



Understanding the Magnitude of Psychological Differences Between Women and Men Requires Seeing the Forest and the Trees

Alice H. Eagly  and William Revelle 

Department of Psychology, Northwestern University

Abstract

Whether women and men are psychologically very similar or quite different is a contentious issue in psychological science. This article clarifies this issue by demonstrating that larger and smaller sex/gender differences can reflect differing ways of organizing the same data. For single psychological constructs, larger differences emerge from averaging multiple indicators that differ by sex/gender to produce scales of a construct's overall typicality for women versus men. For example, averaging self-ratings on personality traits more typical of women or men yields much larger sex/gender differences on measures of the femininity and masculinity of personality. Sex/gender differences on such broad-gauge, thematic variables are large relative to differences on their component indicators. This increased effect magnitude for aggregated scales reflects gains in both their reliability and validity as indicators of sex/gender. In addition, in psychological domains such as vocational interests that are composed of many variables, at least some of which differ by sex/gender, the multivariate distance between women and men is typically larger than the differences on the component variables. These analyses encourage recognition of the interdependence of sex/gender similarity and difference in psychological data.

Keywords

sex differences, gender differences, effect magnitude, gender-similarities hypothesis, femininity, masculinity

The magnitude of differences between women and men is an unsettled issue in psychological science. One prominent claim, known as the *gender-similarities hypothesis*, is that women and men are very similar on most psychological variables (Hyde, 2005, 2014). However, other experts maintain that, on the contrary, larger differences are common even in fundamental areas of human functioning such as personality (e.g., Archer, 2019; Lippa, 2005).

Such inconsistencies in scientific claims can be puzzling because they all rely on quantitative analyses of a large, shared database of psychological research. It is tempting to believe that either similarity or difference is the truth of the matter and to reject the alternative claim. Instead, those who care about this issue should pause and think more deeply. To further such understanding, this article takes crucial, but neglected, psychometric considerations into account. We show that claims of similarity and difference are both valid but

can reflect differing ways of organizing the same data, and thus, metaphorically, they can be two sides of the same psychometric coin.

As a first step, given the lack of consensus among scientists about terminology pertaining to sex and gender, we define our key terms. Difficulties follow from the common view that the term *sex* pertains to biology and *gender* to socialization and culture—that is, sex is to gender as nature is to nurture. However, identifying differences as biologically influenced and/or socially constructed is a work in progress as scientists seek to understand how nature and nurture influence the psychology of women and men. Therefore, to avoid

Corresponding Authors:

Alice H. Eagly, Department of Psychology, Northwestern University
Email: eagly@northwestern.edu

William Revelle, Department of Psychology, Northwestern University
Email: revele@northwestern.edu

prejudging causality, we label differences by the hybrid neologisms of gender/sex (Schudson et al., 2019) and sex/gender (Fausto-Sterling, 2012) and apply these terms interchangeably. Finally, we define the words *feminine* and *femininity* and *masculine* and *masculinity* as referring to human attributes, including traits and behaviors, that are more typical of women or men, respectively.

A large quantity of psychological research has addressed gender/sex differences and similarities. By way of evidence, the American Psychological Association's PsycINFO database (<http://psycnet.apa.org>) includes 89,415 journal articles published from 1950 through 2021 that focus on these comparisons, as indicated by the assignment of the index term *human sex differences* to them. Among these articles are 764 reporting meta-analyses that reviewed portions of the research literature. In addition, many surveys have assessed psychological sex/gender differences, often with representative samples, as have many testing programs, conducted mainly in educational and counseling settings. Findings from these varied sources undergird the claims discussed in this article.¹

The elusiveness of scientific consensus about the magnitude of sex/gender differences is not surprising given the challenges of organizing this massive amount of empirical data. To address this issue, we argue that one important contributor to a lack of consensus is insufficient consideration of relevant psychometric principles. We build our case by first reviewing experts' generalizations about the magnitude of gender/sex differences. Then we explain the psychometric principles by which the aggregation of data underlying effect sizes contributes to their magnitude. This analysis explains why psychological differences between women and men are often large on broad sex/gender-relevant psychological variables and simultaneously small on more narrowly defined indicators of such variables. Our analysis proceeds to consider the overall distance between men and women in psychological domains such as personality traits that are composed of many variables. Finally, the analysis concludes with reflections on the relevance of sex/gender differences and similarities to discourse on diversity in groups and organizations.

In this article, we assiduously avoid discussing particular causal explanations of gender/sex differences and similarities or indicating our personal preferences for any theories about causation. We encourage readers to apply their own interpretations to the patterns of similarity and difference that our analyses reveal. Some will emphasize the dependence of most psychological measures on self-report and the shaping of such responding and of masculinity and femininity more

generally by social and cultural forces. Others will prefer to ascribe the findings we present to essential causes embedded in the evolution of women and men in the human species. Such dissimilar interpretations should inspire scientific research that compares and contrasts predictions from these and other theoretical positions. However, we have no such purpose in this article but instead strive to provide readers with a clearer view of the phenomena that require explanation.

Claims of Sex/Gender Similarities and Differences

The small magnitude of most sex/gender differences is the central theme of Hyde's (2005) pioneering review of 46 meta-analyses, which yielded 128 effect sizes represented as standardized mean differences (i.e., the d statistic; Cohen, 1988); 48% were classified as small ($d = |0.11|$ to $|0.35|$), and 30% were classified as very small ($d = |0.10|$ or smaller). The largest effect size, aside from sexuality or motor behaviors such as throw velocity, was $d = -0.91$ for tender-mindedness, a facet of the personality trait of agreeableness.² Zell et al. (2015) seconded Hyde's similarity conclusion in an expanded project that encompassed 106 meta-analyses reporting 386 effect sizes, which yielded 106 study-level effect sizes. Zell et al.'s mean effect size was $d = |0.21|$ ($SD = 0.14$); 46% were classified as small, and 39% were classified as very small. The largest effect size, $d = |0.73|$, was for masculine versus feminine personality traits.

A subsequent review of sex/gender differences by Archer (2019) encompassed 131 meta-analyses and 89 estimates from other sources, summarized as 146 effect sizes. These other sources consisted of "(i) cross-national surveys of personality traits, social attributes, mate choice and sexuality; (ii) large-sample ($N > 1000$) online studies; (iii) social surveys on attributes related to health and crime; and (iv) crime statistics" (Archer, 2019, p. 1383). The mean effect size was larger, $d = |0.43|$ ($SD = 0.40$); only 24% were classified as small, and 16% were classified as very small. The largest consisted of 12 effect sizes greater than $d = 1.00$, considered very large, many of which pertained to crime and violence (e.g., $d = 2.54$ for same-sex homicide).

Mirroring these contrasting presentations, the conclusions that writers of textbooks on the psychology of sex and gender have offered also have differed considerably. Some authors have emphasized gender/sex similarity (e.g., Bosson et al., 2019), whereas others have instead pointed to the presence of larger differences (e.g., Lippa, 2005).

To shed some light on these issues, we offer psychometric analyses that can promote understanding of the magnitudes of gender/sex differences and similarities.

This analysis includes examples from relevant research but does not provide a general review of psychological differences and similarities or evaluate their causes.

Aggregation of Sex/Gender Differences on Single Dimensions

Research on attitudes and personality traits provides an informative precedent for understanding the magnitude of sex/gender differences. In the 1960s, the relations between psychological dispositions and relevant behaviors gained prominence when Wicker (1969), for attitudes, and Mischel (1968), for personality traits, showed that these relations were typically very weak. Challenges to these initially quite shocking conclusions soon emerged. Demonstrating the importance of aggregation to the magnitude of effects, Fishbein and Ajzen (1974) showed that, despite most specific behaviors' weak relations to relevant attitudes, aggregations of such behaviors related strongly. Likewise, Epstein (1979, 1980) showed that, despite most specific behaviors' weak relations to relevant personality traits, aggregations of such behaviors related strongly. Thus, both attitudes and personality traits do relate strongly to general themes of relevant behaviors, such as when people high in religiosity engage in more religious behaviors than those who are low in this attitude.

Aggregations of behaviors that successfully predict attitudes or personality traits could be cumulated over occasions, contexts, or differing disposition-relevant behaviors. On the basis of such findings, a consensus emerged that behaviors relate more strongly to psychological dispositions to the extent that the behaviors match the generality or scope of a dispositional criterion (Ajzen & Fishbein, 1977; Epstein, 1983).

These demonstrations of the effects of aggregation did not reveal the psychometric principles underlying the improvement in prediction because a predictor encompasses a greater number and breadth of relevant components. Most discussions implicated the venerable psychometric principle that assessments of variables are more reliable when based on a greater number of valid indicators (Spearman, 1904). However, adequate understanding requires considering both the validity and reliability of an aggregated predictor. For predicting sex/gender, for example, aggregating items from a domain, with every item having some validity for predicting sex/gender, increases the reliability of the predictor scale, which asymptotically tends toward 1.00. The validity of a scale as a predictor of sex/gender also increases with the aggregation of items but is limited by the items' average interitem correlation. Specifically, the validity of a scale asymptotically tends toward the average item validity divided

by the square root of the average of the interitem correlations. For a fixed-average item validity, scale validity is a positive function of the number of items and is higher the lower the correlations between the items within the scale (see Appendix for derivation).

What are the implications of these insights for the magnitude of sex/gender differences? In general, aggregates of relevant items should more strongly predict sex/gender than single items in all domains of psychological functioning that differ between men and women. If the items of an aggregated predictor are from the same domain and thus tend to be highly intercorrelated, the validity of the composite predictor increases with the number of items but not as strongly as it would if the items came from different domains and thus were more weakly intercorrelated. This insight thus reveals the principle underlying earlier researchers' claims that dispositions are more strongly predicted by aggregations of items that are not only more numerous but also more diverse in content (Ajzen & Fishbein, 1977; Epstein, 1983). The diversity of a predictor increases as items encompass differing relevant domains and thus tend to be less highly intercorrelated. The implication of these principles for gender/sex research, in plain language, is that women and men differ more on psychological variables that are broadly construed, or thematic, than on more narrowly constituted variables.

A classic example of a large and broad multi-item composite variable relating strongly to gender/sex derives from research by Terman and Miles (1936) that aggregated a very wide range of questionnaire items that differentiated women and men. Their efforts yielded a 456-item masculinity–femininity scale that yielded a gender/sex effect size of $d = 2.53$ in a sample of 696 female and 604 male participants (Terman & Miles, 1936, p. 72). The large magnitude of this gender/sex difference is consistent with the principle that including many relevant items and drawing them from differing domains greatly increase the prediction of gender/sex.

Despite this excellent prediction, Terman and Miles were met with criticisms that their masculinity–femininity scale encompassed a hodgepodge of diverse content and a scoring system that arbitrarily forced a single bipolar masculinity–femininity dimension (e.g., Constantinople, 1973). Addressing these criticisms entailed replacing the single bipolar dimension with two relatively independent unipolar dimensions, one for masculinity and the other for femininity (Bem, 1974; Spence & Helmreich, 1978) and examining these dimensions within various domains of psychological functioning. Therefore, we provide examples of this revised approach and explore their implications for understanding the magnitude of sex/gender differences.

