1 **Reporting and Misreporting of Sex Differences in the Biological Sciences**

2 Yesenia Garcia-Sifuentes[1] and Donna L. Maney[1, 2]

3 [1]Graduate Program in Neuroscience and [2]Department of Psychology, Emory University

4

5

6 *Abstract*

7 As part of an initiative to improve rigor and reproducibility in biomedical research, the U. S.

8 National Institutes of Health now requires the consideration of sex as a biological variable in

9 preclinical studies. This new policy has been interpreted by some as a call to compare males

10 and females with each other. Researchers testing for sex differences may not be trained to do

11 so, however, increasing risk for misinterpretation of results. Using a list of recently published

12 articles curated by Woitowich et al. (eLife, 2020; 9:e56344), we examined reports of sex

13 differences and non-differences across nine biological disciplines. Sex differences were claimed

14 in the majority of the 147 articles we analyzed; however, statistical evidence supporting those

15 differences was often missing. For example, when a sex-specific effect of a manipulation was

16 claimed, authors usually had not tested statistically whether females and males responded

17 differently. Thus, sex-specific effects may be over-reported. In contrast, we also encountered

18 practices that could mask sex differences, such as pooling the sexes without first testing for a

19 difference. Our findings support the need for continuing efforts to train researchers how to test

20 for and report sex differences in order to promote rigor and reproducibility in biomedical

21 research.

Introduction

Historically, biomedical research has not considered sex as a biological variable (SABV). Including only one sex in preclinical studies—or not reporting sex at all—is a widespread issue (Sugimoto et al., 2019). In a cross-disciplinary, quantitative assessment of the 2009 biomedical literature, Beery and Zucker (2011) found a concerning bias toward the use of males only. As awareness of this issue increased, in 2016 the National Institutes of Health (NIH) implemented a policy requiring consideration of SABV in the design, analysis, and reporting of all NIH-funded preclinical research (NIH, 2015; Clayton, 2018). By addressing the long-standing over-representation of male non-human animals and cells, the policy was intended not only to ameliorate health inequities but to improve rigor and reproducibility in biomedical research (Clayton & Collins, 2014).

Although the NIH policy does not explicitly require that males and females be compared directly with each other, the fact that more NIH-funded researchers must now study both sexes should lead to an increase in the frequency of such comparisons (Maney, 2016). For example, there should be more testing for sex-specific responses to experimental treatments. However, in a follow-up to Beery and Zucker's 2011 study, Woitowich et al. (2020) showed evidence to the contrary. Their analysis revealed that between 2011 and 2019, although the proportion of articles that included both sexes significantly increased (see also Will et al., 2017), the number that analyzed the data by sex did not. This finding contrasts sharply with expectations, given not only the NIH mandate but also numerous calls over the past decade to disaggregate all preclinical data by sex and to test for sex differences (e.g., Becker et al., 2016; Potluri et al., 2017; Shansky & Murphy, 2021; Tannenbaum, 2019; Woitowich & Woodruff, 2019).

One potential barrier to SABV implementation is a lack of relevant resources; for example, not all researchers have received training in experimental design and data analysis that would allow them to test for sex differences using appropriate statistical approaches. This barrier is quite important not only because it prevents rigorous consideration of sex in the first

2

48   place, but also because any less-than-rigorous test for sex differences creates risk for

49   misinterpretation of results and dissemination of misinformation to other scientists and to the

50   public (Maney, 2016). In other words, simply calling for the sexes to be compared is not enough

51   if researchers are not trained to do so; if SABV is implemented haphazardly, it has the potential

52   to decrease, rather than increase, rigor and reproducibility.

53   In this study, our goal was to analyze recently published articles to determine how often

54   sex differences are being reported and what statistical evidence is most often used to support

55   findings of difference. To conduct this assessment, we leveraged the collection of articles

56   originally curated by Woitowich et al. (2020) for their analysis of the extent to which SABV is

57   being implemented. Their original list, which was itself generated using criteria developed by

58   Beery & Zucker (2009), included 720 articles published in 2019 across nine biological

59   disciplines and 34 scholarly journals. Of those, Woitowich et al. identified 151 articles that

60   included females and males and that analyzed data disaggregated by sex or with sex as fixed

61   factor or covariate. Working with that list of 151 articles, we asked the following questions for

62   each: First, was a sex difference reported? If so, what statistical approaches were used to

63   support the claim? We focused in particular on studies with factorial designs in which the

64   authors reported that the effect of one factor, for example treatment, depended on sex. Next, we

65   asked whether data from males and females were kept separate throughout the article, and if

66   they were pooled, whether the authors tested for a sex difference before pooling. Finally, we

67   noted whether the authors used the term "sex" or "gender", particularly in the context of

68   preclinical (non-human animal) studies.

69

70   ***Results***

71   We began with 151 articles, published in 2019, that were determined by Woitowich et al.

72   to have (1) included both males and females and (2) reported data by sex (disaggregated or

73   with sex included in the statistical model). Of those, we identified four that contained data from

3

74   only one sex (e.g., animals of the other sex had been used only as stimulus animals or to

75   calculate sex ratios). After excluding those articles, our final sample size was 147. See Table 1

76   for the sample sizes of articles from each discipline. More than one-third of the studies were on

77   humans (35%) and a similarly large proportion on rats or mice (31%). The remainder

78   encompassed a wide variety of species including non-human primates, dogs, cats, pigs, sheep,

79   deer, squirrels, racoons, Tasmanian devils, lemur, lions, meerkats, and mongoose. All codes

80   and results of coding are shown in Tables S1-S3.

