are defined in biology. From a biological standpoint, what distinguishes the males and females of a species is the size of their gametes: Males produce small gametes (e.g., sperm), females produce large gametes (e.g., eggs; Kodric-Brown & Brown, 1987).[1] Dimorphism in gamete size or *anisogamy* is the dominant pattern in multicellular organisms, including animals. The evolution of two gamete types with different sizes and roles in fertilization can be predicted from first principles, as a result of selection to maximize the efficiency of fertilization (Lehtonen & Kokko, 2011; Lehtonen & Parker, 2014). In turn, anisogamy generates a cascade of selective pressures for sexually differentiated traits in morphology, development, and behavior (see Janicke et al., 2016; Lehtonen et al., 2016; Schärer et al., 2012). The biological definition of sex is not just one option among many, or a matter of arbitrary preference: The very *existence* of differentiated males and females in a species depends on the existence of two gamete types. Chromosomes and hormones participate in the mechanics of sex determination and sexual differentiation, but do not play the same foundational role. Crucially, anisogamy gives rise to a true sex binary at the species level: Even if a given individual may fail to produce viable gametes, there are only two gamete types with no meaningful intermediate forms (Lehtonen & Parker, 2014). This dichotomy is functional rather than statistical, and is not challenged by the existence of intersex conditions (regardless of their frequency), nonbinary gender identities, and other apparent exceptions. And yet, anisogamy is rarely discussed—or even mentioned—in the social science literature on sex and gender, with the obvious exceptions of evolutionary psychology and anthropology.

What are the implications for research? If the sex binary is a basic biological fact, arguments that call for rejecting it on scientific grounds (e.g., Hyde et al., 2019) lose much of their appeal. One can speak of sex differences in descriptive terms—as I do in this chapter—without assuming that such differences are "hardwired" or immune from social influences. From a practical standpoint, sex as a categorical variable is also robust to the presence of a small proportion of individuals who, for various reasons, are not easily classified or do not align with the biological definition. This does not mean that exceptions are unimportant, or that sex should *only* be viewed through a categorical lens. For example, there are methods for ranking individuals of both sexes along a continuum of masculinity-femininity or male-female typicality (e.g., Lippa, 2001, 2010; Phillips et al., 2018; more on this in Sect. 1.2.1). Variations in gender identity and sexual orientation can and should be studied in all their complexity regardless of whether sex is a biological binary. More generally, the existence of a well-defined sex binary is perfectly compatible with large amounts of within-sex variation in anatomy, physiology, and behavior. Indeed, sexual selection often amplifies individual variability in sex-related traits, and can favor the evolution of multiple alternative phenotypes in males and females (Geary, 2010, 2015;

[1] Species with *simultaneous hermaphroditism* (mostly plants and invertebrates) do not have distinct sexes, given that any individual can produce both types of gametes at the same time.

Taborsky & Brockmann, 2010; see also Del Giudice et al., 2018). In the remainder of the chapter I discuss how patterns of quantitative variation between the sexes can be measured and analyzed in detail.

1.2 Quantification of Sex Differences/Similarities

There are many possible ways to quantify sex differences and similarities. In this section I review the methods that are most often employed in the literature. I then discuss some methods that are less common but warrant a closer look, either because of their untapped potential or because of their peculiar limitations. I also address the question of how to visualize quantitative findings effectively and intuitively. Note that the various methods and indices discussed in this section are in no way alternative to one another. Different indices can reveal different aspects of the data, and may be used in combination to gain a broader perspective; other times, one of the indices may be better suited to answer the particular question at hand. The basic formulas are reported and explained in Table 1.1. Additional methods to deal with more complex scenarios can be found in the cited references.

1.2.1 Common Indices of Difference/Similarity

1.2.1.1 Univariate Standardized Difference (Cohen's *d*)

The standardized mean difference is by far the most common and versatile effect size (ES) in sex differences research. Cohen's *d* measures the distance between the male and female means in standard deviation units (using the pooled standard deviation; Table 1.1). Confidence intervals on *d* can be calculated with exact formulas or bootstrapped (Kelley, 2007; Kirby & Gerlanc, 2013). Here I follow the convention of using positive *d* values to indicate higher scores in males. For example, $d = -0.50$ indicates that the female mean is half a standard deviation higher than the male mean. In two major syntheses of psychological sex differences, Hyde (2005) and Zell et al. (2015) summarized hundreds of effect sizes from meta-analyses (see Sect. 1.3.4). They found that about 80% of the effects in the psychological literature are smaller than $d = 0.35$; about 95% are smaller than $d = 0.65$; and only about 1–2% are larger than $d = 1.00$ (absolute values, uncorrected for measurement error; the average across domains was $d = 0.21$ in Zell et al., 2015). For comparison, the size of sex differences in adult height is $d = 1.63$ (average across countries; Lippa, 2009).

The substantive interpretation of *d* values is a persistent source of confusion. The problem can be traced to Cohen (1988), who in a popular book on power analysis offered some conventional rules of thumb for *d:* 0.20 for "small" effects, 0.50 for "medium" effects, and 0.80 for "large" effects. These guidelines have been used countless times to interpret empirical findings and evaluate their importance;

Table 1.1 Common indices for the quantification of sex differences/similarities

Univariate	Multivariate
$d = \frac{m_M - m_F}{S} = \frac{m_M - m_F}{\sqrt{\frac{(N_M-1)S_M^2+(N_F-1)S_F^2}{N_M+N_F-2}}}$	$D = \sqrt{(\mathbf{m_M} - \mathbf{m_F})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{m_M} - \mathbf{m_F})} = \sqrt{\mathbf{d}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{d}}$
Cohen's d. Standardized univariate difference (distance between the M and F means). Convention: Positive values for $m_M > m_F$, negative values for $m_F > m_M$[a] m_M, m_F: Male/female means S: Pooled standard deviation S_M, S_F: Male/female standard deviations N_M, N_F: Male/female sample sizes	*Mahalanobis' D*. Standardized multivariate difference (unsigned distance between the M and F centroids along the M-F axis)[a] $\mathbf{m_M}$,$\mathbf{F}$**:** Vectors of male/female means **d***:* Vector of d values **S***:* Pooled covariance matrix **R***:* Pooled correlation matrix
$d_u = g = d\left[1 - \frac{3}{4(N_M+N_F-2)-1}\right]$	$D_u = \sqrt{\max\left[0, \left(\frac{N_M+N_F-k-3}{N_M+N_F-2}D^2 - k\frac{N_M+N_F}{N_M N_F}\right)\right]}$
Small-sample variant of d corrected for bias (approximate formula); also known as *Hedges' g*	Small-sample variant of D corrected for bias k: Number of variables
$OVL = 2\Phi(-\lvert d\rvert/2)$	$OVL = 2\Phi(-D/2)$
Overlapping coefficient. Proportion of overlap relative to a single distribution[a,b] $\Phi(\cdot)$: Normal cumulative distribution function (CDF)	*Overlapping coefficient*. Proportion of overlap relative to a single distribution[a,b] $\Phi(\cdot)$: Normal cumulative distribution function (CDF)
$OVL_2 = \frac{OVL}{2-OVL} = 1 - U_1$	$OVL_2 = \frac{OVL}{2-OVL} = 1 - U_1$
Proportion of overlap relative to the joint distribution[a,b]	Proportion of overlap relative to the joint distribution[a,b]
$U_1 = 1 - \frac{OVL}{2-OVL} = 1 - OVL_2$	$U_1 = 1 - \frac{OVL}{2-OVL} = 1 - OVL_2$
Proportion of nonoverlap relative to the joint distribution[a,b]	Proportion of nonoverlap relative to the joint distribution[a,b]
$U_3 = \Phi(\lvert d\rvert)$	$U_3 = \Phi(D)$
Proportion of individuals in the group with the higher mean who exceed the median individual of the other group[a,b]	Proportion of males who are more male-typical than the median female (= proportion of females who are more female-typical than the median male)[a,b]
$CL = \Phi\left(\lvert d\rvert/\sqrt{2}\right)$	$CL = \Phi\left(D/\sqrt{2}\right)$
Common language effect size. Probability that a randomly picked individual from the group with the higher mean will exceed a randomly picked individual from the other group[a,b]	*Common language effect size*. Probability that a randomly picked male will be more male-typical than a randomly picked female (= probability that a randomly picked female will be more female-typical than a randomly picked male)[a,b]
$PCC = \Phi(\lvert d\rvert/2)$	$PCC = \Phi(D/2)$
Probability of correct classification (predictive accuracy). Probability of correctly classifying a randomly picked individual as male or female with $d/2$ as the decision threshold[a,b,c]	*Probability of correct classification* (predictive accuracy). Probability of correctly classifying a randomly picked individual as male or female with linear discriminant analysis[a–c]
$\eta^2 = \frac{d^2}{d^2+4}$	$\eta^2 = \frac{D^2}{D^2+4}$
Eta squared. Proportion of variance explained by sex[a–c]	*Eta squared*. Proportion of generalized variance explained by sex[a–c]

(continued)

Table 1.1 (continued)

Univariate	Multivariate
$VR = S_M^2/S_F^2$	$VR = \lvert\mathbf{S}_M\rvert/\lvert\mathbf{S}_F\rvert$
Male:Female variance ratio	Male:female generalized variance ratio $\mathbf{S}_M$, $\mathbf{S}_F$: Male/female covariance matrices
$TR_{zSD} = \frac{\Phi(d-z)}{\Phi(-z)}$	$TR_{zSD} = \frac{\Phi(D-z)}{\Phi(-z)}$
Tail ratio. Relative proportion of males:Females in the region located z standard deviations above the female mean (use $-d$ for the relative proportion of females:Males in the region located z standard deviations above the male mean)[a–c]	*Tail ratio*. Relative proportion of males:females in the region located z standard deviations from the female centroid in the male-typical direction (= relative proportion of females:males in the region located z standard deviations from the male centroid in the female-typical direction)[a–c]

[a]The formula assumes equality of variances (univariate case) or covariance matrices (multivariate case) in the population
[b]The formula assumes (multivariate) normality in the population
[c]The formula assumes equal group sizes (i.e., equal proportions of males and females)

unfortunately, this includes the influential papers by Hyde (2005, 2014) and Zell et al. (2015). The irony is that Cohen did *not* intend these numbers as benchmarks to evaluate effect sizes in empirical data, but only as reasonable guesses to use when behavioral scientists want to perform a priori power analysis but have no information about the likely size of the effect.[2] In fact, what counts as "small" or "large" depends entirely on the area of research, the variables under consideration, and the goals of a particular study (Hill et al., 2008; Vacha-Haase & Thompson, 2004). To give just a few examples: A "small" effect can be quite consequential if the phenomenon of interest happens in the tails of the distribution, where average differences are amplified (Sect. 1.2.1.8). Further, the apparent size of an effect can be diminished by measurement error: when measures are contaminated by high levels of noise, differences may appear much smaller than they actually are (Sect. 1.3.3). Even a difference that is genuinely small from a practical standpoint can have significant theoretical implications if rival hypotheses predict no difference at all. In this context, the practice of labeling differences as trivial if they fall below an arbitrary threshold such as $d = 0.10$ (Hyde, 2005, 2014) is especially troubling.[3] Conversely,

[2]In Cohen's own words: "The terms "small," "medium," and "large" are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation [. . .] In the face of this relativity, there is a certain risk inherent in offering conventional definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference *which is recommended only when no better basis for estimating the ES index is available*." (Cohen, 1988, p. 25; emphasis added). This must have been one of the least successful warnings in the history of statistics.

[3]Of course, it is always possible to test the null hypothesis that a given difference is exactly zero, or within a range that makes it practically equivalent to zero for the purpose of a particular study. In contrast with standard significance testing, Bayesian methods can directly quantify the evidence in

effects that are "large" by Cohen's standards can be nearly useless if one needs to make highly accurate predictions or classifications; to illustrate, $d = 0.80$ implies a predictive accuracy of about 66%, which is better than chance but may be too low in some applied contexts (see Sect. 1.2.1.5). Also, a conventionally "large" effect may be *comparatively* small if the other effects in the same domain are consistently larger. This is not just the case for Cohen's *d*: The same principle applies to all the effect sizes discussed in this chapter. The idea that the practical importance of an effect can be determined mechanically using fixed conventional guidelines is tempting, but deeply misguided.