Sex/gender differences in behavioral masculinity and femininity

Our first demonstration pertains to the relations of everyday behaviors to gender/sex, a considerably narrower domain than encompassed by the Terman and Miles (1936) scale. This demonstration relies on Athenstaedt's (2003) development of a multi-item index of behavioral masculinity and femininity.

To identify items that yielded different responses in women and men, Athenstaedt (2003) and her team initially wrote 191 preliminary items describing common observable behaviors from several domains, among them leisure time, occupation, social relationships, and education. Athenstaedt then obtained ratings of these items' typicality for women or men and their desirability for each gender/sex, which enabled item selection for typicality and greater desirability for the more typical gender/sex. These criteria identified two sets of behavioral items, 23 masculine, or male-typical, and 29 feminine, or female-typical.

To assess gender/sex differences in these behaviors, other participants rated how typical each of the behaviors was of themselves (on 7-point scales ranging from *not at all typical* to *very typical*). These participants were Austrians (266 men and 310 women) recruited mainly from a school for vocational training or from sports courses for students and staff members at the University of Graz; additional participants were recruited by snowball sampling conducted by instructors and students of the same university.

On the basis of Athenstaedt's (2003) preliminary analysis of self-rating data, we removed one of these items, "put on makeup," as an outlier because it produced an extreme difference of $d = -2.31$. The remaining unit-weighted items yielded independent masculinity and femininity scales, $r(574) = -.08$. Figure 1 shows the gender/sex difference in the d metric for each item and the two scales. The mean effect size for the individual items was $\bar{d} = |0.67|$. As expected, this effect size was considerably smaller than the effect sizes for the two multi-item scales: Feminine (F) Behavior, $d = -1.82$, and Masculine (M) Behavior, $d = 1.24$. (These effect sizes, as well as all others presented in this article, were not adjusted for potential statistical artifacts.)

To clarify the principles underlying these sizable scale effect sizes, Table 1 presents the properties of three behavioral scales produced from Athenstaedt (2003): F Behavior, M Behavior, and F + M Behavior (composed of the feminine and the reverse-coded masculine items). Following the recommendation of Revelle and Condon (2019), we report three measures of reliability: ω_b , α , and ω_t .³ Although α is best known, it is effectively the average split-half reliability and should

not be interpreted as a measure of internal consistency. The two model-based estimates, ω_b and ω_t , reflect the amount of test variance that is due to one general factor (ω_b) or to all the factors in the test (ω_t). As expected, the reliability values for all scales were high for α and ω_t but lower for ω_b , which dropped to near zero for the scale that contained both the feminine and masculine items.

As also shown in Table 1, the average within-scale item correlations, \bar{r}_i , were low but, as expected, lower for the broader F + M scale that combined the feminine and masculine items. The average item validities, \bar{r}_{iy} (i.e., their prediction of sex/gender), were small and similar across the three scales. The scale validities were moderate for the feminine and the masculine scales and larger for the combined scale, reflecting its broader selection of items. Confirmation of the relevance of both reliability and validity to the scale validities follows from the close match between the empirical scale validities, r_{cy} , and the scale validities modeled by their statistical definition, $r_{cy\ mod}$, which incorporates both validity and reliability (see Appendix).⁴

Expressed in units of Cohen's d , the mean item effect sizes were moderate across the three scales and much smaller than the scale effect sizes. Consistent with the function that relates r to d , the effect sizes expressed in d were each two or more times the value of the corresponding validities expressed in terms of r .⁵

These data demonstrate that, as general trends, women and men differed substantially in their overall tendencies to enact these behaviors, although the differences were much smaller on most of the specific behaviors. Among these behaviors (see Fig. 1), many reflect the conventional domestic division of labor (e.g., shovel snow, do the ironing). Other behaviors pertain to leisure activities (e.g., watch sports on television, go to the ballet), workplace activities (e.g., work overtime, decorate the office with flowers), or social interactions in general (e.g., hold the door open for your partner, listen attentively to others).

The broader meanings that people attach to these behavioral trends stem from the human tendency to spontaneously infer others' psychological attributes as corresponding to their observed behaviors (Uleman et al., 2008). An inspection of Athenstaedt's (2003) behavioral items suggests underlying dispositions—in particular, an orientation of women toward relating to and caring for others and of men toward everyday chivalry and dealing with things. Thus, these data suggest thematic differences in the kinds of behaviors that are more typical of women versus men. However, given that these behaviors are heterogeneous in their psychological content, it is important to determine whether item aggregation also increases the prediction of sex/

Athenstaedt's Behavioral Data Items and Scales

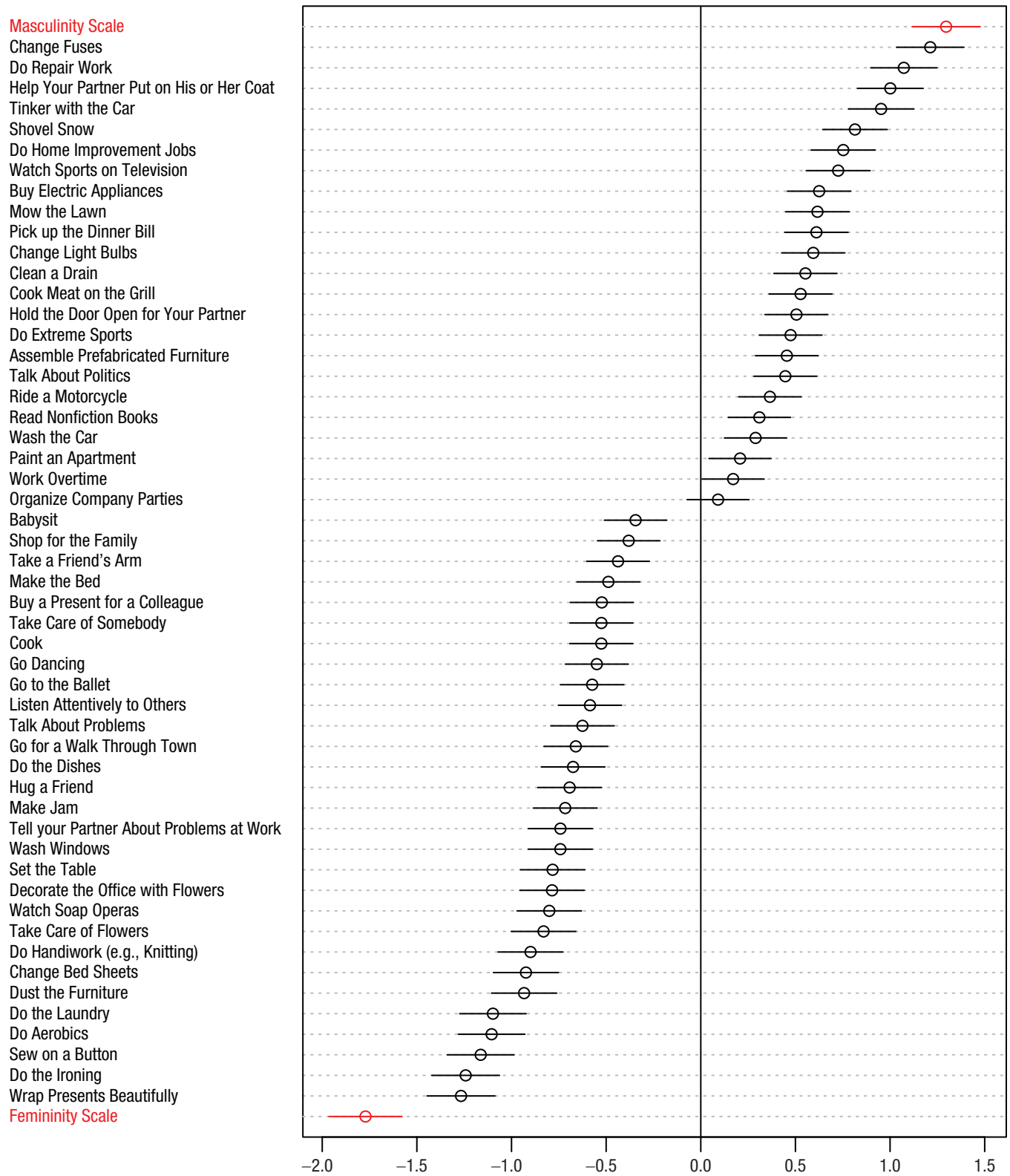


Fig. 1. Effect sizes with 95% confidence intervals for Athenstaedt's (2003) behavioral items and Behavioral Masculinity and Behavioral Femininity Scales.

Table 1. Attributes of Athenstaedt (2003) Scales of Behavioral Femininity and Behavioral Masculinity

Scale	k	ω_b	α	ω_t	\bar{r}_i	\bar{r}_{iy}	r_{cy}	$r_{cy\ mod}$	\bar{d}_i	d
F behavior	29	0.57	0.90	0.91	0.24	-0.34	-0.67	-0.67	-0.75	-1.82
M behavior	23	0.70	0.87	0.89	0.23	0.27	0.53	0.53	0.58	1.24
F + M behavior	52	0.13	0.88	0.90	0.13	-0.31	-0.82	-0.82	-0.67	-2.89

Note: Scales were formed from items loading on femininity (F) and masculinity (M) factors. F + M included all items; the masculinity items were reverse-coded. k = number of items; ω_b = test variance resulting from one general factor; α = coefficient α ; ω_t = test variance resulting from all factors; \bar{r}_i = average within-scale item correlation; \bar{r}_{iy} = average item validity; r_{cy} = observed scale validity; $r_{cy\ mod}$ = modeled scale validity; \bar{d}_i = average-item Cohen's d ; d = scale Cohen's d .

gender when applied to specific domains of psychological functioning.

Sex/gender differences in the femininity and masculinity of personality, cognition, and interests and activities

This second demonstration of aggregation pertains to personality, cognition, and interests and activities. The construction of these scales entailed aggregating sex/gender differences on feminine and masculine items to form two scales in each of these three domains. Sex/gender differences should be larger for each of these scales than for the average of their individual items.

This example relies on an instrument for gender/sex assessment, the Gender-Related Attributes Survey (Gruber et al., 2020). The construction of this self-report instrument followed from the assembly of a preliminary item pool derived from prior research on psychological gender/sex differences as well as a study of self-reported gender identity (Pletzer et al., 2015). Analyses of the factor structure, reliability, and validity of the preliminary items yielded a three-level model. The third level consisted of two general variables, masculinity and femininity, each of which encompassed three domain-specific second-order factors pertaining to personality, cognition, and activities and interests. Although each of these second-order factors encompassed narrower first-order factors, for simplicity and brevity we confine our presentation to the items and scales of the second- and third-order variables.