81

82   **Table 1.** Journals surveyed by discipline.

| DISCIPLINE	JOURNAL 1	JOURNAL 2	JOURNAL 3	JOURNAL 4	NO. ARTICLES
BEHAVIOR	Behavioral Ecology & Sociobiology	Animal Behavior	Animal Cognition	Behavioral Ecology	40
BEHAVIORAL PHYSIOLOGY	Journal of Comparative Psychology	Behavioral Neuroscience	Physiology and Behavior	Hormones and Behavior	20
ENDOCRINOLOGY	European Journal of Endocrinology	Journal of Neuroendocrinology	Endocrinology	American Journal of Physiology – Endocrinology & Metabolism	27
GENERAL BIOLOGY	PLoS Biology	Proceedings of the Royal Society B: Biological Sciences	Nature	Science	9
IMMUNOLOGY	Journal of immunology	Infection and Immunity	Immunity	Vaccine	10
NEUROSCIENCE	Journal of Neuroscience	Neuroscience	Journal of Comparative Neurology	Nature Neuroscience	9
PHARMACOLOGY	Neuropsycho-pharmacology	Journal of Psychopharmacology	Journal of Pharmacology and Experimental Therapeutics	British Journal of Pharmacology	11
PHYSIOLOGY	Journal of Physiology (London)	American Journal of Physiology – Renal Physiology	American Journal of Physiology – Gastrointestinal and Liver Physiology	American Journal of Physiology – Heart and Circulatory Physiology	12
REPRODUCTION	Biology of Reproduction	Reproduction			9

83

84

85 *Question 1: Was a sex difference reported?*

86   Results pertaining to Question 1 are shown in Fig. 1A. Comparing the sexes, either

87 statistically or by assertion, was common, occurring in 80% of the articles. A positive finding of a

88 sex difference was reported in 83 articles, or 56%. Of the articles reporting a sex difference, 41

89 (49%) mentioned that result in the title or the abstract. Thus, in our sample of articles in which

90 data were reported by sex, a sex difference was reported in more than half of the articles and in

91 half of those, the difference was treated as a major finding. In 44% of articles, a sex difference

92 was neither stated nor implied.

93   These results are broken down by discipline in Fig. 1B. The sexes were most commonly

94 compared in the field of Endocrinology (93%) and least often in the field of Neuroscience (33%).

95 When sex differences were found in the field of Endocrinology, however, they were reported in

96 the title or abstract only 32% of the time. In the field of Reproduction, the sexes were compared

97 89% of the time and in 100% of those cases, a sex difference was mentioned in the title or

98 abstract. Sex differences were least likely to be emphasized in the title or abstract in the fields of

99 General Biology and Neuroscience (11% each).

100   Although a sex difference was claimed in a majority of articles (57%), not all of these

101 differences were supported with statistical evidence. In nearly a third of the articles reporting a

102 sex difference, or 24/83 articles, the sexes were never actually compared statistically. In these

103 cases, the authors claimed that the sexes responded differentially to a treatment when the effect

104 of treatment was not statistically compared across sex. This issue is explored in more detail

105 under *Question 2*, below. Finally, we noted at least five articles in which the authors claimed that

106 there was no sex difference, but did not appear to have tested statistically for one.

5
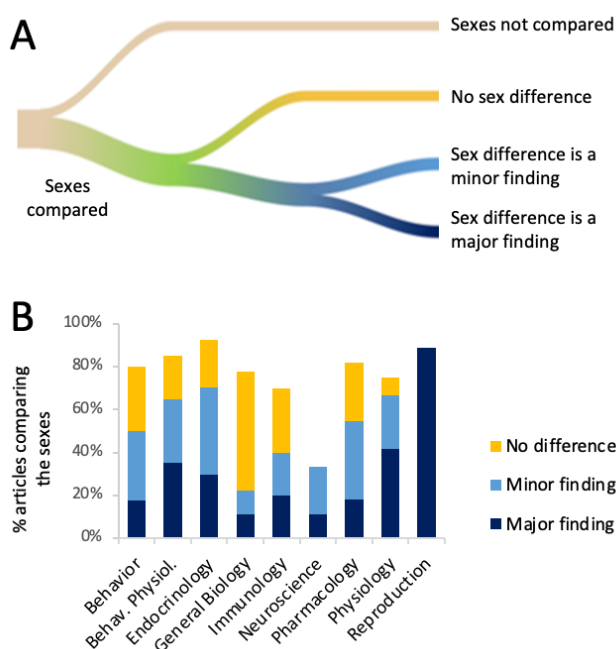
Fig. 1. The sexes were compared in the majority of the articles analyzed. (A) The river plot shows the proportions of articles comparing the sexes, either statistically or qualitatively, and the outcomes of those comparisons. The width of each stream is proportional to the number of articles represented in that stream. If a sex difference was mentioned in the title or abstract, the article was coded as "major finding." For a larger river plot showing how (A) fits into the larger context of the study by Woitowich et al. (2020), please see Fig. S1. (B) The percentage of articles in which sexes were compared is plotted for each discipline. All data are shown in Table S3.

107

108   *Question 2: Did the study have a factorial design with sex as a factor, and if so, did the authors*

109   *test statistically whether the effect of other factors depended on sex?*

110   For each article, we asked whether it contained a study with a factorial design in which

111   sex was one of the factors. This design is common when researchers are interested in testing

112   whether the sexes respond differently to a manipulation such as a drug treatment (Fig 2A).