1.2.1.2 Multivariate Standardized Difference (Mahalanobis' *D*)

Univariate differences are important, but there are situations in which they may easily miss the forest for the trees. Many psychological constructs are intrinsically multidimensional, from personality and cognitive ability to occupational preferences. When investigators are interested in global sex differences within a certain domain, univariate differences calculated for individual variables can be relatively uninformative (or even positively misleading if they are simply averaged together; see Del Giudice, 2009). The reason is that relatively small differences across multiple dimensions can add up to a substantial overall difference. Moreover, the exact way in which multiple variables combine into a global effect size depends on the sign and size of their mutual correlations, and cannot be judged by simply looking at univariate effects. Sex differences in facial morphology nicely illustrate this point (Fig. 1.1a). On average, men and women differ in individual anatomical features such as mouth width, forehead height, and eye size; but univariate

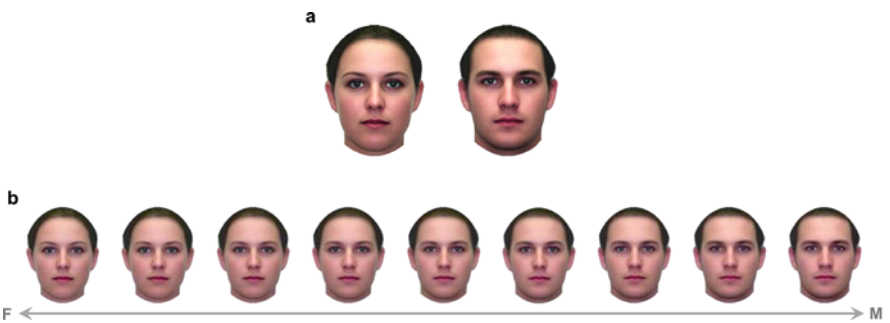

Fig. 1.1 Sex differences in facial morphology. (**a**) Composite male and female faces (averages of 24 pictures each). (**b**) The continuum of male-female typicality in facial features. The figure shows a sequence of morphed faces, from 100% female to 100% male. Adapted with permission from Rhodes et al. (2004). Copyright 2004 by Elsevier Ltd.

support of the null hypothesis (see Dienes, 2016; Kruschke & Liddell, 2018; Wagenmakers et al., 2018).

differences in each of those features (mostly below $d = 1.00$) are too small to accurately distinguish between the sexes. However, the *combination* of multiple features yields two clearly distinct clusters of male vs. female faces, to the point where observers can correctly determine sex from pictures with more than 95% accuracy (Bruce et al., 1993; see Del Giudice, 2013).

The natural metric for measuring global sex differences across multiple variables is Mahalanobis' D, the multivariate generalization of Cohen's d (Huberty, 2005; Olejnik & Algina, 2000; Table 1.1). The value of D is the distance between the centroids (multivariate means) of the male and female distributions, relative to the standard deviation along the axis that connects the centroids. Figure 1.2 illustrates the geometric meaning of D in the case of two variables (for more details see Del Giudice, 2009). The interpretation of D is essentially the same as that of d, with the difference that D is unsigned and cannot take negative values (reflecting the multivariate nature of the comparison). Confidence intervals for D can be obtained with bootstrapping (Kelley, 2005; Hess et al., 2007) or with exact methods, which unfortunately are not always applicable (see Reiser, 2001; Zou, 2007). Procedures for obtaining a pooled correlation matrix are discussed in Furlow and Beretvas (2005). Simple R functions to calculate D with confidence intervals, corrections for bias and measurement error (Sect. 1.3), heterogeneity statistics (see below), and other diagnostics and effect sizes are available at https://doi.org/10.6084/m9.figshare.7934942.

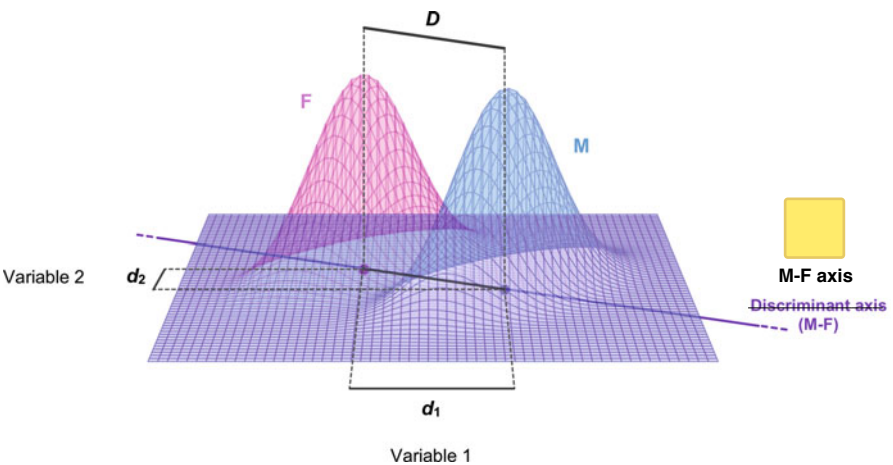

Fig. 1.2 Illustration of Mahalanobis' distance (D) in the bivariate case. D is the standardized distance between the male and female centroids in the bivariate space, taking the correlation between variables into account. (If the variables are uncorrelated, D reduces to the Euclidean distance.) Note that the distributions in the figure are bivariate normal with equal covariance matrices. The axis that connects the male and female centroids can be interpreted as a dimension of male-female typicality or "masculinity-femininity" (M-F) with respect to the relevant variables. Univariate differences are represented as d_1 and d_2

The axis connecting the centroids summarizes the differences between males and females across the entire set of variables, and can be conveniently interpreted as an overall dimension of male-female typicality or masculinity-femininity (M-F) in the domain described by those variables.[4] To illustrate: In the case of facial morphology, the M-F axis would represent a continuum of male-female typicality like the one shown in Fig. 1.1b.[5] This continuum summarizes the combination of anatomical features that make a particular face male- or female-typical. Depending on the size of *D*, the male and female distributions may overlap substantially along the continuum or form largely separate clumps (as in Fig. 1.2). Individual scores on the M-F axis are closely related to the *gender diagnosticity* index proposed by Lippa and Connelly (1990). Gender diagnosticity is the probability that a given individual is male (or, symmetrically, female), estimated with linear discriminant analysis from a set of sexually differentiated variables (e.g., preferences for various occupations or activities). This probability can be used as an index of masculinity-femininity, and is a function of an individual's position along the M-F axis.

In sum, *D* is a convenient index for multivariate differences that generalizes Cohen's *d* and has the same substantive interpretation. Oddly, *D* has been overlooked for decades as a possible measure of group differences (e.g., Huberty, 2002; Vacha-Haase & Thompson, 2004). While *D* has been occasionally discussed as an effect size (Hess et al., 2007; Olejnik & Algina, 2000; Sapp et al., 2007), it has not been used in sex differences research until very recently. An instrumental role in the "rediscovery" of *D* was played by a large-scale analysis of sex differences in personality I performed with my colleagues (Del Giudice et al., 2012), as part of a series of papers on multivariate effect sizes (Del Giudice, 2009, 2013, 2017, 2018). Initial applications of *D* have shown much larger sex differences than previously expected, in domains ranging from personality ($D = 2.71$ in Del Giudice et al., 2012; $D = 2.10$ in Kaiser et al., 2020; uncorrected average $D = 1.12$ in Mac Giolla & Kajonius, 2019; uncorrected average $D = 1.24$ in Lee & Ashton, 2020) and vocational interests ($D = 1.61$ in Morris, 2016) to mate preferences (average $D = 2.41$ in Conroy-Beam et al., 2015). For comparison, the size of multivariate sex differences in facial morphology is about $D = 3.20$ (Hennessy et al., 2005).

An alternative approach followed by some investigators is to combine multiple sex-differentiated variables (e.g., personality items) into a summary score, usually by adding or averaging them together. This method approximates the M-F dimension with a single composite variable; accordingly, effect sizes in these studies are larger than typical univariate differences but smaller than the differences found with

[4] Except in special cases, the M-F axis does not coincide with the discriminant axis. However, the position of an individual point along the M-F axis (i.e., its projection onto the M-F axis in the direction of the classification boundary) is equivalent to its position along the discriminant axis. Thus, scores on the M-F axis provide the same information as discriminant scores.

[5] In this case, "male-female typicality" is arguably preferable to "masculinity-femininity:" studies have shown that when observers make judgements of facial masculinity, they rely on facial cues of body size in addition to sexually dimorphic features (Holzleitner et al., 2014; Mitteroecker et al., 2015).

D in the same domains (e.g., $d = 1.41$ for vocational interests in Lippa, 2010; $d = 1.09$ for personality in Verweij et al., 2016). In a recent paper, Phillips et al. (2018) employed a hybrid method to obtain individual "sex differentiation" scores from brain structure data.[6] First, they computed a differentiation index for each brain feature, based on the ratio of the probability densities in males and females (an approach that is conceptually similar to gender diagnosticity). They then selected a subset of features showing sizable sex differences and averaged them into a summary score. The effect size for this differentiation score was about $d = 1.80$.[7] Depending on how they are constructed, summary scores can be less prone to overfitting the sample data than D (see Sect. 1.3.2); at the same time, they discard information about the correlation structure of the variables and tend to underestimate the overall effect. Note that systematic variation in effect sizes across studies may depend on several factors, from differences in the reference populations (e.g., cross-cultural or age-related effects) to the methods employed to correct for measurement error and other artifacts (more on this in Sect. 1.3.3).

It is worth stressing that multivariate effect sizes like D are not meant to replace univariate indices like Cohen's d. Univariate and multivariate approaches are complementary, and whether one of them provides a more meaningful description of the data is going to depend on the specific question being asked. Criticism of D as an effect size has focused on the supposed lack of interpretability of the M-F axis, and on the fact that D can be inflated by adding large numbers of irrelevant variables (Hyde, 2014; Stewart-Williams & Thomas, 2013). While these points can be readily addressed (see above and Sect. 1.3.2; for a lengthier discussion see Del Giudice, 2013), they do raise the crucial point that D is only meaningful to the extent that it summarizes a coherent, theoretically justified set of variables. A related issue is that many multidimensional constructs in psychology are also hierarchical; for example, the broad-band structure of personality can be usefully described with five broad traits (the *Big Five:* extraversion, openness, agreeableness, conscientiousness, and neuroticism/emotional instability), but each of those traits can be split into multiple narrower traits or "facets" (e.g., the possible facets of extraversion include friendliness, gregariousness, activity, assertiveness, excitement-seeking, and cheerfulness). If sex differences in the lower-order facets of a trait run in opposite directions, they may cancel out at the level of broad traits, leading to underestimates of the actual effect size (see Del Giudice, 2015; Del Giudice et al., 2012). Thus, the choice of the

[6] Of note, Phillips et al. (2018) framed their study as a demonstration that "the sex of the human brain can be conceptualized along a continuum *rather than* as binary" (emphasis added). But this is not what they did: the correlations between sex differentiation scores and other variables were calculated within each sex, meaning that sex was treated as a binary variable and implicitly "controlled for" by analyzing males and females separately.

[7] The paper did not report descriptive statistics for the differentiation score; unfortunately, the raw data were not available for reanalysis (Owen R. Phillips, personal communication, November 2, 2018). I extracted frequencies and central bin values from the histogram in Figure 1 of Phillips et al. (2018) with ImageJ 1.50 (Schneider et al., 2012), and used them to recover approximate sample statistics (females: $M = -0.25$, $SD = 0.29$; males: $M = 0.26$, $SD = 0.27$).

appropriate level of analysis is an important consideration when applying multivariate methods to hierarchical constructs.

Another complication in the interpretation of multivariate indices like D concerns the relative contribution of individual variables to the overall effect. From D values alone, it is impossible to tell whether the multivariate effect reflects the joint contribution of many variables, or the overwhelming contribution of one or a few variables. I have proposed two indices that can be used to aid the interpretation of D (Del Giudice, 2017, 2018). The heterogeneity coefficient H_2 ranges from 0 (maximum homogeneity; all variables contribute equally) to 1 (maximum heterogeneity; the totality of the effect is explained by just one variable). The "equivalent proportion of variables" coefficient EPV_2 (also on a 0–1 scale) estimates the proportion of equally contributing variables that would produce the same amount of heterogeneity, if the other variables in the set made no contribution. Accordingly, smaller values of EPV_2 indicate higher heterogeneity (e.g., $EPV_2 = 0.30$ means that the same amount of heterogeneity would obtain if 30% of the variables contributed equally and the remaining 70% made no contribution to the effect). For example, in the personality dataset analyzed by Del Giudice et al. (2012) the heterogeneity coefficients are $H_2 = 0.90$ and $EPV_2 = 0.16$, suggesting that the overall difference is largely driven by a small subset of variables. Note that there are several possible ways to assign credit to individual variables (e.g., Garthwaite & Koch, 2016); the method used to calculate H_2 and EPV_2 is somewhat ad hoc and will likely be superseded by better alternatives (see Del Giudice, 2018). Still, these indices can be used heuristically to contextualize plain D values and flag patterns that may warrant further attention.