The analyses we present derive from Gruber et al.'s (2020) validation study, which recruited participants from courses and announcement boards at the University of Salzburg and through bulletin boards in local civic centers. The resulting sample consisted of 471 native German speakers (230 men, $M_{age} = 26.11$, $SD = 8.86$; 241 women, $M_{age} = 25.40$, $SD = 8.98$). Approximately 70% were students, and 30% were from the general population. Using 7-point scales ranging from 1 (*not at all*) to 7 (*very*), these participants rated themselves on the attributes included in these scales in

relation to the general population of men and women within the culture with which they predominantly associated themselves.

Figure 2 shows the sex/gender difference effect sizes for the individual items of the scales as well as for the femininity and masculinity scales within the personality, cognition, and activities and interests factors. The content of each of these six scales is distinctive: (a) for personality, the femininity scale focuses on expressivity and neuroticism, and the masculinity scale focuses on risk-taking, assertiveness, and rationality; (b) for cognition, the femininity scale focuses on verbal skills and memory, and the masculinity scale focuses on spatial and mathematical skills;⁶ and (c) for interests and activities, the femininity scale focuses on female-typical social and sports interests, and the masculinity scale focuses on male-typical social and sports interests.

The mean gender/sex item effect size for the masculinity and femininity scales, respectively, were as follows: for personality, $\bar{d}s = 0.28$ and -0.48 ; for cognition, $\bar{d}s = 0.39$ and -0.24 ; and for interests and activities, $\bar{d}s = 0.48$ and -0.82 . As expected, these item effect sizes were smaller than the corresponding scale effect sizes for masculinity and femininity, respectively: for personality, $ds = 0.57$ and -0.86 ; for cognition, $ds = 0.65$ and -0.40 ; and for interests and activities, $ds = 0.83$ and -1.61 .

To further clarify the principles underlying these aggregated effects, Table 2 presents the properties of three types of scales within the masculine and feminine domains and overall: F (feminine items), M (masculine items), and F + M (feminine and reversed-coded masculine items). The table reports the three measures of reliability: ω_b , α , and ω_t . Once again, the reliability values for all scales were higher for α and ω_t than for ω_b , which, as expected, dropped to near zero for the scales that combined the feminine and masculine items. The average within-scale item correlations, \bar{r}_i , were low but lower for the scales that combined feminine and masculine items and lowest for the M + F All scale, which additionally combined items from all three domains. The average item validities, \bar{r}_{iy} , (i.e., their prediction of sex/gender) tended to be small but were

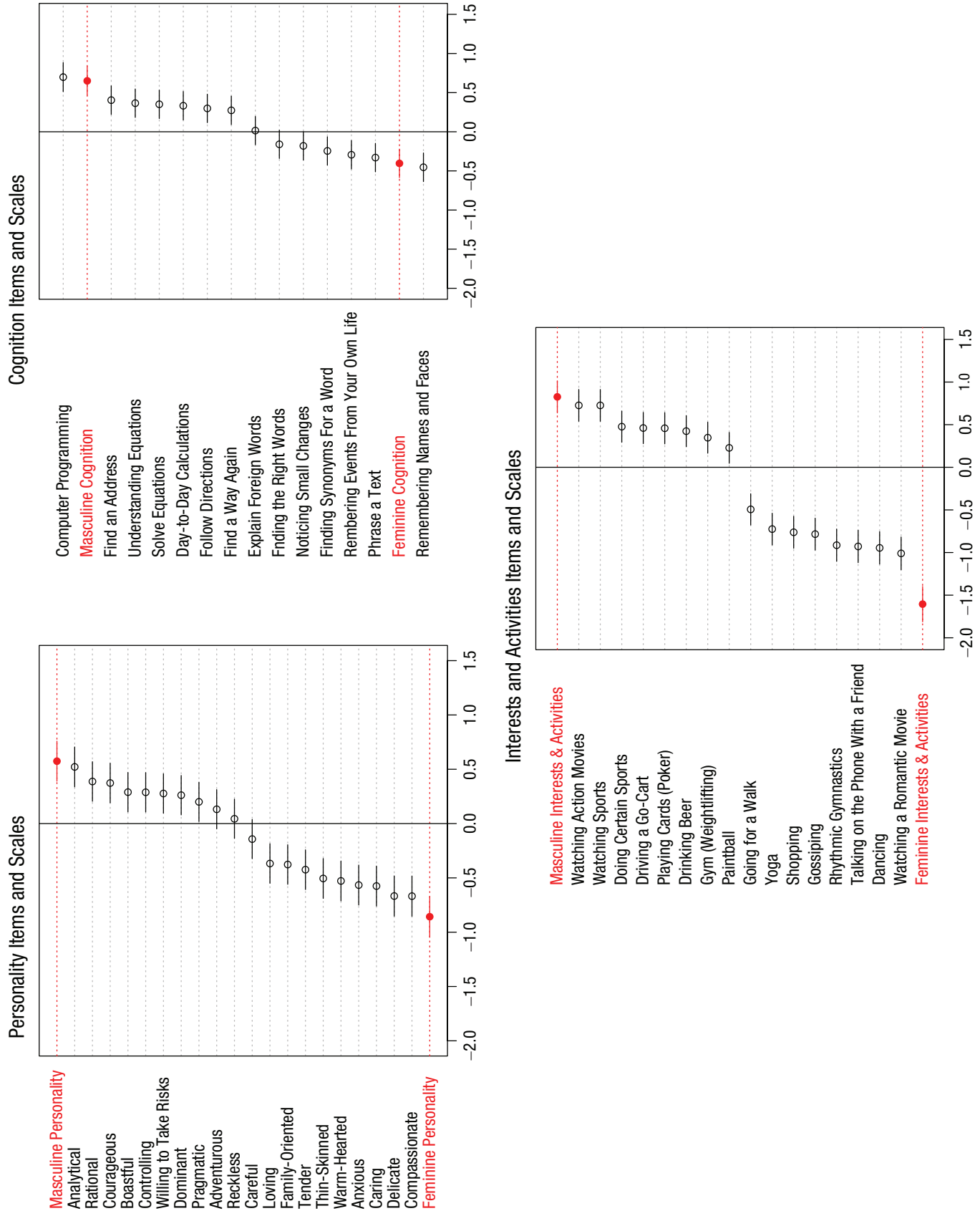


Fig. 2. Effect sizes with 95% confidence intervals for the Gruber et al. (2020) masculinity and femininity of personality, cognition, and interest items and scales.

Table 2. Attributes of Gruber et al. (2020) Femininity and Masculinity Scales of Personality, Cognition, and Interests and Activities

Scale	k	ω_b	α	ω_t	\bar{r}_i	\bar{r}_{iy}	$r_{cy\ mod}$	r_{cy}	\bar{d}_i	d
M personality	10	0.01	0.66	0.65	0.16	0.14	0.27	0.28	0.28	0.57
F personality	10	0.25	0.80	0.84	0.28	-0.23	-0.39	-0.39	-0.48	-0.86
M cognition	7	0.30	0.73	0.84	0.28	0.19	0.31	0.31	0.39	0.65
F cognition	7	0.42	0.70	0.76	0.25	-0.12	-0.19	-0.20	-0.24	-0.40
M interests and activities	8	0.48	0.75	0.78	0.27	0.23	0.38	0.38	0.48	0.83
F interests and activities	8	0.48	0.75	0.79	0.27	-0.38	-0.62	-0.63	-0.82	-1.61
M + F personality	20	0.09	0.77	0.81	0.14	0.18	0.43	0.42	0.38	0.93
M + F cognition	14	0.05	0.67	0.74	0.13	0.15	0.35	0.36	0.31	0.77
M + F interests and activities	16	0.09	0.75	0.79	0.16	0.30	0.66	0.65	0.65	1.73
M all	25	0.25	0.81	0.83	0.15	0.18	0.43	0.43	0.37	0.97
F all	25	0.23	0.83	0.85	0.17	-0.25	-0.55	-0.55	-0.52	-1.41
M + F all	50	0.26	0.85	0.86	0.10	0.21	0.62	0.63	0.45	1.61

Note: Scales were formed from items loading on femininity (F) and masculinity (M) factors for personality, cognition, and interests and activities and for all domains. M + F scales included feminine and masculine items; the masculinity items were reverse-coded. M All, F All, and M + F All included the items from all three domains. k = number of items; ω_b = test variance resulting from one general factor; α = coefficient α ; ω_t = test variance resulting from all factors; \bar{r}_i = average within-scale item correlation; \bar{r}_{iy} = average item validity; $r_{cy\ mod}$ = modeled scale validity; r_{cy} = observed scale validity; \bar{d}_i = average-item Cohen's d ; d = scale Cohen's d .

larger for feminine interests and activities. The scale validities were moderate for the three feminine and three masculine scales but in general larger for the combined M + F scales, reflecting their broader selection of items.

A confirmation of our assumptions about reliability and validity follows from the close match between the empirical scale validities, r_{cy} , and the scale validities modeled by their statistical definition, $r_{cy\ mod}$ (see Appendix). Expressed in units of Cohen's d , the mean item effect sizes were moderate, except for the larger value for the feminine scale of interests and activities, and the scale effect sizes were much larger than the corresponding mean item effect sizes.

In summary, the aggregation of sex/gender-relevant items into broad masculinity and femininity scales produced larger gender/sex differences than the average effect sizes for the individual items. This generalization held whether the underlying items referred to behaviors, as in our first example, or to personality, cognition, or interests and activities, as in our second example. The reason for these gains is that the aggregated scores were a stronger predictor of sex/gender because the number of items in the predictor increased; the gain was enhanced to the extent that the items were not highly intercorrelated, as occurred in the scales that contained both masculine and feminine items.⁷ The meaning of these findings is that men and women differ more strongly in these broadly defined features of personality, cognition, and interests and activities than they do in the narrowly defined tendencies that underlie these broad features.

Other examples of aggregation increasing the magnitude of sex/gender differences

Other examples of the effects of aggregation that are scattered throughout the research literature pertain to psychological variables that show a substantial sex/gender difference but were not designed to do so, as were femininity-masculinity scales. Their component indicators were selected to represent the variable and not for gender/sex typicality. Nevertheless, aggregating the indicators of such a variable can increase the gender/sex difference.

One such example of aggregation occurred in a study of antisocial behavior among children in New Zealand that found boys engaging in more antisocial behavior than girls did. This gender/sex difference was much smaller on the measures of specific antisocial behaviors ($\bar{d} = |0.25|$) than on an index averaged over seven component measures ($d = 0.49$; Moffitt et al., 2001, p. 93). Likewise, a meta-analysis of child temperament found that boys and girls differed mainly in two areas: surgency and effortful control (Else-Quest et al., 2006). Specific indicators of temperament, which typically showed greater surgency in boys and effortful control in girls, yielded mainly small gender/sex differences ($ds = |0.01|$ to $|0.41|$). However, on the aggregated indexes of overall surgency or effortful control reported in some of the studies, differences were larger: $\bar{d} = 0.55$ for boys' greater surgency and $\bar{d} = -1.01$ for girls' greater effortful control.