113   Below, we use the term "treatment" to refer to any non-sex factor in a factorial design. Such

114   factors were not limited to treatment, however; they also included variables such as genotype,

115   season, age, exposure to stimuli, etc. Hypothetical results of a study with such a design are

116   shown in Fig. 2B. In order to draw a conclusion about whether responses to treatment differed

117   between females and males, the effect of the treatment must be compared across sex. Although

118   there are several ways of making such a comparison (see Cumming, 2012; Gelman & Stern,

119   2006), it is typically done by testing for an interaction between sex and treatment in a two-way

120   analysis of variance (ANOVA). If the interaction is significant, then a claim can be made that the

121   sexes responded differently to the treatment. Comparing the treated and control groups within

122   each sex, in other words disaggregating the data by sex and testing for effects of treatment

6

123   separately in females and males, does not test whether the sexes responded differently; that is,

124   it does not test whether the magnitude of the response differs between females and males

125   (Gelman & Stern, 2006; Makin & de Xivry, 2019; Maney, 2016; Nieuwenhuis et al., 2011; Radke

126   et al., 2021).

127

128
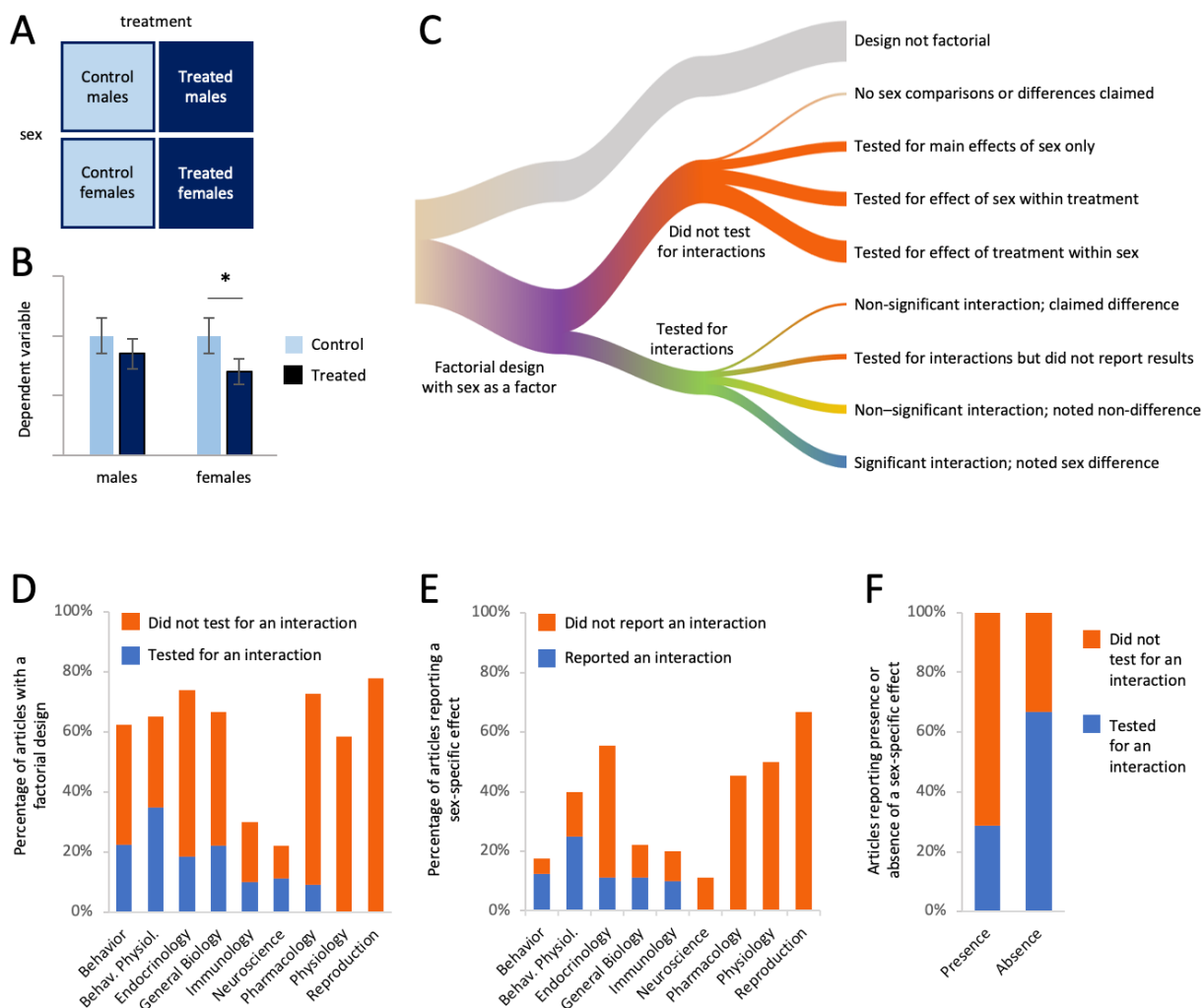
Fig. 2. Factorial designs and sex-specific effects. For each article, we noted whether it contained a study with a factorial design with sex as a factor (A), for example males and females nested inside treated and control groups. (B) In this hypothetical dataset, there was a significant effect of treatment only in females. Some authors would claim that the treatment had a "sex-specific" effect without testing statistically whether the response to treatment depended on sex. In this example, it does not (see Maney, 2016; Nieuwenhuis et al., 2011). (C) The river plot shows the proportion of articles with a factorial design and the analysis strategy for those. The width of each stream is proportional to the number of articles represented in that stream. (D) The percentage of articles with a factorial design is plotted for each discipline. Only a minority tested for an interaction. (E) The percentage of articles reporting a sex-specific effect is plotted for each discipline. Only a minority reported a significant interaction. (F) Testing for an interaction was less common in articles claiming the presence of a sex-specific effect than in articles claiming the absence of such an effect.