1.2.1.3 Indices of Overlap (OVL, OVL_2)

In contrast with difference metrics, indices of overlap focus on similarity, as they quantify the proportion of the distribution area (or volume/hypervolume) that is shared between males and females. When overlap is high, many males have female-typical scores and many females have male-typical scores. The *overlapping coefficient* (OVL) is the proportion of each distribution that is shared with the other (Bradley, 2006). This is a highly intuitive index of overlap; however, many researchers use a somewhat different index (OVL_2), in which overlap is calculated as the shared area relative to the joint distribution.[8] The corresponding value can be calculated as $1–U_1$, where U_1 is Cohen's coefficient of nonoverlap (Cohen, 1988). Typically, the quantity of interest is overlap rather than nonoverlap; for convenience I use the label OVL_2 to indicate $1–U_1$, the proportion of overlap relative to the joint distribution. While OVL_2 is a common index in psychology, its practical

[8]The difference between OVL and OVL_2 can be visualized by looking at Figure 1.5. OVL = (purple area)/(purple area + blue area) = (purple area)/(purple area + pink area). OVL_2 = (purple area)/(purple area + blue area + pink area).

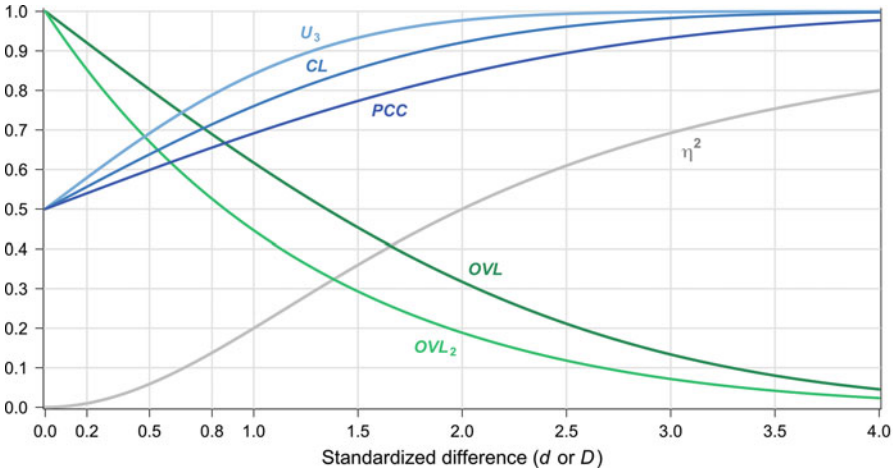

Fig. 1.3 Relations between the standardized mean difference (Cohen's *d* or Mahalanobis' *D*) and various indices of difference/similarity. All conversion formulas assume (multivariate) normality and equality of variances/covariance matrices. See Table 1.1 for details. *OVL* = proportion of overlap on a single distribution. OVL_2 = proportion of overlap on the joint distribution (equals $1–U_1$ in Cohen's terminology). U_3 = proportion of a group above the median of the other group. *CL* = common language effect size ("probability of superiority"). *PCC* = probability of correct classification (assuming equal group sizes). η^2 = proportion of variance explained (assuming equal group sizes)

interpretation is somewhat obscure, and some authors have argued (quite convincingly) that *OVL* is preferable in most contexts (e.g., Grice & Barrett, 2014).

It is easy to convert *d* or *D* values into indices of overlap under the assumption of population normality and equality of variances (in the univariate case), or multivariate normality and equality of covariance matrices (in the multivariate case; Table 1.1). (For brevity, in the remainder of the chapter I will refer to these assumptions as "normality" and "equality of variances/covariances.") The conversion is the same for univariate and multivariate indices, as shown in Fig. 1.3. For example, both $d = 0.50$ and $D = 0.50$ correspond to $OVL = 0.80$ and $OVL_2 = 0.67$, indicating that 80% of each distribution and 67% of the joint distribution are shared between the sexes. Overlap coefficients can also be estimated with nonparametric methods (e.g., Anderson et al., 2012; Schmid & Schmidt, 2006), which may be useful when the standard assumptions are severely violated (see Sect. 1.3.1).

1.2.1.4 Indices of Superiority (U_3, *CL*)

Another way of looking at differences and similarities is to ask what proportion of people in the group with the higher mean would score above the median member of the other group. The answer is provided by Cohen's U_3 coefficient, which can be obtained from *d* or *D* under the same assumptions of overlap indices (Fig. 1.3;

Table 1.1). For example, both $d = 0.50$ and $D = 0.50$ correspond to $U_3 = 0.69$. Following the usual conventions, $U_3 = 0.69$ with a positive d means that 69% of males score above the median female (or, equivalently, that 69% of females score below the median male; Cohen, 1988). The interpretation of U_3 changes slightly when one is dealing with a multivariate distribution. Specifically, U_3 becomes the proportion of males that are more "masculine" or "male-typical" than the median female—or, symmetrically, the proportion of females that are more "feminine" or "female-typical" than the median male.

The *common language effect size* (*CL*; also known as "probability of superiority") is another popular index that translates group differences into probabilities. Specifically, *CL* is the probability that a randomly picked individual from the group with the higher mean will outscore a randomly picked individual from the other group (McGraw & Wong, 1992). By assuming normality and equality of variances/covariances, *CL* can be easily obtained from d or D (Fig. 1.3; Table 1.1). As with U_3, the interpretation of *CL* changes somewhat in a multivariate context, and becomes the probability that a randomly picked male will be more "masculine" or "male-typical" than a randomly picked female (or, symmetrically, the probability that a randomly picked female will be more "feminine" or "female-typical" than a randomly picked male). The original *CL* index can be generalized to discrete distributions (Vargha & Delaney, 2000), and there are procedures to calculate confidence intervals when standard assumptions do not apply (Vargha & Delaney, 2000; Zhou, 2008).

1.2.1.5 Probability of Correct Classification (*PCC*)

The *probability of correct classification* (hereafter *PCC*), *predictive accuracy*, or *hit rate* is the probability that a randomly picked individual will be correctly classified as male or female based on the variable(s) under consideration.[9] The ability to reliably infer the sex of an individual can have considerable practical value and offers an intuitive measure of the degree of statistical separation between two groups. This approach to quantification differs from those reviewed until now in that the probability of success depends (implicitly or explicitly) on the statistical model used to perform the classification task. The problem is greatly simplified when the assumptions of normality and equality of variances/covariances are satisfied. If this is the case, linear discriminant analysis (LDA) approximates the optimal classifier (James et al., 2013), and the *PCC* can be estimated as a simple function of the standardized difference d or D, assuming equal group sizes (Fig. 1.3; Table 1.1; Dunn & Varady, 1966; Hess et al., 2007). For example, both $d = 0.50$ and $D = 0.50$ correspond to $PCC = 0.60$, that is, a 60% probability of correctly classifying a random individual as male or female. Returning to the example of male/female faces discussed earlier, the predictive accuracy of human observers is 0.95 or more; under

[9] This is different from gender diagnosticity (Sect. 1.2.1.2), which is the estimated probability that a *particular* individual is male (or female), regardless of his/her actual sex.

standard assumptions, this would imply a multivariate difference $D \geq 3.30$, a figure very close to the one estimated from face morphology data (about $D = 3.20$ in Hennessy et al., 2005).

If variances/covariances differ between the sexes but normality still applies, the approximately optimal classifier is not LDA but QDA (quadratic discriminant analysis; see James et al., 2013). When distributions are strongly non-normal and patterns of sex differences are characterized by nonlinearity and higher-order interactions, the *PCC* is going to depend on the particular classification model chosen for the analysis. The menu of available methods has been expanding rapidly thanks to advances in machine learning; common options include logistic regression, classification trees, support vector machines (SVMs), and deep neural networks (see Berk, 2016; Efron & Hastie, 2016; James et al., 2013; Skiena, 2017). Sophisticated classification methods can be especially effective in complex datasets with large numbers of variables; it is not a coincidence that many recent applications to sex differences come from neuroscience. To give just a few examples: van Putten et al. (2018) trained a neural network on electroencephalogram signals (EEG) and were able to identify the sex of participants more than 80% of the time. Using regularized logistic regression, Chekroud et al. (2016) achieved 93% accuracy in identifying the sex of adult participants from brain structure. The same accuracy (93%) was reported by Anderson et al. (2018) with SVM and regularized logistic regression, and by Xin et al. (2019) with a neural network. By applying SVM to brain scan data, Joel et al. (2018) obtained 72–80% accuracy in adults, while Sepehrband et al. (2018) achieved 77–83% accuracy in children and adolescents. In all these studies, classification was performed on multivariate data from the whole brain, not on individual brain regions. Interestingly, the sex differentiation score computed by Phillips et al. (2018) from brain structure data (see Sect. 1.2.1.2) yields an expected $PCC = 0.82$ (estimated from $d = 1.80$), which is close to the performance of more complex algorithms.[10]

1.2.1.6 Variance Explained (η^2)

The proportion of variance in the variable of interest that is explained by a categorical predictor (e.g., sex) is usually labeled *eta squared* (η^2; see Lakens, 2013; Olejnik & Algina, 2000). This is a classic effect size but not a very intuitive one; for this reason, it is seldom employed in sex differences research (but see Deaux, 1985). The value of η^2 can be obtained from d or D assuming normality and equality of variances/covariances; for simplicity, the formulas presented in Table 1.1 also assume equal group sizes. As can be seen in Fig. 1.3, $d = 0.50$ and $D = 0.50$

[10] Note that multivariate patterns of sex differences in brain structure are strongly influenced by sex differences in total brain volume. Because different regions show different scaling functions with respect to overall volume, simple linear adjustments do not fully remove the effect of males having larger brains on average. In a recent study that used more sophisticated correction methods, classification accuracy dropped from more than 80% to about 60% (Sanchis-Segura et al., 2020).

correspond to $\eta^2 = 0.06$, or 6% of variance explained by sex. Explaining 50% of the variance requires a male-female difference of two standard deviations. The main problem with indices of variance explained is that values perceived as "small" are easy to underestimate and dismiss as trivial, even when they reflect meaningful or practically important effects (for extended discussion of this point see Abelson, 1985; Breaugh, 2003; Prentice & Miller, 1992; Rosenthal & Rubin, 1979).

1.2.1.7 Variance Ratio (*VR*)

Males and females may differ not only in their mean value on a trait, but also in their *variability* around the mean. When computing most of the indices reviewed in this chapter, unequal variances are treated as a deviation from standard assumptions (Table 1.1); however, systematic differences in variability may be interesting in their own respect, for example because they can have large effects on the relative proportions of males and females at the distribution tails (Sect. 1.2.1.8).

Empirically, males have been found to show larger variance than females in a majority of traits, including most dimensions of personality (except neuroticism; see Del Giudice, 2015), general intelligence (e.g., Arden & Plomin, 2006; Dykiert et al., 2009; Johnson et al., 2008), specific cognitive skills (e.g., Bessudnov & Makarov, 2015; Hyde et al., 2008; Lakin, 2013; Wai et al., 2018), brain size (e.g., Ritchie et al., 2018; Wierenga et al., 2017), and many other bodily and physiological features (see Del Giudice et al., 2018; Lehre et al., 2009). In the human literature, this is known as the "greater male variability hypothesis" (for a historical perspective see Feingold, 1992), but the same general pattern is apparent in most sexually reproducing species (Wyman & Rowe, 2014; Del Giudice et al., 2018). Some of these differences seem to reflect scaling effects: If the variability of a trait increases with its mean level, the sex with the higher mean will also show the larger variance. This is the case for physical traits such as height, body mass, and brain volume. While the variance of these traits is higher in males, the coefficient of variation (i.e., the standard deviation divided by the mean) is very similar in men and women (Del Giudice et al., 2018). However, greater male variance is also found in domains in which average differences are very small or favor females (such as general intelligence and most personality traits).