In general, even for psychological constructs that merely happen to differ by gender/sex but were not designed to do so, the aggregation of their specific indicators increases the magnitude of differences insofar as the underlying indicators also differ consistently between the genders/sexes. Aggregation thus produces a more reliable and valid index of the gender/sex-differentiated attribute that these indicators have in common, which can be a psychological attribute such as effortful control that is greater in one sex/gender.

Another type of aggregation identifies an optimal set of items for predicting respondent sex/gender (Lippa, 2005). This procedure classifies participants by sex/gender on the basis of their responses to items in a particular domain such as occupational preferences or preferred hobbies. Specifically, the method assigns a *gender diagnosticity* score to each participant, defined as the probability that he or she is male (vs. female) as predicted by the weighted average of the items that most successfully classified respondents by gender/sex in a linear discriminant analysis. Gender/sex differences in these gender diagnosticity scores can be very large (e.g., $d = 2.58$ for occupational preferences; Lippa & Connelly, 1990).

What do these varied examples of measure aggregation mean for understanding the psychology of sex and gender? They give evidence of the simultaneous presence of gender/sex similarities and differences in psychological data. Our analyses indicate that women and men (or girls and boys) can differ considerably when their attributes are abstracted to display overall themes, that is, general trends whose components differ across persons. Women and men differ much more in such general female- or male-typical tendencies than in the specific indicators of these tendencies, each of which is influenced by other causes. Similarity and difference are thus compatible and intertwined. These principles apply to aggregation of any appropriate measures, not merely to the self-report measures that predominate in our examples.

As a further reflection on aggregation, consider whether men or women are somewhat more likely to engage in particular behaviors on single occasions. Do such instances actually matter in daily life? The answer is resoundingly “yes” if these behaviors aggregate over occasions and with similar behaviors to produce a notable patterning of behavior. For example, whether a woman or man engages in a single act of dominance such as interrupting someone makes little difference by itself, but consequential sex/gender differences emerge if men or women more often enact the behavior and related dominant behaviors. For example, research on deliberative groups has shown that, especially when men are in the majority and a majority rule for decisions

prevails, they not only engage in hostile (i.e., non-supportive) interruptions of women but also speak more often than women and with a greater feeling of self-efficacy (e.g., Karpowitz & Mendelberg, 2014). This gender/sex inequality in interruptions prevails even in the deliberations of the justices of the U.S. Supreme Court (Patton & Smith, 2017). Given replication of this pattern over time and situations, women and men would reap the consequences that follow from this male-dominant pattern of social interaction (Funder & Ozer, 2019). As this example illustrates, understanding of sex/gender differences and similarities requires taking into account the patterning of behavior as it emerges over time, situations, and variants of the behavior.

Aggregation of Sex/Gender Differences in Multivariate Domains

A different question about psychological sex/gender differences arises in domains such as personality that are composed of distinct variables that usually are components of a single conceptual model. This question is whether women and men differ in general in such a domain. The answer follows from computing an effect size representing the distance between women and men in the multivariate space formed by the component variables.

A first thought might be instead to average the effect sizes of the component variables. To understand why this solution is inadequate, consider the simple example of assessing the distance between two cities, Chicago and Miami, in a two-dimensional space defined by north–south and east–west axes. This distance is not computed as the average of the intercity distances on the two dimensions but by the distance on a straight line connecting the two cities (i.e., the Euclidean distance). Multivariate effect sizes apply this logic to domains composed of two or more variables that are not necessarily independent.

Multivariate effect sizes, like univariate ones, yield a standardized difference between women and men. Specifically, the univariate d represents the difference between the female and male means on a dimension. The multivariate D represents the difference between the female and male centroids in the multivariate space. The centroid, the multivariate analogue of the univariate mean, is the point in multivariate space where the means from the component variables intersect. Computationally, D is the square root of the sum of the squared standardized sex/gender differences on the component variables, which are appropriately weighted so that incorporating new variables increases D only insofar as they contribute unique additional information.⁸ In other words, a new variable cannot enlarge D

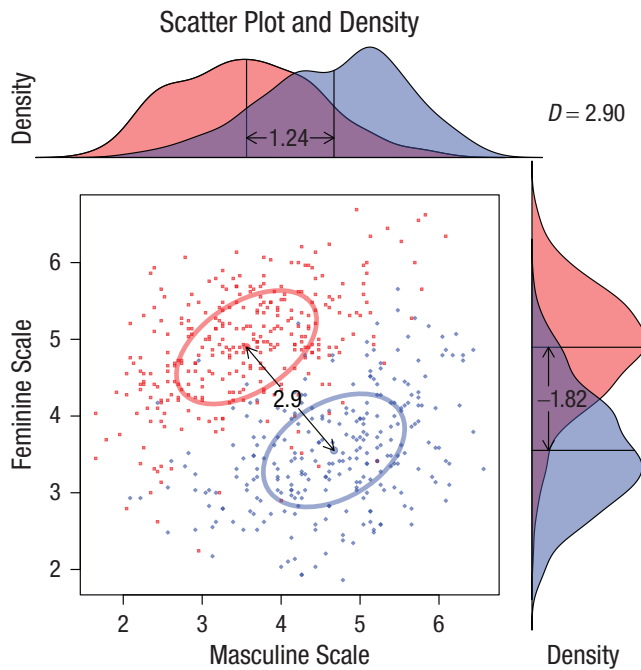


Fig. 3. Scatterplot of the masculine and feminine scales from the Athenstaedt (2003) data set for males (blue) and females (red). Each dot represents an individual participant, the ellipses indicate 1 *SD* from the group centroids, and the arrow between the ellipses represents the Mahalanobis distance (D) between the two centroids. Each plot includes the univariate distributions for the two groups and Cohen's d between the group means. The overall correlation between the scales is $-.08$, and the pooled within-group correlation is 0.45 .

if it is statistically redundant with the variables already included in the sum (see Mahalanobis, 1936).

In summary, the logic of aggregation is different for multivariate and univariate arrays. Univariate aggregation averages responses on multiple indicators (e.g., items) of a single variable to produce a general measure of that variable (e.g., masculine personality). The sex/gender effect size, d , is the standardized difference between the female and male means on the variable. In contrast, the D effect size is also a standardized difference but between the female and male centroids of a multivariate space. This statistic indicates how similar or different women and men are in general on the variables that compose a particular domain. Its interpretation benefits from also considering the univariate d s for the component variables. In other words, D adds information beyond the univariate d s by providing a statistically appropriate summary of them but does not substitute for them.

Multivariate distances in two-dimensional space

Our first examples revisit the Athenstaedt (2003) and the Gruber et al. (2020) data by introducing Mahalanobis D

to represent sex/gender differences in the two-dimensional space defined by the relevant masculinity and femininity scales. For the Athenstaedt data, Figure 3 shows the two-dimensional array defined by scales assessing the masculinity or femininity of behavior. The distributions of the data appear in a scatterplot and density plots for women and men on each of the two scales. The effect sizes for the individual univariate scales appear with the density plots: $d = 1.24$ for M Behavior and $d = -1.82$ for F Behavior. The bivariate scatter plot displays the centroids of the two-dimensional array; Mahalanobis $D = 2.90$ represents the distance between these centroids.

For the Gruber et al. (2020) data, Figure 4 shows the two-dimensional arrays defined by the masculinity and femininity scales for each of Gruber et al.'s three domains—personality, cognition, and interests and activities—as well as for the composite data that combined the three domains. Notable are the univariate and multivariate gender/sex differences for interests and activities, which are larger than those for personality or cognition and even slightly larger than for the combined data. For all four of these displays, the Mahalanobis D is greater than the d effect sizes for the component masculinity or femininity scales.

Now we turn from these bivariate examples to multivariate analyses in two psychological domains for which psychometricians have developed outstanding technologies of assessment: personality and vocational interests.

Multivariate differences in multidimensional space

Personality traits. Personality is an appropriate domain for examining multivariate effect sizes because decades of research have established its multidimensionality. The five-factor, or Big Five, model provides the most popular representation of the dimensions of personality (Digman, 1990; Goldberg, 1992). To varying extents, men and women differ on each of these dimensions. For example, in a large Internet-based survey of 5,417 female and 2,901 male students, Nofle and Shaver (2006, Study 1) found the following gender/sex differences: conscientiousness, $d = -0.28$; agreeableness, $d = -0.22$; neuroticism, $d = -0.49$; openness to experience, $d = 0.08$; and extraversion, $d = -0.16$. These differences produced a mean effect size of $\bar{d} = |0.25|$. A multivariate calculation yielded Mahalanobis $D = 0.84$ (Del Giudice, 2009).

Notably, sex/gender differences tend to be small for the individual Big Five variables, except for neuroticism. These results are not surprising given that the subdimensions, or facets, of Big Five dimensions sometimes show opposite directions for sex/gender differences (Kajonius & Johnson, 2018). For example,