7

129   The results pertaining to Question 2 are shown in Fig 2C-2F. Out of the 147 articles we

130   analyzed, 91 (62%) contained at least one study with a factorial design in which sex was a

131   factor (Fig. 2C). Regardless of whether a sex difference was claimed, we found that the authors

132   explicitly tested for interactions between sex and other factors in only 26 of the 91 articles

133   (29%). Testing for interactions varied by discipline (Fig. 2D). Authors were most likely to test for

134   and report the results of interactions in the field of Behavioral Physiology (54% of relevant

135   articles) and least likely in the fields of Physiology (0%) and Reproduction (0%).

136   Of the studies with a factorial design, 58% reported that the sexes responded differently

137   to one or more other factors. The language used to state these conclusions often included the

138   phrase "sex difference" but could also include "sex-specific effect" or that a treatment had an

139   effect "in males but not females" or vice-versa. Of the 52 articles containing such conclusions,

140   the authors presented statistics showing a significant interaction, in other words appropriate

141   evidence that females and males responded differently, in only 15 (29%). In one of those

142   articles, the authors presented statistical evidence that the interaction was non-significant, yet

143   claimed a sex-specific effect nonetheless. In an additional five articles, the authors mentioned

144   testing for interactions but presented no results or statistics (e.g., p values) for those

145   interactions. In the remainder of articles containing claims of sex-specific effects, the authors

146   took one of two approaches; neither approach included a two-way ANOVA. Instead, authors

147   proceeded to what would normally be the post-hoc tests conducted after finding a significant

148   interaction in the ANOVA. In 24 articles (46% of articles with claims of sex-specific effects)

149   authors reported the effect of treatment within each sex and, reaching different conclusions for

150   each sex (e.g., finding a p value below 0.05 in one sex but not the other), inappropriately argued

151   that the response to treatment differed between females and males (see Fig. 2B). In seven

152   other articles claiming a sex-specific effect (14%), the sexes were compared within treatment;

153   for example authors compared the treated males with the treated females, not considering the

154   control animals. Neither approach tests whether the treatment had different effects in females

8

155   and males. Thus, a substantial majority of articles containing claims of sex-specific effects

156   (71%) did not present statistical evidence to support those claims; in the majority of those

157   articles (24/37), the sexes were never compared statistically at all.

158   The prevalence of reporting sex-specific effects varied by discipline (Fig. 2E). Articles in

159   the field of Reproduction were most likely to contain claims of sex-specific effects (67%) and in

160   Neuroscience least likely (10%). Such claims were most likely to be backed up with statistical

161   evidence in the field of Behavioral Physiology (63%). None of the articles in the fields of

162   Reproduction, Physiology, Pharmacology, or Neuroscience contained statistical evidence for the

163   sex-specific effects that were claimed.

164   The omission of tests for interactions was related to whether researchers were claiming

165   sex differences or not. Among the articles that were missing tests for interactions and yet

166   contained conclusions about the presence or absence of sex-specific effects (40 articles), those

167   claims were in favor of sex differences 88% of the time, compared with only 12% claiming that

168   the responses in females and males were similar. Of all of the articles claiming similar

169   responses to treatment, authors tested for interactions in the majority of cases (68%; Fig. 2F).

170

171   *Question 3: Were the data from males and females pooled for any of the analyses?*

172   In this study we included only articles in which data were reported by sex as previously

173   determined by Woitowich et al. (2020). Thus, any articles in which the sexes were pooled for all

174   analyses were not included here. We assigned each of the 147 articles to one of three

175   categories, as follows (Fig. 3A). In 34 (23%) of the articles, data from males and females were

176   analyzed separately throughout. In 60 (41%) of the articles, males and females were analyzed

177   in the same statistical models, but in those cases sex was included as a fixed factor or a

178   covariate. In most cases when sex was a covariate, authors reported the results of the effect of

179   sex rather than simply controlling for sex. In the remaining 53 (36%) articles, the sexes were

180   pooled for at least some of the analyses.

9

181

182
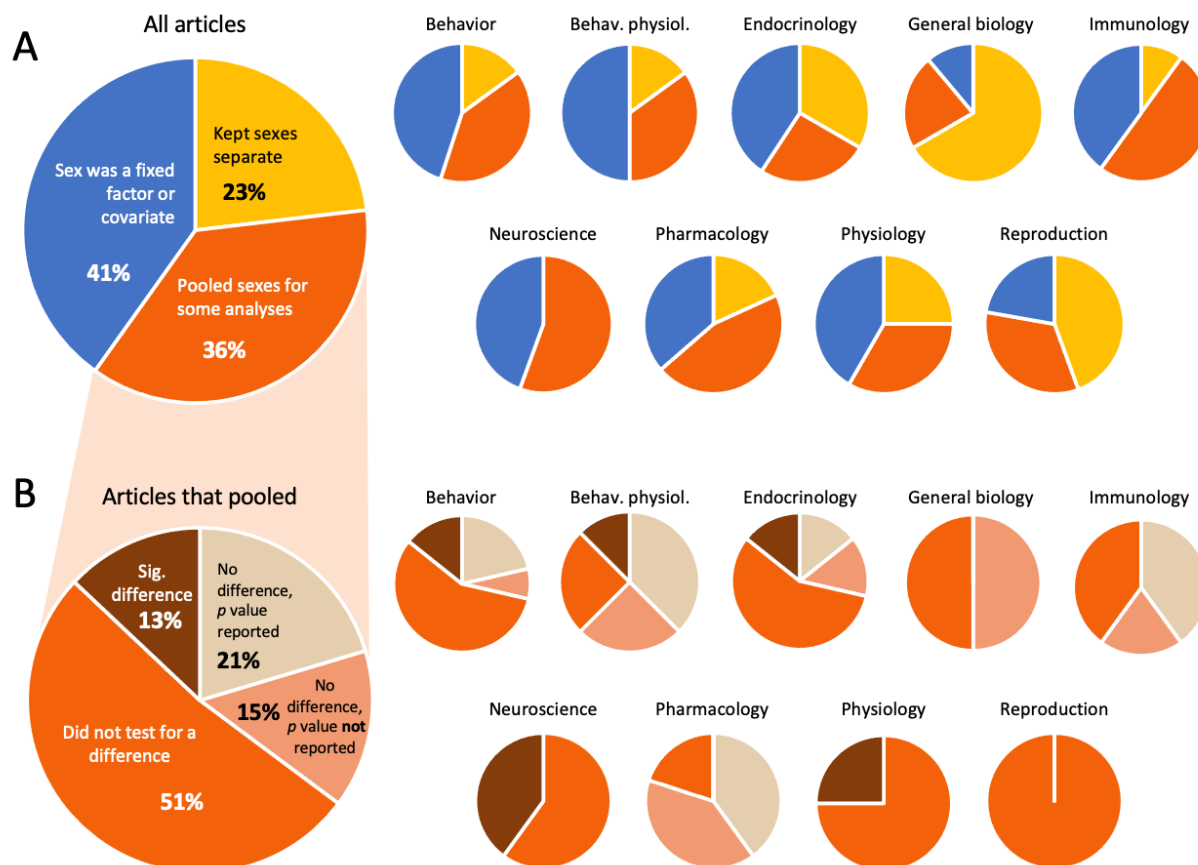
Fig. 3. Proportion of articles in which the sexes were pooled. (A) In our sample, roughly one-third of the articles pooled the sexes for at least some analyses. (B) Among the articles that pooled, more than half did not test for a sex difference before pooling. In both (A) and (B), the smaller pie charts to show the proportions within discipline.