The standard index for sex differences in variability is the *variance ratio* (*VR*), which by convention is the ratio of the male variance to the female variance. In sex differences research, variance ratios are usually calculated on univariate distributions (confidence intervals on *VR* are discussed in Shaffer, 1992). However, the generalized variance of a multivariate distribution is the determinant of the covariance matrix (Sen Gupta, 2004); a generalized variance ratio can be easily obtained as the ratio of the male and female generalized variances (Table 1.1). Equality of variances corresponds to $VR = 1.00$. In the domains of personality and cognition, values of *VR* estimated from large samples are often smaller than 1.20 and rarely larger than 1.50. For neuroticism and related traits, which tend to be more variable in females, *VR* usually ranges between 0.90 and 1.00 (Del Giudice, 2015; Hyde, 2014;

Lakin, 2013; Lippa, 2009). For comparison, the variance ratio for height is estimated at about $VR = 1.11$ (average across countries; Lippa, 2009).

1.2.1.8 Tail Ratio (*TR*)

The relative proportions of males and females in the region around the mean are often less interesting than their representation at the tails of the distribution. This is typically the case when the outcome of interest depends on competition (e.g., selection of the top-ranking applicants for a job), the crossing of a threshold (e.g., selection requiring a minimum passing score), or other nonlinear effects (e.g., the probability of committing violent crimes may increase more steeply at the upper end of the distribution of aggression). Crucially, small differences between means can have a substantial impact as one moves toward the tails of the distribution; and even if males and females have exactly the same mean on a trait, sex differences in variability can produce marked differences at the extremes (Halpern et al., 2007).

When the tails of the distribution are the focus of interest, summary indices such as mean differences and overlap coefficients are uninformative; researchers may wish to calculate a *tail ratio* (*TR*), that is, the relative proportion of the two sexes in the region above (or below) a certain cutoff. Here I adopt a slight variation of the reference group method proposed by Voracek et al. (2013); the alternative approach by Hedges and Friedman (1993) uses the total distribution of the two groups combined. In the standard version of Voracek et al.'s method, the group with the lower mean serves as the reference group, and the cutoff to identify the tail is placed at z standard deviations from the lower mean (where z can be any value). The choice of cutoff is noted as $TR_{z\mathrm{SD}}$: for example, $TR_{2\mathrm{SD}}$ is the tail ratio for a cutoff located $z = 2$ standard deviations above the lower mean; $TR_{2.5\mathrm{SD}}$ is the tail ratio for a cutoff located $z = 2.5$ standard deviations above the lower mean; and so on. In the context of sex differences, it is arguably more useful to pick one of the two sexes as the reference group regardless of the ranking of means; in the following I use females as the reference group, following the standard convention for variance ratios. While Voracek et al. (2013) proposed benchmarks for the interpretation of *TR* modeled on those for Cohen's *d*, fixed conventions are even less meaningful in this context and should probably be avoided.

Tail ratios can be estimated from means and variances assuming normality, or from *d* and *D* with the additional assumption of equal variances/covariances (Table 1.1). However, the resulting estimates can be very sensitive to violations of these assumptions (see Sect. 1.3.1), and researchers working with large samples often calculate tail ratios directly from frequency data rather than from summary statistics (e.g., Lakin, 2013; Wai et al., 2018). Figure 1.4 shows how *d* determines the tail ratios above three common cutoffs. With equal variances ($VR = 1$), an effect size $d = 0.50$ corresponds to $TR_{1\mathrm{SD}} = 1.94$, $TR_{2\mathrm{SD}} = 2.94$, and $TR_{3\mathrm{SD}} = 4.60$. In other words, there are almost twice as many males as females in the region one standard deviation above the female mean ($TR_{1\mathrm{SD}}$); almost three times as many in the region two standard deviations above the female mean ($TR_{2\mathrm{SD}}$); and 4.6 times as

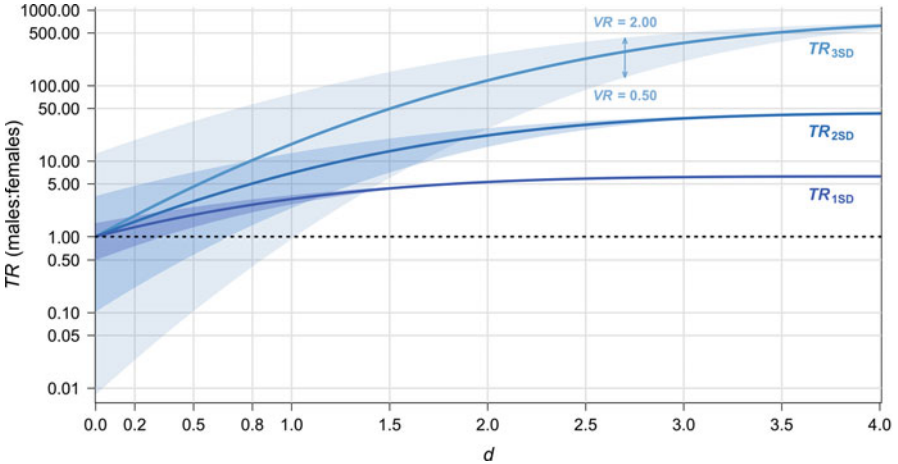

Fig. 1.4 Tail ratios and the effect of unequal variances. The thick lines show the relative proportion of males to females above the cutoffs located at one, two, and three standard deviations from the female mean (TR_{1SD}, TR_{2SD}, and TR_{3SD}) for positive values of *d*. Calculations assume normality, equal group sizes, and equal variances in the two sexes (variance ratio $VR = 1.00$). The shaded areas represent changes in tail ratios when variances are unequal, ranging from $VR = 0.50$ (twice as high in females) to $VR = 2.00$ (twice as high in males). Note that the impact of unequal variances on *TR* is stronger when the difference between means is smaller and/or the cutoff is more extreme

many in the region three standard deviations above the female mean (TR_{3SD}). As the standardized difference increases, *TR* becomes disproportionately larger (note that the vertical axis of Fig. 1.4 is logarithmic). Figure 1.4 also illustrates the major impact of unequal variances, which—depending on how they combine with distribution means—can dramatically amplify sex imbalances in the tails, but also attenuate or even reverse them. While standardized differences and overlap coefficients are robust to minor sex differences in variability, tail ratios can be remarkably sensitive to unequal variances. Specifically, the impact of *VR* is maximized when *d* or *D* values are smaller and/or the chosen cutoff is more extreme (Fig. 1.4).

1.2.2 Other Methods

1.2.2.1 Relative Distribution Methods

A powerful but surprisingly underused approach to group differences employs the statistical concept of a *relative distribution* to compare the distribution of a comparison group to that of a reference group (Handcock & Janssen, 2002; Handcock & Morris, 1998, 1999). A key tool of relative distribution methods is the *relative density plot,* which shows how the ratio of the comparison distribution (e.g., males) to the reference distribution (e.g., females) changes at different levels (quantiles) of the reference distribution. An example of relative density plot is

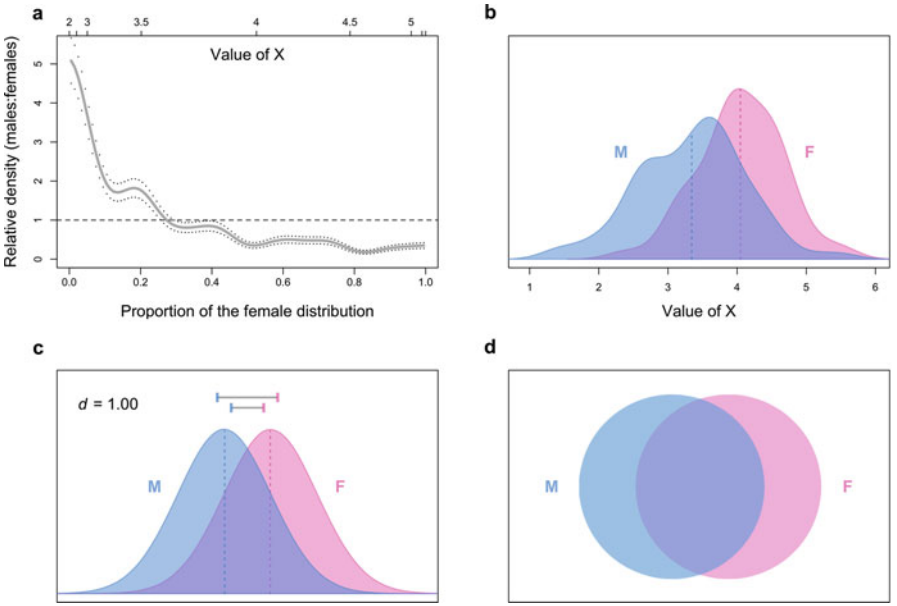

Fig. 1.5 Four visualizations of sex differences/similarities. All plots are based on the same dataset with $d = 1.0$. (**a**) Relative density plot. This plot shows the relative male:female density at different quantiles of the female distribution (bottom axis); the corresponding values of the variable (X) are shown for references on the top axis. Dotted lines represent 95% pointwise confidence intervals. Assuming equal group sizes, a relative density of 1.0 (horizontal dashed line) indicates equal proportions of males and females. Under the same assumption, there are about five time as many males as females with values at the lower extreme of the female distribution (0.0 on the bottom axis; relative density ≈ 5.0). At the median of the female distribution (0.5 on the bottom axis) there are about three times as many females as males (relative density ≈ 0.3), approximately the same proportions found at the upper extreme (1.0 on the bottom axis). (**b**) Overlay density plot of the male and female distributions. This plot shows the shape of the distributions, their overlap, and the location of means (vertical dotted lines). (**c**) Normalized plot of the male and female distributions. This plot shows the standardized mean difference and the corresponding overlap assuming normality and equality of variances (in this case, $OVL = 0.62$ and $OVL_2 = 0.45$). Horizontal bars represent 95% confidence intervals on d; the colors on the bottom bar can be reversed when the interval includes opposite-sign values. (**d**) Venn diagram of the overlap between the male and female distributions. This type of diagram can be used to intuitively communicate the overall size of effects in complex multivariate contexts

shown in Fig. 1.5a (see the figure legend for more details). The relative density function contains all the information about the differences between the male and female distributions, including differences in central tendency (mean or median), differences in variability, and differences in the shape of the two distributions. For this reason, the relative density plot is a remarkably informative display that can be used for high-resolution exploration of the data (for a conceptually similar approach that employs quantile differences instead of relative densities, see Rousselet et al., 2017; Wilcox, 2006).

Besides visual exploration, relative distribution methods also support various types of quantitative inference. Most intriguingly, the relative distribution can be easily decomposed into independent components that separate the effects of location (i.e., differences in means or medians) from those of shape (including, but not limited to, differences in variance). These components of the distribution can be plotted separately to visually examine their characteristics, or quantified and compared using information-theoretic measures (for details and examples see Handcock & Morris, 1998). Despite their many attractive features, relative distribution methods have been largely ignored in sex differences research; the few applications I am aware of—limited to relative density plots—are in Bessudnov and Makarov (2015), Del Giudice (2011), and Del Giudice et al. (2010, 2014).

1.2.2.2 Taxometric Methods

The goal of taxometrics is to use observable indicators to infer the latent structure of a given domain (Beauchaine, 2007; Meehl, 1995; Ruscio et al., 2011, 2013). Specifically, taxometric procedures examine patterns of variation and covariation among indicators to distinguish cases in which differences between individuals are purely dimensional (e.g., a continuum of increasingly severe antisocial behavior) from those in which the data reflect the existence of categories with non-arbitrary boundaries (e.g., psychopaths vs. non-psychopathic individuals)—or, stated otherwise, categories that differ from one another in kind and not just degree (*taxa*). Taxonic and dimensional variation are not mutually exclusive, and often coexist within the same domain (e.g., psychopaths may vary in the severity of their antisocial symptoms; see Ruscio et al., 2013).