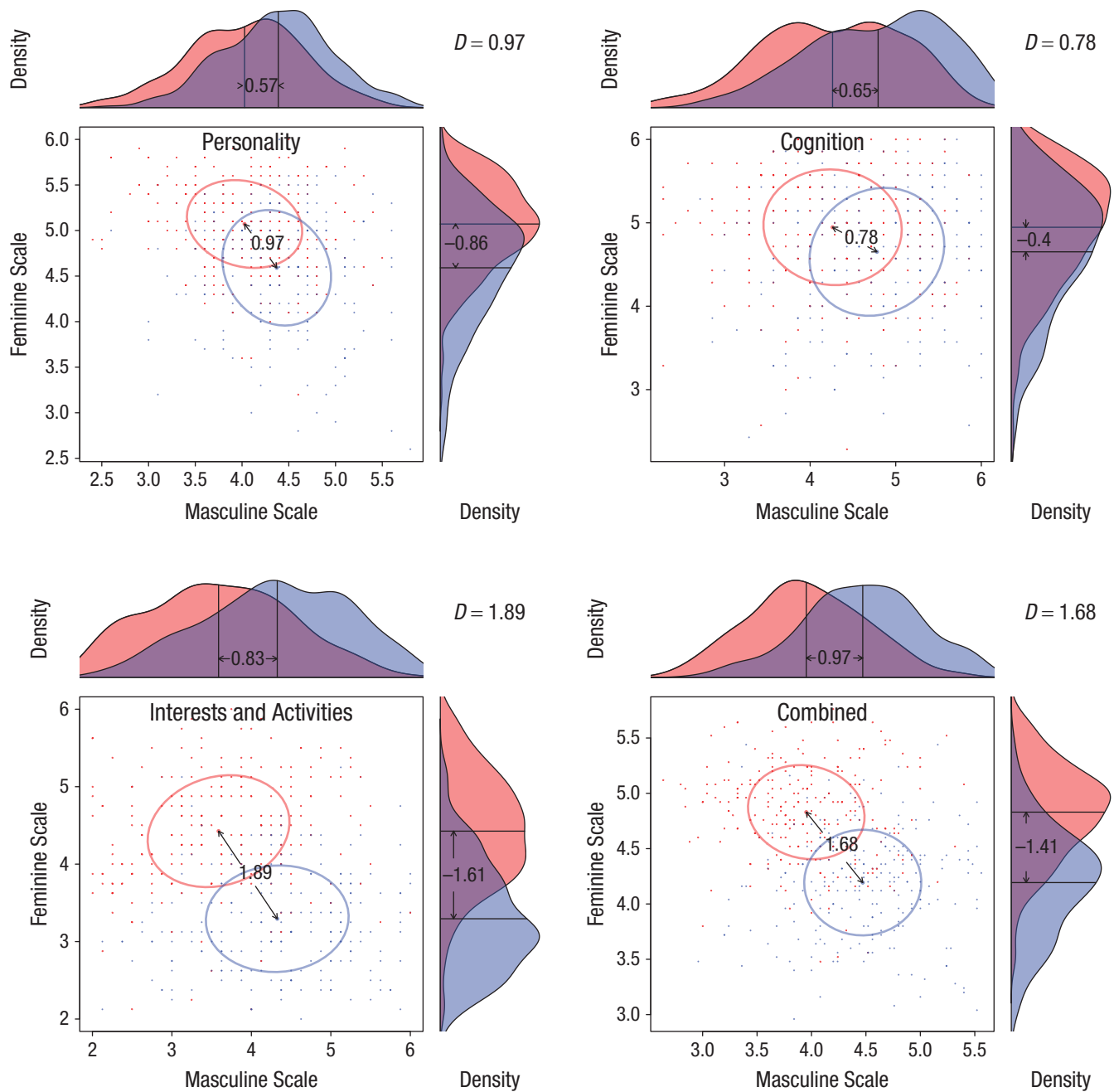


Fig. 4. Scatterplots of the Gruber et al. (2020) masculine and feminine scales of personality, cognition, interests and activities, and all three combined. The three subdomains and overall scores of the Gruber et al. (2020) data set for males (blue) and females (red). Each dot represents an individual participant, the ellipses indicate 1 SD from the group centroids, and the arrow between the ellipses represents the Mahalanobis distance (D) between the two centroids. Each plot includes the univariate distributions for the two groups and Cohen's d between the group means. The pooled within-group correlations are -0.16 (top left), 0.04 (top right), 0.10 (bottom left), and -0.05 (bottom right).

extraversion encompasses the facets of warmth, which is greater in women, and dominance, which is greater in men. Therefore, the overall sex/gender difference in extraversion is small. This mixing of female- and male-typical traits within dimensions also prevails in the Big Two, which forms a stability dimension from

emotional stability (i.e., neuroticism), conscientiousness, and agreeableness and a plasticity dimension from extraversion and openness to experience (Digman, 1997).

Some other systems for representing personality, which less often place male- and female-typical items

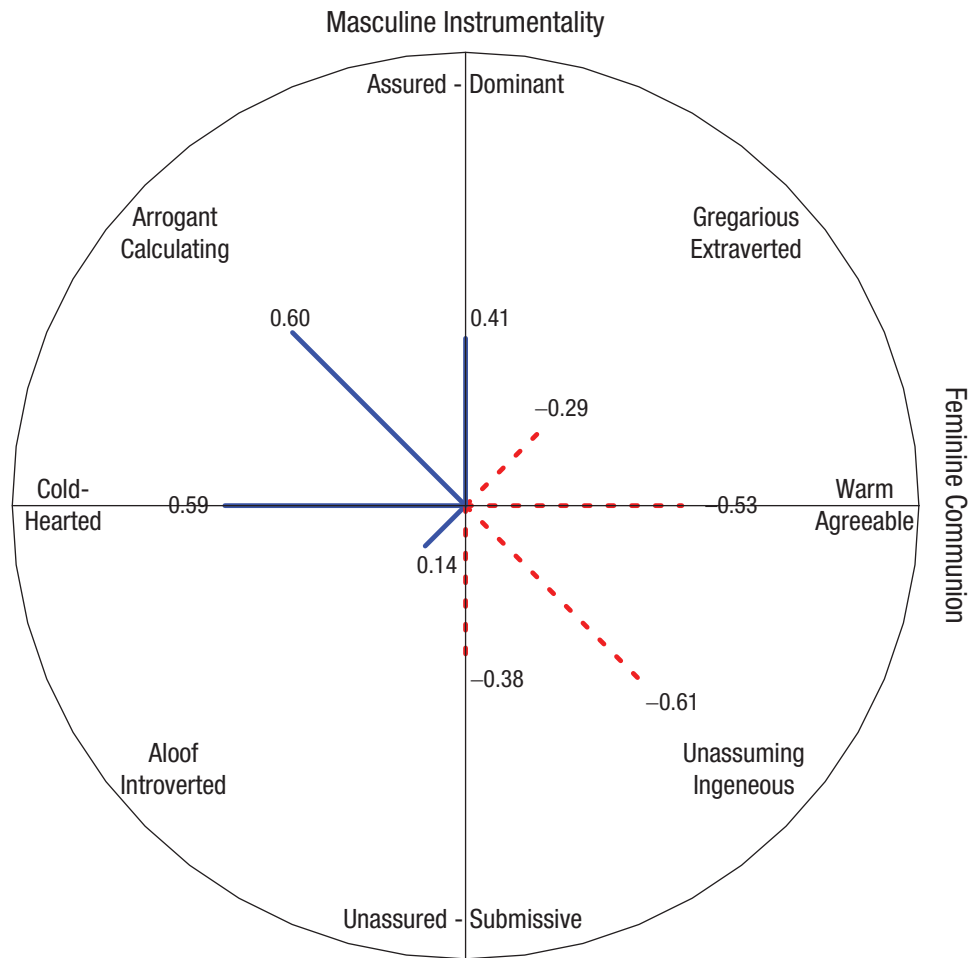


Fig. 5. Sex/gender differences in the interpersonal circumplex. The mean effect sizes (d on the dimensions of the interpersonal circumplex) are from Lippa's (2001) meta-analysis of five studies. The blue lines indicate effects in the male direction, and the red dotted lines indicate effects in the female direction.

on the same scale, predict sex/gender more strongly, as does the Sixteen Personality Factor Questionnaire (16PF; Conn & Rieke, 1994). Del Giudice et al. (2012) thus analyzed data from the 1993 U.S. standardization sample of the fifth edition of the 16PF (5,137 female and 5,124 male respondents). With personality assessed by this measure, the mean sex/gender effect size was $\bar{d} = |0.26|$. The largest univariate effects in the female direction were $d = -1.34$ for sensitivity and $d = -0.59$ for apprehension; in the male direction the largest univariate effects were $d = 0.32$ for emotional stability and $d = 0.27$ for dominance. With the dimensions represented in multivariate space, Del Giudice et al. calculated Mahalanobis distance as $D = 1.49$, thus much larger than the D for the Big Five and larger than any of the univariate effect sizes for the 16PF.

Feminine and masculine themes have also emerged prominently in a representation of personality traits by

a circular structural model known as the *circumplex* (Gurtman, 2009). In this model, personality traits form a circular array defined by two principal, higher-order dimensions, which in Wiggins's (1996) interpersonal circumplex are agency (or dominance/instrumentality) and communion (or warmth/expressiveness). Component variables are located on a circular continuum of similarity and difference so that each variable represents a blend of these two orthogonal dimensions that define the space (see Fig. 5).

Because the interpersonal circumplex has a two-dimensional agency-communion structure by design, it should produce gender/sex differences that correspond to those for the masculinity (or agency) and femininity (or communion) scales in classic personality measures of gender identity (e.g., Bem, 1974). Of interest therefore is Lippa's (2001, p. 293) meta-analysis of five studies that assessed gender/sex differences in the

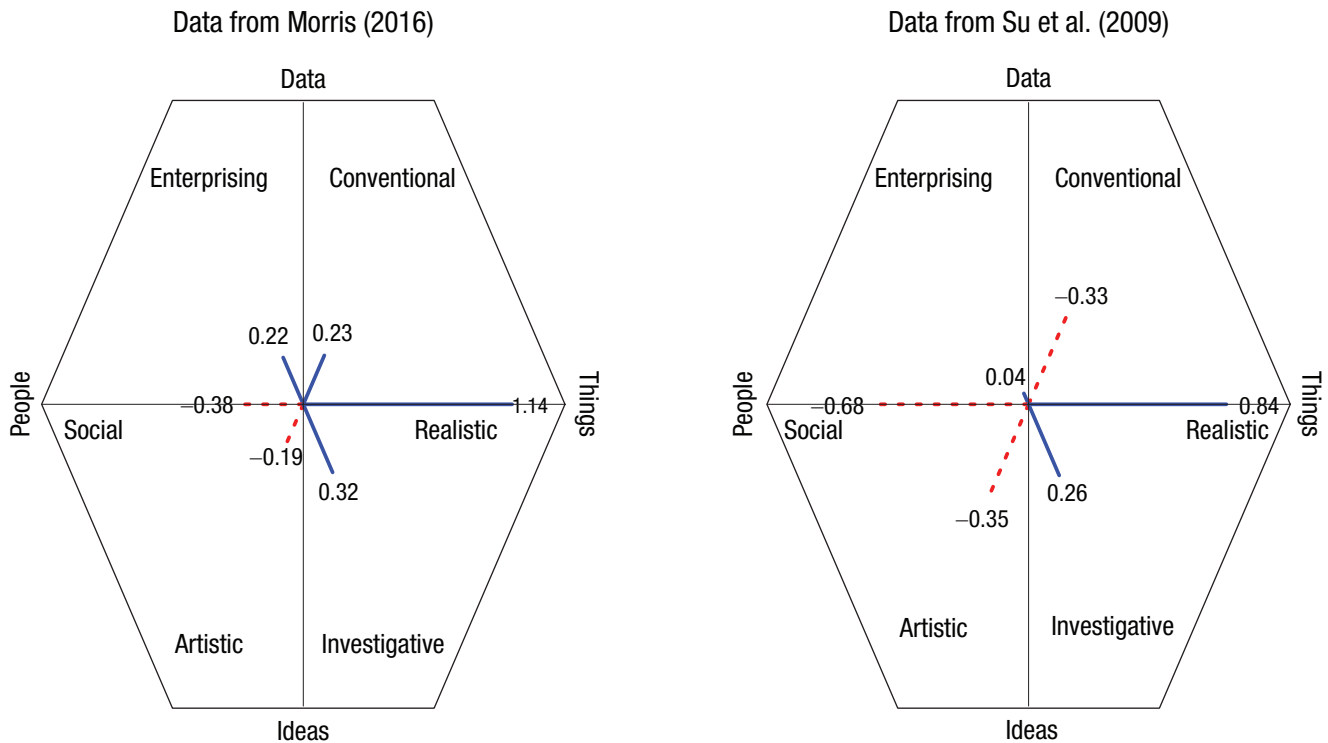


Fig. 6. Six dimensions of vocational interests organized in the Holland (1997) model. The blue lines indicate effects in the male direction, and the red dotted lines indicate effects in the female direction. The labels of the two axes (people vs. things and ideas vs. data) are from Prediger (1982).

circumplex (two for the interpersonal circumplex and three for the closely related circumplex of interpersonal problems). As shown in Figure 5, the meta-analytic effect sizes for the gender/sex differences on the individual circumplex variables ranged from $d = |0.14|$ to $|0.61|$ and averaged to $\bar{d} = |0.44|$. Our computation of Mahalanobis D for one of the data sets from Lippa (2001) yielded $D = 0.67$. The small increment of D beyond the component univariate effect sizes reflects the high correlation between the dimensions with larger effect sizes.