Among the articles in which the sexes were pooled, the authors did so without testing for

a sex difference more than half of the time (51%; Fig. 3B). When authors did test for a sex

difference before pooling, they sometimes found a significant difference yet pooled the sexes

anyway; this occurred in 13% of the articles that pooled. When the sexes were pooled after

finding no significant difference (36% of the articles that pooled), authors presented *p* values for

the sex difference about two-thirds of the time (11 out of 18 articles). Those *p* values ranged

from 0.15 to >0.999. Effect sizes were never reported before pooling.

10

195   Across disciplines, pooling was most prevalent in Neuroscience (56%) and least

196   prevalent in General Biology (22%). Males and females were most likely to be kept separate in

197   General Biology (67%) and most likely to be included in statistical models in the field of

198   Behavior (50%). When females and males were pooled, authors in the field of Reproduction

199   were least likely to have tested for a sex difference before pooling (0%) and most likely to do so

200   in Pharmacology (80%). Pooling after finding a significant difference was most common in the

201   field of Neuroscience (40% of articles that pooled).

202
203   *Question 4: Was the term "gender" used for non-human animals?*

204   To refer to the categorical variable comprising male/female or man/woman (all were

205   binary), the term "sex" was used exclusively in 69% of the articles (Fig. 4). "Gender" was used

206   exclusively in 9%, and both "sex" and "gender" were used in 19%. When both terms were used,

207   they usually seemed to be used interchangeably. In 4% of the articles, neither term was used.

208   Of the articles in which the term "gender" was used, 20% of the time it referred to non-

209   human animals, such as mice, rats, and pigs. In one case, both "sex" and "gender" were used to

210   refer to non-human animals in the title. In another case, "gender" was used to refer to human

211   cells.
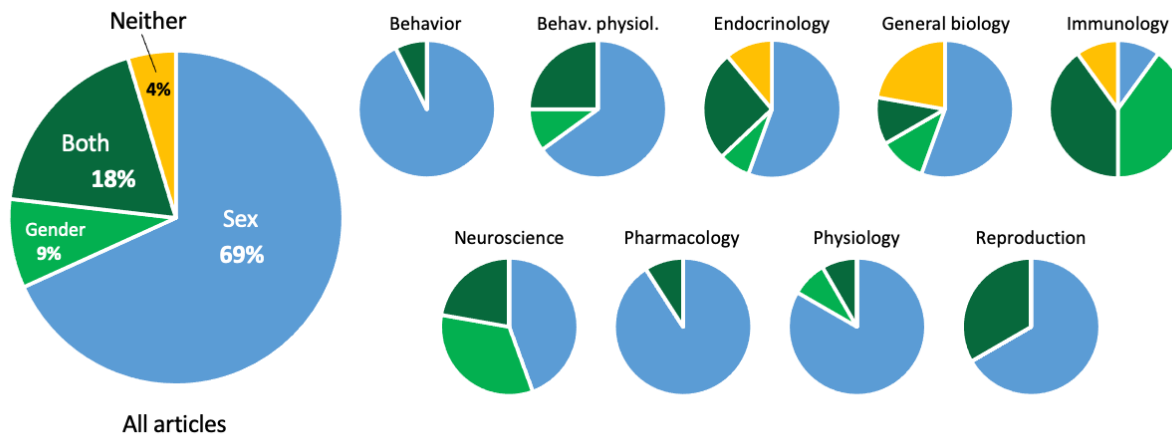
212

213   **Fig. 4. Proportions of articles using the terms "sex" and "gender".** The smaller pie charts show the
214   proportions within discipline.
215

11

216 ***Discussion***

217 *Finding sex differences*

218   Woitowich et al. (2020) found that over the past decade, the proportion of biological

219 studies that included both females and males has increased, but the proportion reporting data

220 by sex has not. Here, we have taken a closer look at the studies determined by those authors to

221 have reported data by sex, that is, to have conformed to NIH guidelines on SABV. We found

222 that in this subset of studies, authors typically also compared the sexes either statistically or

223 qualitatively (80% of cases). Thus, the authors that complied with NIH guidelines to

224 disaggregate data usually went *beyond* NIH guidelines to explicitly compare the sexes with each

225 other. This finding is consistent with a larger analysis of articles in the field of Neuroscience from

226 2010 to 2014; when authors disaggregated data by sex, they usually proceeded to compare the

227 sexes as well (Will et al., 2017). It is important to note, however, that both Will et al. (2017) and

228 Woitowich et al. (2020) found that data were not analyzed by sex in the majority of articles that

229 included both sexes (see Fig. S1). Thus, our current finding that the sexes were usually

230 compared should be interpreted in the context of the subset of articles following NIH guidelines.