Carothers and Reis (2013; Reis & Carothers, 2014) performed a taxometric analysis on various putative indicators of gender, which they distinguished from biological sex: measures of sexuality, mating preferences, empathy, intimacy, and personality (including the Big Five). They found overwhelming support for a dimensional model and concluded that the latent structure of gender—in contrast with that of sex—is not a binary but a continuum. They also argued that average sex differences are "not consistent or big enough to accurately diagnose group membership" (p. 401). However, a simpler interpretation of these findings is that the indicators used in the study were too weak to detect the underlying taxa. As also noted by the authors, taxometric procedures quickly lose sensitivity as group differences on the indicators become smaller than $d = 1.20$ (Beauchaine, 2007; Ruscio et al., 2011); but almost all the effect sizes in the study were below this threshold, and often substantially so. Because the indicators were inadequate to detect taxonic differences, the analysis predictably indicated a dimensional structure. The only set of psychological indicators with adequate effect sizes was a list of preferences for sex-typed activities (e.g., boxing, hair styling, playing golf). Predictably, sex-typed activities showed clear evidence of taxonicity, but this result was not treated as part of the main analysis. Also, the authors' claim that sex differences are too small and inconsistent to infer a person's sex from psychological measures is

unfounded: Personality traits alone can correctly classify males and females with high probability, provided they are measured at the level of narrow traits and aggregated with multivariate methods. For example, $D = 2.71$ (Del Giudice et al., 2012) yields $PCC = 0.91$ using the standard formula.[11] In contrast, the Big Five lack the resolution to accurately differentiate the sexes, and the corresponding effect sizes ($d = 0.19$–0.56 in the study) are too small to regard these traits as valid taxometric indicators. In light of these limitations, the findings by Carothers and Reis (2013) are hard to interpret with any confidence.

Beyond this particular study, it is unclear whether taxometric methods can make a substantive contribution to sex differences research. The purpose of taxometrics is to probe for the existence of taxa that cannot be directly observed, as is often the case with mental disorders (Meehl, 1995). In meaningful applications, one does not know a priori whether the hypothetical taxa exist or not, and there is a genuine possibility that the underlying structure of the data is fully dimensional. But in the case of sex differences, the taxa (males and females) are already known to the investigators, and indicator variables are chosen precisely because they can distinguish between males and females. Given these premises, studies that use sufficiently strong indicators (e.g., sex-typed activities) can be expected to confirm the existence of two sexes, whereas studies that use weak indicators will be uninterpretable because of their lack of sensitivity, as in Carothers and Reis (2013). Either way, the results are going to be uninformative, unless the goal is to look for *additional* taxonic distinctions within each sex (e.g., discrete categories related to sexual orientation; Gangestad et al., 2000; Norris et al., 2015).

1.2.2.3 Internal Consistency Analysis

Internal consistency analysis was introduced by Joel et al. (2015) in a famous study of sex differences in brain structure. The first step of this procedure is to select a subset of variables showing the largest sex differences (e.g., volumes of particular brain regions) and split each of them in three equal-sized categories—the most male-typical third, the most female-typical third, and the middle third. Each participant is then classified based on his/her combination of variables: Participants who fall in the male-typical, female-typical, or intermediate category on all the variables are deemed "internally consistent"; those with at least one male-typical and one female-typical variable are said to show "substantial variability" and are regarded as "mosaics." Joel et al. (2015) found very low proportions of internally consistent individuals (ranging from 0.1% to 10.4%), not only in brain structure but also in personality, attitudes, and preferences for sex-typed activities. Based on these

[11] Of course, this effect size is based on latent variables, and the corresponding *PCC* assumes error-free measurement (Sect. 1.3.3). The point remains valid: in principle, a combination of narrow personality traits can accurately discriminate between males and females. Note that Carothers and Reis' claim concerned the *actual* amount of overlap between the sexes, not the attenuating effects of measurement error.

findings, they claimed that most people are characterized by a mosaic of male and female brain features, a pattern that undermines any attempt to distinguish between "male" and "female" brain types. In later work, Joel and others have argued that extensive brain mosaicism calls into question the use of sex as an independent variable in neuroscience (Joel & Fausto-Sterling, 2016; Hyde et al., 2019).

Unfortunately, the method devised by Joel et al. (2015) is seriously flawed. The threshold for consistency is both arbitrary and exceedingly high: It is easy to show that, in realistic conditions, the method *always* returns a small proportion of "internally consistent" individuals, regardless of the pattern of differences and correlations among variables (Del Giudice et al., 2015, 2016). This remains true even when the variables show unrealistically high levels of consistency (i.e., all correlations among variables equal to 0.90). In light of this, it is not surprising that Joel et al. (2015) found only 1.2% of internally consistent individuals in the domain of sex-typed activities, with the same data that showed clear evidence of taxonicity in Carothers and Reis' (2013) analysis.[12] While "substantially variable" profiles are more sensitive to variations in the data (Del Giudice et al., 2015; Joel et al., 2016), the percentages returned by this method can be quite misleading if taken at face value. The authors have continued to present their findings as evidence that most brains are "gender/sex mosaics" (Joel & Fausto-Sterling, 2016; Hyde et al., 2019). The question they address is without doubt an important one; patterns of consistency/inconsistency among sex-related traits can be both theoretically interesting and practically important. However, their method is designed to show invariably low levels of internal consistency, and I cannot recommend it as a useful analytic tool.

1.2.3 Visualization

There are many possible ways to visualize sex differences/similarities in plots and diagrams; the most appropriate type of display is going to depend on the researchers' aims and their intended audience. Figure 1.5a shows a relative density plot with females as the reference group (Sect. 1.2.2.1). This plot does not depict the original distributions but only their relative differences, and highlights the behavior of the variable in the tail regions. While relative density plots can be very informative, they are not immediately intuitive and require some technical background to interpret. A similar type of plot based on quantile differences instead of relative densities is discussed in Rousselet et al. (2017) and Wilcox (2006). In Fig. 1.5b, the male and female probability densities are overlaid on the same plot (e.g., Ritchie et al., 2018). This straightforward display conveys a lot of information, including the shape of the

[12] To see why, consider a fictional man who hates talk shows and cosmetics and is passionate about boxing and video games (male-typical values), but does not particularly like golf (intermediate). He would be classified as showing an "intermediate" profile of gendered interests. If he happened to dislike golf (female-typical value), he would be classified as a sex/gender mosaic with a "substantially variable" interest profile (see Del Giudice et al., 2015).

two distributions, the difference between means, and the amount of male-female overlap—though it is less effective than the relative density plot in showing differences in the tail regions. Overlay density plots are similar to split violin plots, in which densities are displayed side by side instead of overlaid (e.g., Wai et al., 2018); however, split violin plots make it hard to visualize the overlap between distributions. Both density and relative density plots can be used to visually detect obvious deviations from standard assumptions.

When effect sizes are mapped on normal distributions (with equal or unequal variances), normalized density plots (Fig. 1.5c) offer an intuitive display of standardized differences and overlaps (e.g., Maney, 2016[13]). Plots of actual or normalized distributions can be easily augmented with confidence intervals on *d*, as shown in Fig. 1.5c. Still, this kind of plot is inherently univariate, and can be misleading when one wants to present the results of multivariate analyses. In complex multivariate contexts, the overlap between distributions is usually the most intuitive metric; overlap coefficients can be visualized with Venn diagrams (Fig. 1.5d) in which areas represent proportions of overlap and nonoverlap (e.g., Del Giudice et al., 2012).

1.3 Statistical and Methodological Issues

1.3.1 Assumption Violations

Many of the standard formulas presented in this chapter make the assumptions of normality and equality of variances/covariances in the population. These formulas are useful because they allow investigators to calculate a wide range of indices from commonly reported statistics such as means, standard deviations, correlations, and values of *d* or *D*. Moreover, some non-standard indices (e.g., multivariate overlap between non-normal distributions) may be complicated to obtain even if raw data are available. Still, deviations from normality are quite common: Empirical data are frequently skewed, have heavier tails than expected under a normal distribution, and so on (e.g., Limpert & Stahel, 2011). The size of indices like *d* and *D* is sensitive to both non-normality and the presence of outliers (Wilcox, 2006); moreover, exact formulas for confidence intervals are only accurate when normality can be assumed. Remedies to these distorting effects include bootstrap confidence intervals and robust variants of Cohen's *d* that eliminate the influence of extreme values (e.g., Algina et al., 2005; see Kirby & Gerlanc, 2013). Deviations from normality may also change the amount of overlap between distributions. When this is the case, robust

[13] Note that some of the normalized plots in Maney (2016) show atypically large differences in variance between males and females, up to about $VR = 23$. However, those plots are based on very small samples, and the extreme differences in variability they display are most likely due to sampling error.

nonparametric methods can be used to estimate the *OVL* coefficient in place of the usual formulas (Anderson et al., 2012; Schmid & Schmidt, 2006). As noted in Sect. 1.2.1, when variances/covariances are markedly unequal it is possible to use QDA instead of LDA to estimate the *PCC*; however, both models are quite sensitive to non-normality (Eisenbeis, 1977), which limits the utility of standard formulas when normality assumptions are not met.

The most widely used test of univariate normality is the Shapiro-Wilk test (Garson, 2012; Yap & Sim, 2011). Multivariate normality is harder to assess, and no single method performs well in all conditions (Mecklin & Mundfrom, 2004, 2005). Thus, the recommended approach is to combine multiple tests (which do not always agree with one another) and supplement them with graphical displays (Holgersson, 2006; Korkmaz et al., 2014; see Mecklin & Mundfrom, 2004). Levene's test is the standard procedure for comparing variances, and there are robust versions of the test that are less sensitive to non-normality (Gastwirth et al., 2009). The equality of covariance matrices is usually evaluated with Box's M test. Unfortunately, the M test suffers from a high rate of false positives (i.e., it rejects homogeneity too often) and is very sensitive to departures from multivariate normality; the latter problem can be lessened by using robust variants of the test (Anderson, 2006; O'Brien, 1992). More generally, using significance tests to evaluate assumptions is not without problems. With small samples, many tests have low power to detect violations; but when sample size is large, very small deviations from perfect normality/homogeneity may cause a test to reject the assumption, even if the practical consequences may be negligible.

In sex differences research, the phenomenon of greater male variability (complemented by some instances of greater female variability) implies that the assumption of equal variances is literally false in a majority of cases. If so, it makes little sense to perform significance tests of strict equality: If equality is not expected, a non-significant result may just mean that the test was underpowered. At the same time, sex differences in variance are relatively mild—as noted in Sect. 1.2.1, variance ratios are often lower than 1.20 and rarely higher than 1.50. Large discrepancies between male and female variances typically occur as a consequence of non-normality (e.g., skewed distributions with long tails), the presence of outliers, ceiling/floor effects, and other artifacts. With variance ratios in the usual range and approximately normal distributions, the results of the formulas in Table 1.1 are very close to the actual values even when variances differ between the sexes (with the exception of tail ratios; see below). Because equality of variances cannot be generally assumed, one can test the equality of correlation matrices (which are standardized and do not contain information on variance) instead of that of covariance matrices. This can be done with various significance tests (e.g., Jennrich, 1970; Steiger, 1980; see Revelle, 2018). However, these tests suffer from the usual problems of low sensitivity in small samples and excessive sensitivity in large samples (see above). An alternative that does not rely on significance is to compare sample correlation matrices with Tucker's *congruence coefficient* (φ or *CC*; Abdi, 2007). The *CC* coefficient in an index of matrix congruence that ranges from −1.00–1.00. Lorenzo-Seva and ten Berge (2006) proposed benchmarks for *CC*

based on expert judgments; following their recommendations, values of 0.85 or more indicate fair similarity, while values above 0.95 indicate high similarity. A high value of *CC* implies that there are no major discrepancies between the correlation matrices of males and females. In many applications, this justifies the use of multivariate indices, with the caveat that the resulting values are best regarded as reasonable approximations. Inspection of the correlation matrices (and their difference) may point to specific variables that seem to behave differently in the two sexes. Yet another strategy is to employ structural equation modeling (SEM) to fit a multigroup factor model of the variables (see below), and use model fit indices to evaluate the equivalence of correlations in the two sexes (e.g., Del Giudice et al., 2012).

While most of the standard formulas are robust to minor violations of their assumptions, this is emphatically *not* the case of tail ratios. The formulas used to estimate *TR* from effect sizes or summary statistics are very sensitive to small deviations from the hypothesized distributions, particularly when differences between groups are small and/or cutoffs are extreme (Fig. 1.4). Thus, estimates of *TR* based on standard formulas should be treated with special caution unless the underlying assumptions can be reasonably justified.