Lippa’s (2001) meta-analysis obtained the expected tendencies of men toward agency and women toward communion (see Fig. 5). However, these trends were somewhat smaller on the masculine instrumentality versus feminine expressiveness axis than on the related but more evaluatively negative axis of arrogant/calculating ($\bar{d} = 0.60$) versus unassuming/ingenuous ($\bar{d} = -0.61$). A similar pattern emerged in Gurtman and Lee’s (2009) study that implemented a different method of assessing gender/sex differences (*octant scores* within the circumplex).

In conclusion, models of personality have produced differing estimates of the overall similarity versus difference of women and men. These discrepancies tend to reflect how well each measurement model isolates

the communal or agentic tendencies that most differentiate female and male personality. In all models, the meaning of the D multivariate distance statistic emerges from considering it in conjunction with its component univariate effect sizes.

Vocational interests. Vocational interests provide another illustration of gender/sex differences in a multivariate domain. To take account of multidimensionality, Holland (1959, 1997) proposed a typology of six variables designed to bridge between personality traits and work environments: realistic (interest in working with things or outdoors), investigative (interest in science, including mathematics and the physical, social, biological, and medical sciences), artistic (interest in creative expression, including writing and the visual and performing arts), social (interest in working with and helping others), enterprising (interest in leadership or persuasive roles directed toward economic objectives), and conventional (interest in working in structured environments, especially in business settings). A hexagon representing these variables places the more highly related ones in closer proximity (see Fig. 6).

To integrate research comparing the vocational interests of men and women, Su et al. (2009) meta-analyzed norming data for U.S. and Canadian samples reported in technical manuals for 47 interest inventories published

between 1964 and 2007 ($N = 503,188$ respondents; sample mean ages between 12.50 and 42.55 years). The results showed that men scored higher than women did on realistic ($\bar{d} = 0.84$) and investigative ($\bar{d} = 0.26$) interests and that women scored higher than men did on artistic ($\bar{d} = -0.35$), social ($\bar{d} = -0.68$), and conventional ($\bar{d} = -0.33$) interests. Enterprising showed little difference ($\bar{d} = 0.04$). The grand mean of these effect sizes was $\bar{d} = |0.45|$.

The most popular measure of vocational interests is the Strong Interest Inventory. Morris (2016) reported sex/gender differences on this measure for a large cross-sectional sample of U.S. residents who completed this test between 2005 and 2014 ($N = 1,283,110$ respondents; ages between 14 and 63 years). These analyses found that men scored higher on realistic ($d = 1.14$), investigative ($d = 0.32$), conventional ($d = 0.23$), and enterprising ($d = 0.22$) interests, and women scored higher on artistic ($d = -0.19$) and social ($d = -0.38$) interests. Most of these effect sizes were thus consistent with the Su et al. (2009) meta-analysis: Men scored higher on realistic and investigative interests, and women scored higher on artistic and social interests. The mean of Morris's effect sizes was $d = |0.41|$.

Comparing women and men in this multivariate space yielded $D = 1.50$ for Morris's (2016) study of the Strong Interest Inventory and $D = 1.40$ for Su et al.'s (2009) meta-analysis (R. Su, personal communication, November 25, 2019). As expected, these multivariate effect sizes were thus much larger than the average of the differences for the six individual scales.

Providing a different type of multivariate summary of interest scores, Prediger (1982) derived two bipolar higher-order dimensions within Holland's hexagon: preferences for working with (a) things versus people and (b) data versus ideas (see Fig. 6).⁹ Assessments of women and men on these metadimensions have consistently produced a substantial difference only on the things-people dimension; men preferred things and women preferred people (see Su & Rounds, 2015). Su et al.'s (2009) meta-analysis thus found a large things-people gender/sex difference ($\bar{d} = 0.93$), as did Morris's (2016) analysis of the Strong Interest Inventory ($d = 1.01$). These values were larger than all but one of the gender/sex differences on the individual dimensions (i.e., realistic in the Morris analysis).

The even larger Mahalanobis D summarized the overall distance between the vocational interests of women and men. To identify which particular interests differed in women versus men, interpreters should refer to the things-people dimension and the six specific dimensions. Once again, the multivariate effect size supplements but does not substitute for univariate (and, if available, bivariate) effect sizes.

Discussion

We invite our readers to embrace the complexity of the psychology of sex and gender by taking into account gender/sex similarities and differences at the differing levels of analysis explored in this article. One lesson is that gender/sex differences become larger by averaging relevant individual indicators that differ by gender/sex to yield measures of broader masculine or feminine psychological tendencies. A second lesson is that differences are larger on assessments of the overall difference between women and men in multidimensional domains such as personality, in which they differ on the component dimensions. Although these insights about magnitude do not speak to the causes of similarity and difference, they clarify the phenomena that require explanation.

The first method by which gender/sex differences increase in magnitude relies on the principle that individual indicators of a variable consist of true score (what the researcher intends to measure) and error score (irrelevant influences; Lord & Novick, 1968). Averaging responses across relevant indicators ordinarily provides a more precise estimate of true-score variance while reducing error-score variance, producing a more reliable aggregated measure of a target psychological variable. In addition, given that the validity of an aggregated criterion shows greater gains to the extent that its components are not highly correlated, sex/gender differences tend to be larger on broader sex/gender-relevant criteria that draw from differing psychological domains.

The second method by which effect sizes estimating gender/sex differences become larger assesses the distance between men and women in psychological domains that comprise two or more variables. This method appropriately combines the differences on the component variables to yield a multivariate difference (Mahalanobis D) between the female and male centroids of the multivariate space. Such analyses answer the question of how different or similar women and men are in general in a psychological domain such as personality.

These insights about effect magnitudes are important, although our presentation has some limitations. One is that, for brevity and simplicity, we have reported effect magnitudes only in the popular metric of standardized average differences. However, readers can easily transform these univariate and multivariate effect sizes into any of several alternative metrics, such as the percentage of overlap of the female and male distributions (see Del Giudice, in press; Revelle, 2021).

Another limitation is that psychological research on sex/gender differences and similarities has almost always assessed a binary comparison; however, a small

percentage of people are biologically intersexed (Sax, 2002) or self-identify as nonbinary or transgendered (e.g., Meerwijk & Sevelius, 2017). Because the proportion of individuals not in the cisgendered binary is increasing, at least in the United States (Jones, 2021), future researchers may routinely take account of these other categories, but this practice is rare in current or earlier presentations. Likewise, the intersectionalities of sex/gender with age, race, sexualities, and other social categories only occasionally appear in psychological studies of sex/gender differences and similarities. Nevertheless, the psychometric principles presented in this article would also illuminate the magnitudes of category differences that result from more complex arrays of gender/sex social categories.

Claims of gender/sex similarity and difference

A return to the opening theme of this article is in order: Scientists disagree about magnitude of sex/gender differences in research findings; some maintain that similarity is the correct overall description of these findings, and others maintain that large differences are common. Our analyses have transcended this debate by explaining how smaller and larger differences can be linked: Small differences on specific variables can function as components of a larger difference on a broader, thematic variable such as the femininity of personality, and small differences on variables within a multivariate psychological domain such as vocational interests can contribute to a larger multivariate difference for the domain as a whole.

Despite the importance of these insights, inconsistencies in aggregation are surely not the only cause of the discrepancies between the existing large-scale quantitative reviews of gender/sex differences, that is, between the generally larger differences in the Archer (2019) review than the Hyde (2005) and the Zell et al. (2015) reviews. The major reason for inconsistency in magnitudes no doubt pertains to what these reviews included given untold degrees of freedom in assembling their databases from hundreds of published meta-analyses. The authors' decision rules about inclusion and exclusion differed, consistent with this freedom. Consider, for example, Hyde's (2005) simple, but ambiguous, selection criterion of "the major meta-analyses that have been conducted on psychological gender differences" (p. 582) and Archer's (2019) exclusion of "studies on attributions or attitudes, except where these relate to core topics, such as sexuality and interests" (p. 1385). Moreover, Archer searched beyond meta-analyses to include information such as crime statistics. For example, Archer included statistics pertaining to

violent crime, homicide, partner homicide, rape, and violent computer game use, all contributing effect sizes greater than 1.00.

The aggregation issues analyzed in this article would have some influence on the effect magnitudes reported in these three general syntheses. Unfortunately, however, meta-analysts have seldom coded the aggregation of indicators underlying the measures reported in primary studies. In addition, the synthesized meta-analyses rarely included both aggregative and component measures, allowing later reviewers to choose which effect sizes to import. In an exceptional example, the Else-Quest et al. (2006) meta-analysis of child temperament did present both types of measures. In her review, Hyde (2014) did not include the effect sizes for the aggregative measures of effortful control ($\bar{d} = -1.01$) and surgency ($\bar{d} = 0.55$) but cited some of the smaller component effect sizes. Archer (2019) included the aggregative effect size for effortful control, as well as several of the smaller component effect sizes. Zell et al. (2015) reported only a grand mean effect size ($\bar{d} = |0.16|$) averaged over the temperament effect sizes in Else-Quest et al.

In another example of contrasting selections from meta-analyses, Hyde (2005) included effect sizes for six of the 12 domains from Table 2 of the Hedges and Nowell (1995) meta-analysis of cognitive abilities, omitting the larger effect sizes (e.g., mechanical reasoning; $ds = 0.83, 0.72$) as well as those for writing ability reported in Table 3 of Hedges and Nowell (1995; $ds = -0.55$ to -0.61). Zell et al. (2015) omitted this meta-analysis entirely, but Archer (2019) incorporated eight of its findings, including the relatively large effect sizes for mechanical reasoning and writing.

As suggested by these examples, reviewers' decisions to include or exclude effect sizes differing in magnitude and sometimes in aggregation can be inconsistent. To help clarify these issues, it would be helpful if meta-analysts coded the degree of aggregation of studies' measures. We also recommend that researchers who synthesize meta-analyses report mean effect sizes for various categories of studies that can exist within individual meta-analyses as well as overall as Hyde (2005, 2014) and Archer (2019) have done.

We declined to correct effect sizes reported in this article for reliability, and thus our data understate the size of the true effects (Booth & Irwing, 2011; Del Giudice et al., 2012; Schmidt & Hunter, 1999). We made this choice for two reasons. The first reason is to express the aggregated effect sizes in the same units (observed rather than latent) as the item effect sizes. The second reason pertains to the standard problem of which adjustment to use (Revelle & Condon, 2019). The tradition of adjusting effects by the square root of the reliability estimated by

the coefficient α necessarily overinflates the effect size because α underestimates reliability. Adjusting by a model-based estimate of reliability, such as ω_i , is perhaps better, but it is still an underestimation. Adjusting for ω_i would increase the effect-size estimates that we have reported by roughly 5% to 7%. Readers should thus realize that the aggregated effect sizes presented in this article would be somewhat larger with the application of such corrections.