231 In the set of articles analyzed here, sex differences were claimed in a majority and were often

232 highlighted in the title or abstract. We therefore found little evidence that researchers—at least

233 those who comply with NIH guidelines—are uninterested in sex differences. Conversely, our

234 finding could indicate that researchers interested in sex differences are primarily the ones

235 following NIH guidelines.

236

237 *Testing for interactions in a factorial design*

238   Testing whether the sexes respond differently to a treatment requires statistical

239 comparison between the two effects, which is typically done by testing for a sex $\times$ treatment

240 interaction. In our analysis, however, tests for interactions were missing 71% of the time (Fig.

241 2C, D). In these cases, the most common method for detecting differential effects of treatment

12

242   was to compare qualitatively the conclusions drawn for each sex; that is, to assert that a *p* value

243   below 0.05 for one sex but not the other (Fig. 2B) represents a meaningful difference between

244   the effects. But null hypothesis significance testing does not allow for such conclusions

245   (Cumming, 2012). This error, and the frequency with which it is made, has been covered in

246   multiple publications; for example Gelman & Stern (2006) titled their commentary "The

247   difference between 'significant' and 'not significant' is not itself statistically significant." Makin &

248   de Xivry (2019) included the error in their "Top ten list of common statistical mistakes". In an

249   analysis of 520 articles in the field of neuroscience, Nieuwenhuis et al. (2011) found that the

250   error was committed in about half of articles containing a factorial design. The current analysis

251   showed that, even a decade later, the frequency of this error in the field of neuroscience has not

252   changed (Fig. 2D), at least when sex is one of the factors under consideration. The frequency of

253   the error was even higher in most of the other disciplines, particularly Physiology and

254   Reproduction, for which we found that authors never tested for interactions.

255   Statements such as the following, usually made without statistical evidence, were

256   common: "The treatment increased expression of gene X in a sex-dependent manner"; "Our

257   results demonstrate that deletion of gene X produces a male-specific increase in the behavior";

258   "Our findings indicate that females are more sensitive to the drug than males". In some of these

259   cases, the terms "sex-specific", "sex-dependent" or "sexual dimorphism" were used in the title of

260   the article despite a lack of statistical evidence supporting the claim. In many of these articles,

261   some of which stated that finding a sex difference was the major goal of the study, the sexes

262   were not statistically compared at all. Thus, a lack of statistical evidence for sex-specific effects

263   did not prevent authors from asserting such effects. In fact, we found that authors failing to test

264   for interactions were far more likely to claim sex-specific effects than not (88% vs. 12%; Fig.

265   2F); they were also more likely to do so than were authors that did test for interactions (88% vs.

266   62%; Table S3). Together, these results suggest a bias toward finding sex differences. In the

267   absence of evidence, differences were claimed more often than not. A bias toward finding sex

268   differences, where there are none, could artificially inflate the importance of sex in the reporting

269   of biological data. Given that findings of sex × treatment interactions are rare in the human

270   clinical literature, with false positives outnumbering false negatives (Wallach et al., 2016), and

271   given also that sex differences are often misrepresented to the public (Maney, 2014), it is

272   especially important to base conclusions from preclinical research on solid statistical evidence.

273

274   *Pooling across sex*

275   The set of articles we analyzed was pre-screened by Woitowich (2020) to include only

276   studies in which sex was considered as a variable. Nonetheless, even in this sample, data were

277   often pooled across sex for some of the analyses (Fig. 3A). In a majority of these articles,

278   authors did not test for a sex difference before pooling (Fig. 3B). Thus, for at least some

279   analyses represented here, the data were not disaggregated by sex, sex was not a factor in

280   those analyses, and we do not know whether there might have been a sex difference. Even

281   when authors did test for a sex difference before pooling, the relevant statistics were often not

282   presented. Finding and reporting a significant sex difference did not seem to reduce the

283   likelihood that the sexes would be pooled. Note that the original sample of 720 articles in the

284   study by Woitowich et al. included 251 articles in which sex was either not specified or the

285   sexes were pooled for all analyses (Fig. S1). Thus, the issue is more widespread than could be

286   represented in the current study. Pooling is not consistent with the NIH mandate to disaggregate

287   data by sex, and can prevent detection of meaningful differences. We note further that effect

288   sizes were not reported before pooling; in addition to p values, effect sizes would be valuable for

289   any assessment of whether data from males and females can be pooled without masking a

290   potentially important difference (Beltz et al., 2019; Diester et al., 2019).