1.3.2 Biases in Effect Sizes

When they are calculated from sample data, d and D are not unbiased estimators of the corresponding population parameters but exhibit a certain amount of bias away from zero (i.e., their expected value overestimates the absolute size of the effect). Bias is typically negligible in large samples, but can be substantial in small studies; it transmits to other indices when conversion formulas are used (Table 1.1), and may lead investigators to overestimate the size of sex differences in their data. The bias in d arises from the fact that the pooled sample variance slightly underestimates the population variance, and is only an issue when sample size is very small: It amounts to less than 5% of the absolute value when the total N is $\geq$18, and less than 1% when $N \geq 78$. The bias-corrected variant of Cohen's d is known as d_u or Hedges' g; a simple correction formula is reported in Table 1.1 (see Hedges, 1981; Kelley, 2005). The bias in D is a bigger concern, because random deviations from zero in the univariate effects (caused by sampling error) add up and collectively inflate the value of D. In a previous paper (Del Giudice, 2013), I suggested a simple rule of thumb based on simulations: The bias in D can be kept to acceptable levels (i.e., less than 0.05 in absolute value) by having at least 100 cases for each variable in the analysis (e.g., $N \geq 500$ when calculating D from 5 variables). The rule works as advertised when $D \geq 0.45$, but bias can still be substantial for smaller values of D. A better alternative when N is small relative to the number of variables is to use the correction formula reported in Table 1.1, which yields the small-sample variant D_u (Lachenbruch & Mickey, 1968; Hess et al., 2007).

Capitalization on chance is also an issue with η^2, which tends to systematically overestimate the amount of variance explained. The index ω^2 (*omega squared*) provides a less biased variant of η^2 that can be useful when working with small samples (Lakens, 2013; Olejnik & Algina, 2000). More generally, multivariate methods tend to overfit the sample data, leading to overestimate both the proportion of variance they can explain and the accuracy of their predictions. This is obviously the case when standard formulas are used to estimate *PCC* from inflated values of *D* (see Glick, 1978). However, all kinds of predictive models—from logistic regression to classification trees and SVMs—tend to overfit the sample on which they are trained; to the extent that they do, their performance can be expected to drop when they are applied to a new, different set of data. Reducing overfit to improve out-of-sample predictions and obtain correct estimates of a model's performance is a major concern in the field of machine learning. Common tools employed to this end include cross-validation, regularization, and model selection based on information criteria (see Berk, 2016; Efron & Hastie, 2016; Hooten & Hobbs, 2015; James et al., 2013).

1.3.3 Measurement Error and Other Artifacts

While upward bias increases the apparent size of sex differences, measurement error has the opposite effect. When variables are measured with error, the raw difference between group means remains approximately the same but the standard deviation is inflated by noise; as a consequence, standardized indices like *d* and *D* become proportionally smaller. When measurement is unreliable, this reduction (*attenuation*) can be substantial. In classical test theory, the reliability of a measure is the proportion of variance attributable to the construct being measured ("true score variance," as contrasted with "error variance"). Assuming that sex is measured without error, the true value of *d* is attenuated by the square root of the reliability: $d = 1.00$ becomes 0.95 if the measure has 90% reliability, 0.84 with 70% reliability, and 0.71 with 50% reliability (Schmidt & Hunter, 2014; see also Schmidt & Hunter, 1996). In the case of *D*, measurement error reduces both the univariate differences and the correlations among variables; these effects may either reinforce or oppose one another depending on the correlation structure and the direction of the univariate effects. In the field of sex differences, the large majority of individual studies and meta-analyses fail to correct for attenuation due to measurement error, and as a result yield downward biased estimates of effect sizes. This is also the case of the literature syntheses compiled by Hyde (2005) and Zell et al. (2015).

There are two main approaches to correcting for measurement error. The first and simpler method is to estimate the reliability of measures from sample data, then disattenuate *d* by dividing it by the square root of the reliability coefficient. For example, consider a standardized difference $d = 0.50$ on a variable with reliability 0.77. The square root of 0.77 is 0.88, and the disattenuated *d* is $0.50/0.88 = 0.57$. To calculate *D*, both univariate effect sizes and correlations need to be disattenuated. To

disattenuate a correlation, one divides it by square root of the product of the two reliabilities. For example, consider a correlation $r = 0.30$ between two variables with reliabilities 0.77 and 0.82. The product of these reliabilities is 0.63, its square root is 0.79, and the disattenuated r is 0.30/0.79 = 0.38. While this method is an improvement over no correction at all, reliability is typically estimated with *Cronbach's alpha* (α), an index with substantial methodological limitations. In realistic conditions, α tends to yield deflated values when applied to unidimensional scales (Dunn et al., 2014; McNeish, 2018; Revelle & Condon, 2018). More worryingly, values of α do not reflect the unidimensionality of a test: If the items measure more than one construct, or tap additional specific factors on top of the general factor they are supposed to measure, α can be substantially inflated (Cortina, 1993; Crutzen & Peters, 2017; Schmitt, 1996). For other ways to estimate reliability and a review of alternative indices, see McNeish (2018), Revelle and Condon (2018), and Zinbarg et al. (2005). Also note that disattenuated effect sizes have larger sampling errors than their attenuated counterparts; this should be taken into account when calculating confidence intervals (see Schmidt & Hunter, 2014).

The second and more sophisticated approach is to use latent variable methods (most commonly SEM) to explicitly model the factor structure of the measures, and obtain estimates of sex differences on latent variables instead of observed scores (e.g., Del Giudice et al., 2012; for a different approach to factor analysis with SEM see Marsh et al., 2014). This applies to both univariate and multivariate differences. If the factor structure is correctly specified, latent variable modeling sidesteps the many problems of α and can achieve nearly error-free estimates of the underlying effects (Brown, 2015; Kline, 2016; Rhemtulla et al., 2018). Typically, SEM estimates of sex differences are notably larger than those obtained with reliability-based disattenuation. In Del Giudice et al. (2012), we examined the effect of different correction methods on the same dataset (15 personality facets in a large United States sample). With uncorrected raw scores, we obtained $D = 1.49$. Disattenuation with α raised the estimate to $D = 1.72$; fitting a multigroup SEM and calculating the effect size from latent mean differences and correlations yielded $D = 2.71$. Similarly, Mac Giolla and Kajonius (2019) calculated D on 30 facets of the Big Five, with no error correction; their average estimate across countries was $D = 1.12$. Of course, the use of SEM raises additional methodological issues, primarily that of measurement invariance between the sexes (or lack thereof; see Brown, 2015; Kline, 2016). Note that while invariance is desirable, the practical impact of statistically significant violations may be small enough to be tolerable or even negligible (especially in large samples; e.g., Schmitt et al., 2011). Nye and Drasgow (2011) developed methods to quantify the effects of measurement non-invariance at the item level and estimate its impact on observed (not latent) group differences. In presence of sizable distortions, it may still be possible to estimate latent differences by fitting a partially invariant model (Guenole & Brown, 2014; Schmitt et al., 2011). As an alternative to SEM, models based on item response theory (IRT) can also be used to estimate sex differences on latent variables (e.g., Liddell & Kruschke, 2018).

Measurement error is not the only artifact researchers should guard against. Floor and ceiling effects can severely distort measurement, and either inflate or deflate sex

differences depending on the direction of the effect, the direction of the artifact (floor vs. ceiling), and the relative variances of males and females (Wilcox, 2006; see also Liddell & Kruschke, 2018). Range restriction is another insidious artifact that occurs in a variety of research contexts: When the participants of a study are (directly or indirectly) selected from the original population on the basis of their personal characteristics, the resulting effect sizes can be substantially biased. There are several methods and formulas that attempt to correct for range restriction, though they are not without limitations (see Schmidt & Hunter, 2014; Johnson et al., 2017).

1.3.4 Meta-Analysis

Meta-analysis plays a prominent role in contemporary research on sex differences and similarities. The main function of meta-analysis is to aggregate evidence across studies, and correct for variation caused by sampling error to obtain accurate, reliable estimates of effect sizes (see Borenstein et al., 2009; Cooper et al., 2009; Schmidt & Hunter, 2014). With enough studies in the meta-analytic dataset one can examine the effect of moderators, both substantive (e.g., age of the participants) and methodological (e.g., different questionnaires or testing procedures). Standard methods of meta-analysis take individual effect sizes at face value; the main exception is the psychometric approach developed by Schmidt and Hunter (2014), which emphasizes the need to correct effect sizes for measurement error, range restriction, and other artifacts before meta-analyzing them.

Meta-analysis is a vast improvement over old-fashioned "vote counting" of significant vs. non-significant results (for an unfortunate example see Ellis et al., 2008), but it is not a panacea. As with primary research, the methodological quality of published meta-analyses is highly variable (Nakagawa et al., 2017 provide useful evaluation guidelines); the tendency to regard the results of meta-analytic studies as "definitive" should be tempered in view of the many levels of judgment involved in their design and execution. While aggregation can effectively deal with sampling error, it does nothing to correct the other artifacts reviewed in this section, which have to be deliberately addressed (see Schmidt & Hunter, 2014). Moreover, meta-analyses may overlook important moderators of a given effect, leading to a distorted picture of its size. This problem is exacerbated when the findings of multiple meta-analyses are aggregated into a "meta-synthesis." For example, Zell et al. (2015) obtained a summary effect size for each meta-analysis included in their synthesis by averaging all the effect sizes reported in the same meta-analysis. At this level of aggregation, the risk of obtaining meaningless results increases dramatically, especially when estimates that pertain to widely different variables and domains are pooled into a single model. Moreover, the magnitude of sex differences in some domains can be drastically underestimated if effect sizes that would be best aggregated with multivariate methods (Sect. 1.2.1) are simply averaged together.

A persistent problem in meta-analysis is the distorting influence of publication and reporting bias. Low statistical power (primarily due to small sample size) and

selective reporting of results can generate large amounts of statistically significant "false positives"; more troubling from the standpoint of meta-analysis, even the true effects in the published literature are likely to be systematically inflated (Ioannidis, 2005, 2008a). For example, a recent study found evidence indicating positive reporting bias in brain imaging studies of sex differences (David et al., 2018). Note that the same statistical and methodological factors can also promote "reverse bias" (the selective non-publication, non-reporting, and deflation of effect sizes) if certain findings go against the social/ideological preferences of the field (Ioannidis, 2005, 2008a; see also Coburn & Vevea, 2015). This is not an unreasonable concern, if one considers that claims of large sex differences in psychology are often denounced as dangerous and socially harmful (e.g., Fine, 2010; Hyde, 2005; Reis & Carothers, 2014). In principle, several methods can be used to detect publication and/or reporting bias in meta-analytic datasets (Jin et al., 2015). Unfortunately, the standard tests are easy to misapply, suffer from high rates of false negatives unless the dataset includes a large number of studies, and may mistake other sources of heterogeneity for evidence of bias (Ioannidis, 2008b; Ioannidis & Trikalinos, 2007; Jin et al., 2015). Thus, common tests of bias can be meaningfully applied only in the relatively few cases in which effect sizes are fairly homogeneous across studies (Ioannidis & Trikalinos, 2007).

In recent years, standard procedures based on the distribution of effect sizes have been joined by *p-curve* and *p-uniform* analyses, two methods that rely on the distribution of significant *p* values in a set of studies to detect selective publication and/or reporting (Simonsohn et al., 2014a, 2015; van Assen et al., 2015). The same methods can be used to estimate the average effect size of a set of studies from their significant *p* values (Simonsohn et al., 2014b; van Assen et al., 2015), thus complementing standard meta-analytic techniques. However, both *p*-curve and *p*-uniform may overestimate the population effect when studies are highly heterogeneous (van Aert et al., 2016). There are also some concerns about the validity of *p*-curve methods in non-experimental research, when changes in significance may depend on the selective inclusion of covariates in the analysis (see Bruns & Ioannidis, 2016).