We again emphasize that our limited goals precluded providing a general review of psychological gender/sex differences. Instead, our purpose is to demonstrate that understanding the magnitude of gender/sex differences benefits from taking into account both the aggregation of relevant indicators of single variables and the estimation of overall effects in multivariate domains. The principles revealed by this analysis thus inform the science of sex and gender beyond the particular variables used for illustrations.

The psychometric principles invoked in this article would hold for the data of any time or place or culture despite variation in the content of gender/sex differences. As an extreme example, Mead (1935) described a culture in which agency and communion were reversed from Western cultures, yielding agentic women and communal men. Others have challenged Mead's claim (see Shankman, 2009), but if there is or was such a culture, aggregation would still work just fine. The agency-communion content of the personality traits typical of men and women would merely be opposite from that of Western (and many other) cultures.

In summary, our conclusion is that sex/gender comparisons in psychological research produce both large and small differences; the smaller differences are often components of the larger differences. These findings raise questions about the correspondence between people's beliefs about women and men in relation to gender/sex differences at the differing levels of analysis in scientific findings that our analyses have displayed. Social-cognitive research has shown that in general people do aggregate their observations to more abstract beliefs, with varying degrees of accuracy (Kunda & Nisbett, 1986). Manifesting intuitive aggregation, people volunteer primarily personality-type attributes when asked to describe women or men in their own words, although physical characteristics, social roles, and cognitive abilities also emerge (e.g., Broverman et al., 1972; Ghavami & Peplau, 2013). Moreover, suggesting higher-level abstraction, the majority of the personality traits volunteered in these studies cohere thematically into the two families of agency and communion (e.g., Broverman et al., 1972; Williams & Best, 1990), with agentic traits ascribed more to men and communal

more to women (Eagly et al., 2020). Such trends invite closer consideration of the relations between gender stereotypes and scientific findings, an important topic that is beyond the scope of this article.

Another concern is that recognizing the influence of aggregation on effect magnitude might discourage researchers from trusting effect sizes at all. Such a reaction could follow from the overly simple interpretations prevailing in psychology of what magnitudes should be considered small, medium, and large and therefore less or more important. These interpretations ignore how measures are constituted. The solution does not lie in returning to a reliance on statistical significance to evaluate effect magnitude. Rather, progress in understanding effect magnitude requires taking into account the properties of measures along with effect sizes, and we hope that this article will encourage progress in this direction.

Reflections on difference and similarity

We now briefly depart from our focus on the psychometrics of sex/gender comparisons because we suspect that our insights that sex/gender differences can be simultaneously large and small might appear to some readers to threaten gender equality. However, just as we have urged more complex thinking about the magnitude of sex/gender differences, we urge more complex thinking about gender equality.

Consider that acknowledging gender/sex differences can sometime help to redress inequalities. For example, evidence that women's life goals tend to be more communal than those of men has inspired efforts to attract women into STEM by incorporating collaborative, pro-social principles into the practice of science (Diekmann et al., 2017). In addition, recognizing women's greater communion may improve their access to employment in view of U.S. labor-market analyses showing that more jobs increasingly require a higher level of social skills (Deming, 2017). Furthermore, research on the so-called female advantage in leadership has shown that women's tendencies toward more collaborative and participative leadership can confer benefits for groups and organizations in many contexts (Eagly, 2007; Post, 2015). Such considerations display the limitations of assuming that sex/gender difference necessarily disables women or that gender equality requires the psychological similarity of women and men.

Psychological differences between identity groups can be consistent with social equality to the extent that groups and organizations respect and value diversity, not only in demographic characteristics but also in the attitudes, values, personality, preferences, and

competencies that are sometimes correlated with these characteristics. In fact, advocates for diversity have championed the idea that diverse groups are more effective in solving problems and predicting events than are homogeneous groups. Their reasoning follows from the assumption that *cognitive heterogeneity*—differences between identity groups in knowledge, perspectives, preferences, and heuristics—yields more tools and resources for doing the work of groups and organizations (e.g., Page, 2008). Even if such diversity gains can be overstated (Eagly, 2016), they would not follow from gender/sex diversity if women were the psychological clones of men.

Regardless of any advantages or disadvantages that follow from gender/sex differences and similarities, responsible scientists act as honest brokers by producing and communicating valid findings to increase scientific knowledge and contribute to evidence-based policy (Eagly, 2016). To this end, we recommend recognizing the forest and the trees of sex/gender differences and similarities. It is necessary to step away from the individual trees, perhaps to a hilltop, to observe the patterning of trees in a forest. Likewise, the patterning of psychological gender/sex differences can be difficult to discern in narrowly defined attributes but emerges more strongly in general trends. It follows that neither similarity nor difference prevails but instead a more complex intertwining of these two types of findings.

Appendix

Derivation of equation for predicting scale validity from item characteristics

The benefit of aggregating items to form more reliable composites has been known since Spearman (1904) and thoughtfully reviewed by Epstein (1983). Unfortunately, most interpretations of aggregation emphasize the increase in reliability that results from adding more items and assume that validity increases in parallel. This is not the case. Although conventional measures of internal consistency (e.g., α) increase as a function of the number of items in a domain (k) and the average correlation within that domain (\bar{r}_x),¹⁰ the aggregated validity (r_{jx}) increases as a function of the number of items, the average item validity (\bar{r}_y), and the average correlation of the predicting items (\bar{r}_x). Any correlation between two variables is just their covariance divided by the square root of the product of their variances. In the case of predicting a criterion from an aggregation of items, the covariance will be the sum of the individual predictor times the criterion covariance (Σr_{yi}); the variance of the predictor is the sum of the interitem variances and covariances ($\Sigma \Sigma \sigma_{ij}$). In the case of standardized

predictors (i.e., correlations), the covariance with the criterion will be the number of items \times the average item validity ($k\bar{r}_y$); the variance of the predictor is the sum of the number of items (k) and the $k \times (k - 1)$ intercorrelations ($k + k \times (k - 1) \bar{r}_x$). Equation 1 is the result of replacing the sums by the products of averages and the number of items. For a similar derivation, see Equation 4 in Chapter 9 of Gulliksen (1950).

$$r_{jx} = \frac{k\bar{r}_y}{\sigma_x} = \frac{k\bar{r}_y}{\sqrt{k + k \times (k - 1)\bar{r}_x}} \quad (1)$$

If the predicting items all come from the same domain, then r_{jx} increases with the number of items. However, if the aggregation is taken across different domains, then the aggregated prediction can be much larger than that from any single domain because of the decrease in the average correlation of the prediction set. Thus, as k grows, the limit of the validity is

$$r_{jx} = \frac{\bar{r}_y}{\sqrt{\bar{r}_x}}$$

In the case of predicting gender/sex differences, scales coded for masculine or feminine aspects predict gender/sex as the number of items increases. But forming composite M + F scales increases this effect beyond what would be expected by the mere addition of items. Because of the independence of the M and F scales, their composite is a much stronger predictor than either scale by itself. We consider this effect for two different data sets.

Transparency

Action Editor: Laura A. King


Editor: Laura A. King

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iDs

Alice H. Eagly  <https://orcid.org/0000-0003-4938-9101>

William Revelle  <https://orcid.org/0000-0003-4880-9610>

Acknowledgments

We thank Ursula Athenstaedt, Katie Badura, Freya Gruber, Phillip Lemaster, Richard Lippa, Tuulia Ortner, Rong Su, and Ethan Zell for performing analyses or sharing the data presented in this article. We also thank Amanda Diekman, David Funder, Marco Del Giudice, Judith Hall, Richard Lippa, Chris Petsko, Agnieszka Pietraszkiewicz, Sabine Sczesny, and Marcel Zentner for comments on previous versions of this article.

Notes

1. For comparison, consider the benchmarks provided by the following PsycINFO index terms for the same time period—*cognitive behavior therapy*: 17,090 articles, 598 meta-analyses; *leadership*: 23,176 articles, 179 meta-analyses; *cognitive development*: 28,855 articles, 146 meta-analyses; *organizational behavior*: 24,212 articles, 214 meta-analyses; *memory*: 102,779 articles, 670 meta-analyses; *personality*: 120,569 articles, 882 meta-analyses.

2. To indicate the direction of sex/gender comparisons, a positive sign indicates larger male scores, and a negative sign indicates larger female scores. Vertical lines surrounding a numeral indicate an absolute value.

3. The ω function in the psych package (Version 2.1.6; Revelle, 2021) for the R software environment (Version 4.1.0; R Core Team, 2021) provides these estimates.

4. We also calculated an F + M Behavior Short scale from the first halves of the femininity items and reverse-coded masculinity items. Its scale effect size ($d = -1.25$) was larger than those of the separate F Behavior and M Behavior scales (d s = -1.12 and 0.92), which had similar numbers of items. This comparison showed that the gain of prediction for the F + M Behavior scale did not follow only from the greater number of items producing a more reliable scale. Because the items are more highly correlated within than between the M Behavior and F Behavior scales, the broader selection of items in the combined F + M Behavior scale increased the magnitude of the effect size because the validity of the scale also increased.

5. The r to d function is the following: $d = \frac{2r}{\sqrt{1-r^2}}$.

6. Self-report items are unusual assessments of cognition, a domain in which most measures present ability-relevant tasks (see Miller & Halpern, 2014).

7. Another example of the aggregation on single dimensions appears in the Supplemental Materials available online.

8. The Mahalanobis D differs from the Euclidean distance between the centroids by taking into account the correlations between the variables. D combines bivariate distances (i.e., the gender/sex difference on each dimension) in a manner similar to multivariate R combining bivariate correlations. Specifically, the individual distances are weighted by their independent effects by multiplying by the inverse of the pooled correlation matrix of the various predictors. For an exposition on how the Mahalanobis distance compares to a simple Euclidean distance between two centroids, see Del Giudice (2021). Depending on the pattern of correlations, the Mahalanobis distance can be greater or less than the Euclidean distance.

9. The computation of these dimensions maps the six interest types onto the two higher-order dimensions: (a) things-people = $(2.0 \times R) + (1.0 \times I) - (1.0 \times A) - (2.0 \times S) - (1.0 \times E) + (1.0 \times C)$ and (b) data-ideas = $(0.0 \times R) - (1.7 \times I) - (1.7 \times A) + (0.0 \times S) + (1.7 \times E) + (1.7 \times C)$.

10. This is the well-known generalization of the Spearman-Brown formula $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2} = \alpha = \frac{k}{k-1} \frac{\sigma_x^2 - \sum(\sigma_i^2)}{\sigma_x^2}$ for τ equivalent items (Cronbach, 1951; Guttman, 1945).