291

292   *Correcting for multiple comparisons*

293   In their article on "Ten statistical mistakes…," Makin and de Xivry (2019) list another

294   issue that we found to be prevalent, although we did not collect systematic data on it. Many

295   authors compared a large number of dependent variables across sex without correcting for

296   multiple comparisons. The omission of the correction increases the risk of false positives, that

297   is, making a type I error, which would result in over-reporting of significant effects. This problem

298   is particularly important for researchers trying to comply with SABV, who may feel compelled to

299   test for sex differences in every measured variable. For example, we noted articles in which

300   researchers measured expression of multiple genes in multiple tissues at multiple time points,

301   resulting in a large number of comparisons across sex. In one such study, authors made 90

302   separate comparisons in the same set of animals and found 5 significant differences, which is

303   exactly the number one would expect to find by chance. The prevalence of this issue is difficult

304   to estimate because opinions vary about when corrections are necessary; nonetheless,

305   omission of such corrections, when they are clearly needed, is likely contributing to over-

306   reporting of sex differences broadly across disciplines.

307

308   *Usage of "sex" and "gender"*

309   We found that a large majority of studies on non-human animals used "sex" to refer to

310   the categorical variable comprising females and males (Fig. 4). In eight articles, we noted usage

311   of the word "gender" for non-human animals. This usage appears to conflict with current

312   recommendations regarding usage of "gender", that is, gender should refer to socially

313   constructed identities or behaviors rather than biological attributes (Clayton & Tannenbaum,

314   2016; Holmes & Monks, 2019; Woitowich & Woodruff, 2019). We did not, however, investigate

315   the authors' intended meaning of either term. Although definitions of "gender" vary, the term

316   might be appropriate for non-human animals under certain circumstances, such as when the

317   influence of social interactions is a main point of interest (Cortes et al., 2019). Operational

318   definitions, even for the term "sex", are important and, in our experience conducting this study,

319   almost never included in publications. As others have done (e.g., Duchesne et al., 2020; Cortes

320   et al., 2019; Holmes & Monks, 2019; Johnson et al., 2009), we emphasize the importance of

321   clear operational definitions while recognizing the limitations of binary categories.

322

323   *Limitations of this study*

324   This study was underpowered for examining these issues within any particular discipline.

325   For most disciplines, fewer than a dozen articles were in our starting sample; for Neuroscience

326   and Reproduction, only nine. As a result, after we coded the articles, some categories contained

327   few or no articles in a given discipline (see Table S3). The within-discipline analyses, particularly

328   the pie charts in Fig. 3B, should therefore be interpreted with caution. Firm conclusions about

329   whether a particular practice is more prevalent in one discipline than another cannot be drawn

330   from the data presented here.

331   As is the case for any analysis, qualitative or otherwise, our coding was based on our

332   interpretation of the data presentation and wording in the articles. Details of the statistical

333   approach were sometimes left out, leaving the author's intentions ambiguous. Although our

334   approach was as systematic as possible, a small number of articles may have been coded in a

335   way that did not completely capture those intentions. We believe our sample size, particularly in

336   the overall analyses across disciplines, was sufficient to reveal the important trends.

337

338   *Conclusion*

339   SABV has been hailed as a game-changing policy that is already bringing previously

340   ignored sex-specific factors to light, particularly for females. In this study, we have shown that a

341   substantial proportion of claimed sex differences, particularly sex-specific effects of

342   experimental manipulations, are not supported by statistical evidence. Although only a minority

343   of studies that include both sexes actually report data by sex (Woitowich et al., 2020), our

16

344 findings suggest that when data *are* reported by sex, critical statistical analyses are often

345 missing and the findings likely to be interpreted in misleading ways. Note that in most cases, our

346 findings do *not* indicate that the conclusions were incorrect; they may have been supported by

347 appropriate statistical analyses. Our results emphasize the need for resources and training,

348 particularly those relevant to the study designs and analyses that are commonly used to test for

349 sex differences. Such training would benefit not only the researchers doing the work, but also

350 the peer reviewers, journal editors, and program officers who have the power to hold

351 researchers to a higher standard. Without better awareness of what can and cannot be

352 concluded from separate analysis of males and females, SABV may have the undesired effect

353 of reducing, rather than enhancing, rigor and reproducibility.

354

355 *Materials and Methods*

356 　　We conducted our analysis using journal articles from a list published by Woitowich et al.

357 (2020). In their study, which was itself based on a study by Beery and Zucker (2011), the

358 authors selected 720 articles from 34 journals in nine biological disciplines. Each discipline was

359 represented by four journals, with the exception of Reproduction, which was represented by two

360 (Table 1). To be included, articles needed to be primary research articles not part of a special

361 issue, describe studies conducted on mammals, and be published in English. For each journal,

362 Woitowich et al. selected the first 20 articles meeting these criteria published in 2019 (40 articles

363 for Reproduction). For most disciplines, all articles were published between January and April,

364 2019; for others, articles could have been published as late as June, August, or October for

365 Endocrinology, Behavioral Physiology, and Behavior, respectively.

366 　　Woitowich et al. (2020) coded each article with respect to whether it contained data

367 analyzed by sex, defined as either that the sexes were kept separate throughout the analysis or

368 that sex was included as a fixed factor or covariate. Of the original 720 articles analyzed, 151

369 met this criterion. We began our study with this list of 151 articles. Four articles were excluded

370   because they contained data from only one sex, with animals of the other sex used as stimulus

371   animals or to calculate sex ratios.

372   All articles were initially scanned by the first author (YGS) to ascertain the experimental

373   designs in each, and a subset of the articles was discussed between the authors to develop an

374   analysis strategy. All articles were then coded by the second author (DLM). The final strategy

375   consisted of four decision trees (Table S1) used to assign articles to hierarchical categories

376   pertaining to each of four central questions (see below). Each article was assigned to only one

377   final category per question (Table S2). A subset of the articles was independently coded by

378   YGS and any discrepancies discussed between the authors until agreement was reached.

379   *Question 1: Was a sex difference reported?* Because we were interested in the

380   frequency with which sex differences were found, we first identified articles in which the sexes

381   were explicitly compared. We counted as a comparison any of the following: (1) sex was a fixed

382   factor in a statistical model; (2) sex was included as a covariate in a statistical model *and* a p

383   value for the effect of sex was reported; (3) a *p* value for a comparison of means between males

384   and females was presented; (4) the article contained wording suggestive of a comparison, e.g.