1.4 Conclusion

In concluding this chapter it may be useful to point out that, important as it is, successful quantification is only the beginning of understanding. Research on sex differences and similarities relies on an exceptionally rich toolkit of methods, ranging from experimental studies to developmental, cross-cultural, and even comparative research across species. Together, these methods can be used to understand how sex differences in various domains vary systematically across contexts, and what are the main factors that reduce or amplify them. At a deeper level, an emphasis on measurement should not blind investigators to the possibility that males and females may differ in *qualitative* rather than purely quantitative ways. For example,

the same traits may be influenced by different causal factors in the two sexes, or predict different patterns of outcomes. If multiple sexually differentiated traits interact with each other in complex patterns, they may give rise to configural or "gestalt" effects that are not well captured by their linear combination (as implicitly assumed by *D* or discriminant analysis). Other nonlinear relations between traits and outcomes (e.g., threshold effects) may turn graded quantitative differences into discrete transitions. In some cases, males and females may possess different psychological specializations that follow qualitatively different rules of operation. No doubt, the study of sex differences and similarities will remain an exciting enterprise for a long time to come, and it is easy to predict that high-quality measurement will play an ever more central role in the future of the field.

Acknowledgments I am grateful to Drew Bailey, Mike Bailey, Alex Byrne, Tom Booth, Doug VanderLaan, and Ivy Wong for their many thoughtful comments on earlier drafts of this chapter.

References

Abdi, H. (2007). RV coefficient and congruence coefficient. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 849–853). Sage.

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*, 129–133.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10*, 317–328.

Anderson, G., Linton, O., & Whang, Y. J. (2012). Nonparametric estimation and inference about the overlap of two distributions. *Journal of Econometrics, 171*, 1–23.

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics, 62*, 245–253.

Anderson, N. E., Harenski, K. A., Harenski, C. L., Koenigs, M. R., Decety, J., Calhoun, V. D., & Kiehl, K. A. (2018). Machine learning of brain gray matter differentiates sex in a large forensic sample. *Human Brain Mapping, 40*, 1496–1506.

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences, 41*, 39–48.

Beauchaine, T. P. (2007). A brief taxometrics primer. *Journal of Clinical Child and Adolescent Psychology, 36*, 654–676.

Bentley, M. (1945). Sanity and hazard in childhood. *American Journal of Psychology, 58*, 212–246.

Berk, R. A. (2016). *Statistical learning from a regression perspective* (2nd ed.). Springer.

Bessudnov, A., & Makarov, A. (2015). School context and gender differences in mathematical performance among school graduates in Russia. *International Studies in Sociology of Education, 25*, 63–81.

Blackless, M., Charuvastra, A., Derryck, A., Fausto-Sterling, A., Lauzanne, K., & Lee, E. (2000). How sexually dimorphic are we? Review and synthesis. *American Journal of Human Biology, 12*, 151–166.

Blakemore, J. E. O., Berenbaum, S., & Liben, L. S. (2009). *Gender development*. Psychology Press.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bradley, E. L. (2006). Overlapping coefficient. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (p. 1900). Wiley.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29*, 79–97.
Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford.
Bruce, V. A., Burton, M., Hanna, E., Healey, P., Mason, O., Coombes, A., . . . Linney, A. (1993). Sex discrimination: How well do we tell the difference between male and female faces? *Perception, 22*, 131–152.
Bruns, S. B., & Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. *PLoS One, 11*, e0149144. https://doi.org/10.1371/journal.pone.0149144
Buss, D. M. (1995). Psychological sex differences: Origins through sexual selection. *American Psychologist, 50*, 164–171.
Carothers, B. J., & Reis, H. T. (2013). Men and women are from earth: Examining the latent structure of gender. *Journal of Personality and Social Psychology, 10*, 385–407.
Chekroud, A. M., Ward, E. J., Rosenberg, M. D., & Holmes, A. J. (2016). Patterns in the human brain mosaic discriminate males from females. *Proceedings of the National Academy of Sciences, 113*, E1968–E1968. https://doi.org/10.1073/pnas.1523888113
Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods, 20*, 310–330.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
Conroy-Beam, D., Buss, D. M., Pham, M. N., & Shackelford, T. K. (2015). How sexually dimorphic are human mate preferences? *Personality and Social Psychology Bulletin, 41*, 1082–1093.
Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.
Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.
Crutzen, R., & Peters, G. J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review, 11*, 242–247.
David, S. P., Naudet, F., Laude, J., Radua, J., Fusar-Poli, P., Chu, I., . . . Ioannidis, J. P. (2018). Potential reporting bias in neuroimaging studies of sex differences. *Scientific Reports, 8*, 6082.
Davies, A. P. C., & Shackelford, T. K. (2008). Two human natures: How men and women evolved different psychologies. In C. Crawford & D. Krebs (Eds.), *Foundations of evolutionary psychology* (pp. 261–280). Erlbaum.
Deaux, K. (1985). Sex and gender. *Annual Review of Psychology, 36*, 49–81.
Del Giudice, M. (2009). On the real magnitude of psychological sex differences. *Evolutionary Psychology, 7*, 264–279.
Del Giudice, M. (2011). Sex differences in romantic attachment: A meta-analysis. *Personality and Social Psychology Bulletin, 37*, 193–214.
Del Giudice, M. (2013). Multivariate misgivings: Is *D* a valid measure of group and sex differences? *Evolutionary Psychology, 11*, 1067–1076.
Del Giudice, M. (2015). Gender differences in personality and social behavior. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed., pp. 750–756). Elsevier.
Del Giudice, M. (2017). Heterogeneity coefficients for Mahalanobis' *D* as a multivariate effect size. *Multivariate Behavioral Research, 52*, 216–221.
Del Giudice, M. (2018). Addendum to: Heterogeneity coefficients for Mahalanobis' *D* as a multivariate effect size. *Multivariate Behavioral Research, 53*, 571–357.
Del Giudice, M., Angeleri, R., Brizio, A., & Elena, M. R. (2010). The evolution of autistic-like and schizotypal traits: A sexual selection hypothesis. *Frontiers in Psychology, 1*, 41.
Del Giudice, M., Barrett, E. S., Belsky, J., Hartman, S., Martel, M. M., Sangenstedt, S., & Kuzawa, C. W. (2018). Individual differences in developmental plasticity: A role for early androgens? *Psychoneuroendocrinology, 90*, 165–173.

Del Giudice, M., Booth, T., & Irwing, P. (2012). The distance between Mars and Venus: Measuring global sex differences in personality. *PLoS One, 7*, e29265. https://doi.org/10.1371/journal.pone.0029265

Del Giudice, M., Klimczuk, A. C. E., Traficonte, D. M., & Maestripieri, D. (2014). Autistic-like and schizotypal traits in a life history perspective: Diametrical associations with impulsivity, sensation seeking, and sociosexual behavior. *Evolution and Human Behavior, 35*, 415–424.

Del Giudice, M., Lippa, R. A., Puts, D. A., Bailey, D. H., Bailey, J. M., & Schmitt, D. P. (2015). *Mosaic brains? A methodological critique of Joel et al. (2015).* https://doi.org/10.13140/RG.2.1.1038.8566.

Del Giudice, M., Lippa, R. A., Puts, D. A., Bailey, D. H., Bailey, J. M., & Schmitt, D. P. (2016). Joel et al.'s method systematically fails to detect large, consistent sex differences. *Proceedings of the National Academy of Sciences USA, 113*, E1965–E1965.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78–89.

Dunn, O. J., & Varady, P. D. (1966). Probabilities of correct classification in discriminant analysis. *Biometrics, 22*, 908–924.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*, 399–412.

Dykiert, D., Gale, C. R., & Deary, I. J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence, 37*, 42–47.

Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science, 8*, 340–357.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University press.

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *Journal of Finance, 32*, 875–900.

Ellis, L. (2011). Identifying and explaining apparent universal sex differences in cognition and behavior. *Personality and Individual Differences, 51*, 552–561.

Ellis, L., Hershberger, S., Field, E., Wersinger, S., Pellis, S., Geary, D., … Karadi, K. (2008). *Sex differences: Summarizing more than a century of scientific research*. Psychology Press.

Fausto-Sterling, A. (2012). *Sex/gender: Biology in a social world*. Routledge.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62*, 61–84.

Fine, C. (2010). *Delusions of gender: How our minds, society, and neurosexism create difference*. Norton.

Furlow, C. F., & Beretvas, S. N. (2005). Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. *Psychological Methods, 10*, 227–254.

Gangestad, S. W., Bailey, J. M., & Martin, N. G. (2000). Taxometric analyses of sexual orientation and gender identity. *Journal of Personality and Social Psychology, 78*, 1109–1121.

Garson, G. D. (2012). *Testing statistical assumptions*. Statistical Associates Publishing.

Garthwaite, P. H., & Koch, I. (2016). Evaluating the contributions of individual variables to a quadratic form. *Australian & New Zealand Journal of Statistics, 58*, 99–119.

Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science, 24*, 343–360.

Geary, D. C. (2010). *Male, female: The evolution of human sex differences* (2nd ed.). American Psychological Association.

Geary, D. C. (2015). *Evolution of vulnerability: Implications for sex differences in health and development*. Academic Press.

Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition, 10*, 211–222.

Grice, J. W., & Barrett, P. T. (2014). A note on Cohen's overlapping proportions of normal distributions. *Psychological Reports, 115*, 741–747.
Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology, 5*, 980.
Haig, D. (2004). The inexorable rise of gender and the decline of sex: Social change in academic titles, 1945–2001. *Archives of Sexual Behavior, 33*, 87–96.
Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*, 1–51.
Handcock, M. S., & Janssen, P. L. (2002). Statistical inference for the relative density. *Sociological Methods & Research, 30*, 394–424.
Handcock, M. S., & Morris, M. (1998). Relative distribution methods. *Sociological Methodology, 28*, 53–97.
Handcock, M. S., & Morris, M. (1999). *Relative distribution methods in the social sciences*. Springer.
Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.
Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research, 63*, 94–105.
Helgeson, V. S. (2016). *Psychology of gender* (5th ed.). Routledge.
Hennessy, R. J., McLearie, S., Kinsella, A., & Waddington, J. L. (2005). Facial surface analysis by 3D laser scanning and geometric morphometrics in relation to sexual dimorphism in cerebral–craniofacial morphogenesis and cognitive function. *Journal of Anatomy, 207*, 283–295.
Hess, M. R., Hogarty, K. Y., Ferron, J. M., & Kromrey, J. D. (2007). Interval estimates of multivariate effect sizes: Coverage and interval width estimates under variance heterogeneity and nonnormality. *Educational and Psychological Measurement, 67*, 21–40.
Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172–177.
Holgersson, H. E. T. (2006). A graphical method for assessing multivariate normality. *Computational Statistics, 21*, 141–149.
Holzleitner, I. J., Hunter, D. W., Tiddeman, B. P., Seck, A., Re, D. E., & Perrett, D. I. (2014). Men's facial masculinity: When (body) size matters. *Perception, 43*, 1191–1202.
Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs, 85*, 3–28.
Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*, 227–240.
Huberty, C. J. (2005). Mahalanobis distance. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1110–1111). Wiley.
Hull, C. L. (2003). Letter to the editor: How sexually dimorphic are we? Review and synthesis. *American Journal of Human Biology, 15*, 112–116.
Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581–592.
Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*, 373–398.
Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist, 74*, 171–193.
Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*, 494–495.
Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124. https://doi.org/10.1371/journal.pmed.0020124
Ioannidis, J. P. (2008a). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648.
Ioannidis, J. P. (2008b). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice, 14*, 951–957.