References

- Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918. <https://doi.org/10.1037/0033-2909.84.5.888>
- Archer, J. (2019). The reality and evolutionary significance of human psychological sex differences. *Biological Reviews*, *94*(4), 1381–1415. <https://doi.org/10.1111/brv.12507>
- Athenstaedt, U. (2003). On the content and structure of the gender role self-concept: Including gender-stereotypical behaviors in addition to traits. *Psychology of Women Quarterly*, *27*, 309–318. <https://doi.org/10.1111/1471-6402.00111>
- Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*, 155–162. <https://doi.org/10.1037/h0036215>
- Booth, T., & Irwing, P. (2011). Sex differences in the 16PF5, test of measurement invariance and mean differences in the US standardisation sample. *Personality and Individual Differences*, *50*(5), 553–558. <https://doi.org/10.1016/j.paid.2010.11.026>
- Bosson, J. K., Vandello, J. A., & Buckner, C. E. (2019). *The psychology of sex and gender*. SAGE.
- Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal. *Journal of Social Issues*, *28*, 59–78. <https://doi.org/10.1111/j.1540-4560.1972.tb00018.x>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Conn, S. R., & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual*. Institute for Personality and Ability Testing.
- Constantinople, A. (1973). Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin*, *80*(5), 389–407. <https://doi.org/10.1177/0959-353505057611>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.
- Del Giudice, M. (2009). On the real magnitude of psychological sex differences. *Evolutionary Psychology*, *7*(2). <https://doi.org/10.1177/147470490900700209>
- Del Giudice, M. (2021). *Individual and group differences in multivariate domains: What happens when the number of traits increases?* PsyArXiv. <https://doi.org/10.31234/osf.io/rgzd2>
- Del Giudice, M. (in press). Measuring sex differences and similarities. In D. P. VanderLaan & W. I. Wong (Eds.), *Gender and sexuality development: Contemporary theory and research*. Springer.
- Del Giudice, M., Booth, T., & Irwing, P. (2012). The distance between Mars and Venus: Measuring global sex differences in personality. *PLOS ONE*, *7*(1), Article e29265. <https://doi.org/10.1371/journal.pone.0029265>
- Deming, D. J. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics*, *132*, 1593–1640. <http://doi.org/10.1093/qje/qjx022>
- Diekmann, A. B., Steinberg, M., Brown, E. R., Belanger, A. L., & Clark, E. K. (2017). A goal congruity model of role

- entry, engagement, and exit: Understanding communal goal processes in STEM gender gaps. *Personality and Social Psychology Review*, 21(2), 142–175. <https://doi.org/10.1177/1088868316642141>
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1), 417–440. <https://doi.org/10.1146/annurev.ps.41.020190.002221>
- Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, 73, 1246–1256. <https://doi.org/10.1006/jesp.2001.1511>
- Eagly, A. H. (2007). Female leadership advantage and disadvantage: Resolving the contradictions. *Psychology of Women Quarterly*, 31(1), 1–12. <https://doi.org/10.1111/j.1471-6402.2007.00326.x>
- Eagly, A. H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance? *Journal of Social Issues*, 72, 199–222. <https://doi.org/10.1111/josi.12163>
- Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist*, 75(3), 301–315. <https://doi.org/10.1037/amp0000494>
- Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin*, 132, 33–72. <https://doi.org/10.1037/0033-2909.132.1.33>
- Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37, 1097–1126. <https://doi.org/10.1037/0022-3514.37.7.1097>
- Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist*, 35, 790–806. <https://doi.org/10.1037/0003-066X.35.9.790790>
- Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, 51(3), 360–392. <https://doi.org/10.1111/j.1467-6494.1983.tb00338.x>
- Fausto-Sterling, A. (2012). *Sex/gender: Biology in a social world*. Routledge. <https://doi.org/10.4324/9780203127971>
- Fishbein, M., & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review*, 81, 59–74. <https://doi.org/10.1037/h0035872>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168. <https://doi.org/10.1177/2515245919847202>
- Ghavami, N., & Peplau, L. A. (2013). An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, 37, 113–127. <http://doi.org/10.1177/0361684312464203>
- Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1), 26–42. <https://doi.org/10.1037/1040-3590.4.1.26>
- Gruber, F. M., Distlberger, E., Scherndl, T., Ortner, T. M., & Pletzer, B. (2020). Psychometric properties of the multifaceted gender-related attributes survey (GERAS). *European Journal of Psychological Assessment*, 36(4), 612–623. <https://doi.org/10.1027/1015-5759/a000528>
- Gullikens, H. (1950). *Theory of mental tests*. John Wiley & Sons.
- Gurtman, M. B. (2009). Exploring personality with the interpersonal circumplex. *Social and Personality Psychology Compass*, 3(4), 601–619. <https://doi.org/10.1111/j.1751-9004.2009.00172.x>
- Gurtman, M. B., & Lee, D. L. (2009). Sex differences in interpersonal problems: A circumplex analysis. *Psychological Assessment*, 21, 515–527. <https://doi.org/10.1037/a0017085>
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282.
- Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269(5220), 41–45. <https://doi.org/10.1126/science.7604277>
- Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology*, 6, 35–45. <https://doi.org/10.1037/h0040767>
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources.
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60, 581–592. <https://doi.org/10.1037/0003-066X.60.6.581>
- Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, 65, 373–398. <https://doi.org/10.1146/annurev-psych-010213-115057>
- Jones, J. M. (2021, February 24). LGBT identification rises to 5.6% in latest U.S. estimate. *Gallup News*. <https://news.gallup.com/poll/329708/lgbt-identification-rises-latest-estimate.aspx>
- Kajonius, P. J., & Johnson, J. (2018). Sex differences in 30 facets of the Five Factor Model of personality in the large public ($N = 320,128$). *Personality and Individual Differences*, 129, 126–130. <https://doi.org/10.1016/j.paid.2018.03.026>
- Karpowitz, C. F., & Mendelberg, T. (2014). *The silent sex: Gender, deliberation, and institutions*. Princeton University Press. <https://doi.org/10.23943/princeton/9780691159751.001.0001>
- Kunda, Z., & Nisbett, R. E. (1986). The psychometrics of everyday life. *Cognitive Psychology*, 18(2), 195–224. [https://doi.org/10.1016/0010-0285\(86\)90012-5](https://doi.org/10.1016/0010-0285(86)90012-5)
- Lippa, R. A. (2001). On deconstructing and reconstructing masculinity–femininity. *Journal of Research in Personality*, 35, 168–207. <https://doi.org/10.1006/jrpe.2000.2307>
- Lippa, R. A. (2005). *Gender, nature, and nurture* (2nd ed.). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9781410612946>
- Lippa, R. A., & Connolly, S. (1990). Gender diagnosticity: A new Bayesian approach to gender-related individual differences. *Journal of Personality and Social Psychology*, 59(5), 1051–1065. <https://doi.org/10.1037/0022-3514.59.5.1051>

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 2(1), 49–55.
- Mead, M. (1935). *Sex and temperament in three primitive societies*. William Morrow and Company.
- Meerwijk, E. L., & Sevelius, J. M. (2017). Transgender population size in the United States: A meta-regression of population-based probability samples. *American Journal of Public Health*, 107(2), e1–e8. <https://doi.org/10.2105/AJPH.2016.303578>
- Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences*, 18(1), 37–45. <https://doi.org/10.1016/j.tics.2013.10.011>
- Mischel, W. (1968). *Personality and assessment*. John Wiley & Sons.
- Moffitt, T. E., Caspi, A., Rutter, M., & Silva, P. A. (2001). *Sex differences in antisocial behavior: Conduct disorder, delinquency, and violence in the Dunedin longitudinal study*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511490057>
- Morris, M. L. (2016). Vocational interests in the United States: Sex, age, ethnicity, and year effects. *Journal of Counseling Psychology*, 63, 604–615. <http://doi.org/10.1037/cou0000164>
- Noftle, E. E., & Shaver, P. R. (2006). Attachment dimensions and the Big Five personality traits: Associations and comparative ability to predict relationship quality. *Journal of Research in Personality*, 40, 179–208. <https://doi.org/10.1016/j.jrp.2004.11.003>
- Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press. <https://doi.org/10.1515/9781400830282>
- Patton, D., & Smith, J. L. (2017). Lawyer, interrupted: Gender bias in oral arguments at the US Supreme Court. *Journal of Law and Courts*, 5(2), 337–361. <https://doi.org/10.1086/692611>
- Pletzer, B., Petasis, O., Ortner, T. M., & Cahill, L. (2015). Interactive effects of culture and sex hormones on the sex role self-concept. *Frontiers in Neuroscience*, 9, Article 240. <https://doi.org/10.3389/fnins.2015.00240>
- Post, C. (2015). When is female leadership an advantage? Coordination requirements, team cohesion, and team interaction norms. *Journal of Organizational Behavior*, 36(8), 1153–1175. <https://doi.org/10.1002/job.2031>
- Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior*, 21, 259–287. [https://doi.org/10.1016/0001-8791\(82\)90036-7](https://doi.org/10.1016/0001-8791(82)90036-7)
- R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1.0) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org>
- Revelle, W. (2021). *Psych: Procedures for personality and psychological research* (Version 2.1.6) [Computer software]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Revelle, W., & Condon, D. M. (2019). Reliability from alpha to omega: A tutorial. *Psychological Assessment*, 31(12), 1395–1411. <https://doi.org/10.1037/pas0000754>
- Sax, L. (2002). How common is intersex? A response to Anne Fausto-Sterling. *Journal of Sex Research*, 39, 174–178. <https://doi.org/10.3389/fnint.2011.00057>
- Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, 27, 183–198. [https://doi.org/10.1016/S0160-2896\(99\)00024-0](https://doi.org/10.1016/S0160-2896(99)00024-0)
- Schudson, Z. C., Beischel, W. J., & van Anders, S. M. (2019). Individual variation in gender/sex category definitions. *Psychology of Sexual Orientation and Gender Diversity*, 6(4), 448–460. <https://doi.org/10.1037/sgd0000346>
- Shankman, P. (2009). *The trashing of Margaret Mead: Anatomy of an anthropological controversy*. University of Wisconsin Press.
- Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15(2), 201–293. <https://doi.org/10.2307/1412107>
- Spence, J. T., & Helmreich, R. L. (1978). *Masculinity & femininity: Their psychological dimensions, correlates, and antecedents*. University of Texas Press.
- Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology*, 6, Article 189. <https://doi.org/10.1037/t02466-000>
- Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, 135, 859–884. <https://doi.org/10.1037/a0017364>
- Terman, L. M., & Miles, C. C. (1936). *Sex and personality: Studies in masculinity and femininity*. McGraw-Hill.
- Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, 59, 329–360. <http://doi.org/10.1146/annurev.psych.59.103006.093707>
- Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, 25, 41–78. <https://doi.org/10.1111/j.1540-4560.1969.tb00619.x>
- Wiggins, J. S. (1996). An informal history of the interpersonal circumplex tradition. *Journal of Personality Assessment*, 66, 217–233. https://doi.org/10.1207/s15327752jpa6602_2
- Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multination study* (Rev. ed.). SAGE.
- Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist*, 70(1), 10–20. <https://doi.org/10.1037/a0038208>