Ioannidis, J. P., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal, 176*, 1091–1096.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.
Janicke, T., Häderer, I. K., Lajeunesse, M. J., & Anthes, N. (2016). Darwinian sex roles confirmed across the animal kingdom. *Science Advances, 2*, e1500983. https://doi.org/10.1126/sciadv.1500983
Janssen, D. F. (2018). Know thy gender: Ethymological primer. *Archives of Sexual Behavior, 47*, 2149–2154.
Jennrich, R. I. (1970). An asymptotic χ^2 test for the equality of two correlation matrices. *Journal of the American Statistical Association, 65*, 904–912.
Jin, Z. C., Zhou, X. H., & He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine, 34*, 343–360.
Joel, D. (2012). Genetic-gonadal-genitals sex (3G-sex) and the misconception of brain and gender, or, why 3G-males and 3G-females have intersex brain and intersex gender. *Biology of Sex Differences, 3*, 27.
Joel, D., Berman, Z., Tavor, I., Wexler, N., Gaber, O., Stein, Y., . . . Liem, F. (2015). Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences USA, 112*, 15468–15473.
Joel, D., & Fausto-Sterling, A. (2016). Beyond sex differences: New approaches for thinking about variation in brain structure and function. *Philosophical Transaction of the Royal Society of London B, 371*, 20150451.
Joel, D., Persico, A., Hänggi, J., Pool, J., & Berman, Z. (2016). Reply to Del Giudice et al., Chekroud et al., and Rosenblatt: Do brains of females and males belong to two distinct populations? *Proceedings of the National Academy of Sciences USA, 113*, E1969–E1970.
Joel, D., Persico, A., Salhov, M., Berman, Z., Oligschlager, S., Meilijson, I., & Averbuch, A. (2018). Analysis of human brain structure reveals that the brain 'types' typical of males are also typical of females, and vice versa. *Frontiers in Human Neuroscience, 12*, 399.
Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3*, 518–531.
Johnson, W., Deary, I. J., & Bouchard, T. J., Jr. (2017). Have standard formulas correcting correlations for range restriction been adequately tested? Minor sampling distribution quirks distort them. *Educational and Psychological Measurement, 78*, 1021–1055.
Jordan-Young, R., & Rumiati, R. I. (2012). Hardwired for sexism? Approaches to sex/gender in neuroscience. *Neuroethics, 5*, 305–315.
Kaiser, T., Del Giudice, M., & Booth, T. (2020). Global sex differences in personality: Replication with an open online dataset. *Journal of Personality, 88*, 415–429.
Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement, 65*, 51–69.
Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*, 1–24.
Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods, 45*, 905–927.
Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford.
Kodric-Brown, A., & Brown, J. H. (1987). Anisogamy, sexual selection, and the evolution and maintenance of sex. *Evolutionary Ecology, 1*, 95–105.
Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal, 6*, 151–162.
Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*, 155–177.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10*, 1–11.
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.
Lakin, J. M. (2013). Sex differences in reasoning abilities: Surprising evidence that male-female ratios in the tails of the quantitative reasoning distribution have increased. *Intelligence, 41*, 263–274.
Lee, K., & Ashton, M. C. (2020). Sex differences in HEXACO personality characteristics across countries and ethnicities. *Journal of Personality, 88*, 1075–1090.
Lehre, A. C., Lehre, K. P., Laake, P., & Danbolt, N. C. (2009). Greater intrasex phenotype variability in males than in females is a fundamental aspect of the gender differences in humans. *Developmental Psychobiology, 51*, 198–206.
Lehtonen, J., & Kokko, H. (2011). Two roads to two sexes: Unifying gamete competition and gamete limitation in a single model of anisogamy evolution. *Behavioral Ecology and Sociobiology, 65*, 445–459.
Lehtonen, J., & Parker, G. A. (2014). Gamete competition, gamete limitation, and the evolution of the two sexes. *Molecular Human Reproduction, 20*, 1161–1168.
Lehtonen, J., Parker, G. A., & Schärer, L. (2016). Why anisogamy drives ancestral sex roles. *Evolution, 70*, 1129–1135.
Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348.
Limpert, E., & Stahel, W. A. (2011). Problems with using the normal distribution–and ways to improve quality and efficiency of data analysis. *PLoS One, 6*, e21403. https://doi.org/10.1371/journal.pone.0021403
Lippa, R. A. (2001). On deconstructing and reconstructing masculinity–femininity. *Journal of Research in Personality, 35*, 168–207.
Lippa, R. A. (2005). *Gender, nature, and nurture* (2nd ed.). Lawrence Erlbaum Associates.
Lippa, R. A. (2009). Sex differences in sex drive, sociosexuality, and height across 53 nations: Testing evolutionary and social structural theories. *Archives of Sexual Behavior, 38*, 631–651.
Lippa, R. A. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: Testing evolutionary and social-environmental theories. *Archives of Sexual Behavior, 39*, 619–636.
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.
Lorenzo-Seva, U., & ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*, 57–64.
Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology, 54*, 705–711.
Maney, D. L. (2016). Perils and pitfalls of reporting sex differences. *Philosophical Transactions of the Royal Society B, 371*, 20150119.
Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85–110.
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361–365.
McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412–433.
Mecklin, C. J., & Mundfrom, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review, 72*, 123–138.
Mecklin, C. J., & Mundfrom, D. J. (2005). A Monte Carlo comparison of the type I and type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation, 75*, 93–107.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50*, 266–275.
Mitteroecker, P., Windhager, S., Müller, G. B., & Schaefer, K. (2015). The morphometrics of "masculinity" in human faces. *PLoS One, 10*, e0118374. https://doi.org/10.1371/journal.pone.0118374
Money, J. (1955). Hermaphroditism, gender and precocity in hyperadrenocorticism: Psychologic findings. *Bulletin of the Johns Hopkins Hospital, 96*, 253–264.
Morris, M. L. (2016). Vocational interests in the United States: Sex, age, ethnicity, and year effects. *Journal of Counseling Psychology, 63*, 604–615.
Nakagawa, S., Noble, D. W., Senior, A. M., & Lagisz, M. (2017). Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC Biology, 15*, 18.
Norris, A. L., Marcus, D. K., & Green, B. A. (2015). Homosexuality as a discrete class. *Psychological Science, 26*, 1843–1853.
Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*, 966–980.
Oakley, A. (1972). *Sex, gender, and society*. Harper Colophon.
O'Brien, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics, 48*, 819–827.
Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241–286.
Phillips, O. R., Onopa, A. K., Hsu, V., Ollila, H. M., Hillary, R. P., Hallmayer, J., . . . Singh, M. K. (2018). Beyond a binary classification of sex: An examination of brain sex differentiation, psychopathology, and genotype. *Journal of the American Academy of Child & Adolescent Psychiatry, 58*, 787–798.
Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160–164.
Reiser, B. (2001). Confidence intervals for the Mahalanobis distance. *Communications in Statistics: Simulation and Computation, 30*, 37–45.
Reis, H. T., & Carothers, B. J. (2014). Black and white or shades of gray: Are gender differences categorical or dimensional? *Current Directions in Psychological Science, 23*, 19–26.
Revelle, W. (2018). *An introduction to psychometric theory with applications in R*. Retrieved on October 24, 2018 from the personality project website http://personality-project.org/r/book/
Revelle, W., & Condon, D. M. (2018). Reliability. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 709–749). Wiley.
Rhemtulla, M., van Bork, R., & Borsboom, D. (2018). *Worse than measurement error: Consequences of inappropriate latent variable measurement models*. Retrieved on October 24, 2018 from the Open Science framework website https://osf.io/27bxg/
Rhodes, G., Jeffery, L., Watson, T. L., Jaquet, E., Winkler, C., & Clifford, C. W. G. (2004). Orientation-contingent face aftereffects and implications for face-coding mechanisms. *Current Biology, 14*, 2119–2123.
Rippon, G., Jordan-Young, R., Kaiser, A., & Fine, C. (2014). Recommendations for sex/gender neuroimaging research: Key principles and implications for research design, analysis, and interpretation. *Frontiers in Human Neuroscience, 8*, 650.
Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., . . . Liewald, D. C. (2018). Sex differences in the adult human brain: Evidence from 5216 UK biobank participants. *Cerebral Cortex, 28*, 2959–2975.
Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology, 9*, 395–396.
Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2017). Beyond differences in means: Robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience, 46*, 1738–1748.

Ruscio, J., Haslam, N., & Ruscio, A. M. (2013). *Introduction to the taxometric method: A practical guide*. Routledge.
Ruscio, J., Ruscio, A. M., & Carney, L. M. (2011). Performing taxometric analysis to distinguish categorical and dimensional variables. *Journal of Experimental Psychopathology, 2*, 170–196.
Sanchis-Segura, C., Ibañez-Gual, M. V., Aguirre, N., Cruz-Gómez, Á. J., & Forn, C. (2020). Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. *Scientific Reports, 10*, 12953. https://doi.org/10.1038/s41598-020-69361-9
Sapp, M., Obiakor, F. E., Gregas, A. J., & Scholze, S. (2007). Mahalanobis distance: A multivariate measure of effect in hypnosis research. *Sleep and Hypnosis, 9*, 67–70.
Sax, L. (2002). How common is intersex? A response to Anne Fausto-Sterling. *Journal of Sex Research, 39*, 174–178.
Schärer, L., Rowe, L., & Arnqvist, G. (2012). Anisogamy, chance and the evolution of sex roles. *Trends in Ecology & Evolution, 27*, 260–264.
Schmid, F., & Schmidt, A. (2006). Nonparametric estimation of the coefficient of overlapping—Theory and empirical application. *Computational Statistics & Data Analysis, 50*, 1583–1596.
Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.
Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Sage.
Schmitt, D. P. (2015). The evolution of culturally-variable sex differences: Men and women are not always different, but when they are… it appears not to result from patriarchy or sex role socialization. In T. K. Shackelford & R. D. Hansen (Eds.), *The evolution of sexuality* (pp. 221–256). Springer.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.
Schmitt, N., Golubovich, J., & Leong, F. T. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using big five and RIASEC measures. *Assessment, 18*, 412–427.
Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH image to ImageJ: 25 years of image analysis. *Nature Methods, 9*, 671–675.
Sen Gupta, A. (2004). Generalized variance. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (p. 6053). Wiley. https://doi.org/10.1002/0471667196.ess6053
Sepehrband, F., Lynch, K. M., Cabeen, R. P., Gonzalez-Zacarias, C., Zhao, L., D'arcy, M., Kesselman, C., Herting, M. M., Dinov, I. D., Toga, A. W., & Clark, K. A. (2018). Neuroanatomical morphometric characterization of sex differences in youth using statistical learning. *NeuroImage, 172*, 217–227.
Shaffer, J. P. (1992). Caution on the use of variance ratios: A comment. *Review of Educational Research, 62*, 429–432.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534–547.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666–681.
Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking. A reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*, 1146–1152.
Skiena, S. S. (2017). *The data science design manual*. Springer.
Steiger, J. H. (1980). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. *Multivariate Behavioral Research, 15*, 335–352.
Stewart-Williams, S., & Thomas, A. G. (2013). The ape that thought it was a peacock: Does evolutionary psychology exaggerate human sex differences? *Psychological Inquiry, 24*, 137–168.

Stoller, R. J. (1968). *Sex and gender: The development of masculinity and femininity*. Science House.
Taborsky, M., & Brockmann, H. J. (2010). Alternative reproductive tactics and life history phenotypes. In P. Kappeler (Ed.), *Animal behavior: Evolution and mechanisms* (pp. 537–586). Springer.
Unger, R. K. (1979). Toward a redefinition of sex and gender. *American Psychologist, 34*, 1085–1094.
Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*, 473–481.
van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*, 713–729.
van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*, 293–309.
van Putten, M. J., Olbrich, S., & Arns, M. (2018). Predicting sex from brain rhythms with deep learning. *Scientific Reports, 8*, 3069.
Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*, 101–132.
Verweij, K. J., Mosing, M. A., Ullén, F., & Madison, G. (2016). Individual differences in personality masculinity-femininity: Examining the effects of genes, environment, and prenatal hormone transfer. *Twin Research and Human Genetics, 19*, 87–96.
Voracek, M., Mohr, E., & Hagmann, M. (2013). On the importance of tail ratios for psychological science. *Psychological Reports, 112*, 872–886.
Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57.
Wai, J., Hodges, J., & Makel, M. C. (2018). Sex differences in ability tilt in the right tail of cognitive abilities: A 35-year examination. *Intelligence, 67*, 76–83.
Wierenga, L. M., Sexton, J. A., Laake, P., Giedd, J. N., Tamnes, C. K., & Pediatric Imaging, Neurocognition, and Genetics Study. (2017). A key characteristic of sex differences in the developing brain: Greater variability in brain structure of boys than girls. *Cerebral Cortex, 28*, 2741–2751.
Wilcox, R. R. (2006). Graphical methods for assessing effect size: Some alternatives to Cohen's *d*. *Journal of Experimental Education, 74*, 351–367.
Wyman, M. J., & Rowe, L. (2014). Male bias in distributions of additive genetic, residual, and phenotypic variances of shared traits. *The American Naturalist, 184*, 326–337.
Xin, J., Zhang, Y., Tang, Y., & Yang, Y. (2019). Brain differences between men and women: Evidence from deep learning. *Frontiers in Neuroscience, 13*, 185. https://doi.org/10.3389/fnins.2019.00185
Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*, 2141–2155.
Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist, 70*, 10–20.
Zhou, W. (2008). Statistical inference for P(X<Y). *Statistics in Medicine, 27*, 257–279.
Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133.
Zou, G. Y. (2007). Exact confidence interval for Cohen's effect size is readily available. *Statistics in Medicine, 26*, 3054–3056.