385   "males were larger than females". We also included articles with wording suggestive of a sex

386   difference in response to a treatment, for example "the treatment affected males *but not*

387   females" or "the males responded to treatment, *whereas* the females did not", or "the treatment

388   had a *sex-specific* effect". Similarly, we included here articles with language referring to a non-

389   difference, for example "we detected no sex differences in size" or "the response to treatment

390   was similar in males and females." Articles in which sex was included as a covariate for the

391   purposes of controlling for sex, rather than comparing the sexes, were not coded as having

392   compared the sexes (see Beltz et al., 2019). When the sexes were compared but no results of

393   those comparisons, e.g., *p* values, were reported, that omission was noted and the article was

394   coded accordingly. Each article in which the sexes were compared was then further coded as

18

395   either reporting a sex difference or not, and if so, whether a sex difference was mentioned in the

396   title or abstract.

397   *Question 2: Did the article contain a study with a factorial design?* We looked for studies

398   with a 2X2 factorial design (Fig. 2A) in which sex was one of the factors. Sex did not need to be

399   explicitly identified as a fixed factor; we included here all studies comparing across levels of one

400   factor that comprised females and males with each of those levels. In some cases that factor

401   was a manipulation, such as a drug treatment or a gene knockout; these factors also included

402   variables such as age, season, presentation of a stimulus, etc. For simplicity, we refer in this

403   article to the other factor as "treatment". Any article containing at least one such study was

404   coded as having a factorial design. The other articles were coded as containing no comparisons

405   across sex or as containing group comparisons across sex. The latter category included studies

406   with sex as a covariate of interest in a model such as a multiple regression, if the authors were

407   not making any claims about potential interactions between sex and other variables.

408   For studies with a factorial design, we further coded the authors' strategy of data

409   analysis. First, we noted whether authors tested for an interaction between sex and treatment.

410   We included one study in which the effect of treatment was explicitly compared across sex

411   using a method other than a classic ANOVA (the magnitude of the differences between treated

412   and control groups were compared across sex). If authors tested statistically whether the effect

413   of treatment depended on sex, we noted the outcome of that test and the interpretation. Articles

414   containing no tests for interactions were assigned to one of several sub-categories in the

415   following order (coded as the first category on this list for which the description was met for any

416   analysis in the article): tested for effects treatment within sex, tested for effects of sex within at

417   least one level of treatment, or tested for main effects of sex only. Within each of those

418   categories we further coded the outcome/interpretation, e.g., sex difference or no sex

419   difference. Any articles containing statements that the sexes responded differently to treatment

420   or that the response was "sex-specific" were coded as reporting a sex-specific effect. We also

19

421   noted when authors reported an absence of such a result. Articles not comparing across sex at

422   all, with statistical evidence or by assertion, were coded accordingly.

423   *Question 3: Did the authors pool males and females?* We assigned articles to one of

424   three categories: analyzed males and females separately throughout, included sex in the

425   statistical model for at least some analyses (with the rest analyzed separately), or pooled for at

426   least some analyses. The second category, included sex in the model, included articles in which

427   AIC or similar statistic was used to choose among models that included sex, although sex may

428   not have been in the model ultimately chosen. This category did not distinguish between

429   analyses including sex as a fixed factor vs. a covariate; this distinction is noted where relevant

430   in Table S2. Any article containing pooled data was coded as pooled, even if some analyses

431   were conducted separately or with sex in the model. For articles that pooled, we further noted

432   whether the authors tested for a sex difference before pooling and, if so, whether *p* values or

433   effect sizes were reported.

434   *Question 4: Did the authors use the term "sex" or "gender"?* We searched the articles for

435   the terms "sex" and "gender" and noted whether the authors used one or the other, both, or

436   neither. Terms such as "sex hormones" or "gender role", which did not refer to sex/gender

437   variables in the study, were excluded from this assessment. For the articles using "gender" we

438   further noted when the term was used for non-human animals.

439   To visualize the data, we used river plots (Weiner, 2017), stacked bar graphs, and pie

440   charts based on formulae and data presented in Table S3.

441

446

447   *Competing interests*: The authors have no competing interests.

448

449   *References*

450   Becker, J. B., Prendergast, B. J., & Liang, J. W. (2016). Female rats are not more variable than
451   male rats: a meta-analysis of neuroscience studies. Biology of Sex Differences, 7(1), 1-
452   7.

453   Beltz, A. M., Beery, A. K., & Becker, J. B. (2019). Analysis of sex differences in pre-clinical and
454   clinical data sets. Neuropsychopharmacology, 44(13), 2155-2158.

455   Beery, A. K., & Zucker, I. (2011). Sex bias in neuroscience and biomedical research.
456   Neuroscience & Biobehavioral Reviews, 35(3), 565-572.

457   Clayton, J. A. (2018). Applying the new SABV (sex as a biological variable) policy to research
458   and clinical care. Physiology & Behavior, 187, 2-5.

459   Clayton, J. A., & Collins, F. S. (2014). Policy: NIH to balance sex in cell and animal studies.
460   Nature News, 509(7500), 282.

461   Clayton, J. A., and Tannenbaum, C. (2016). Reporting sex, gender, or both in clinical research?
462   JAMA 316, 1863–1864.

463   Cortes, L. R., Cisternas, C. D., & Forger, N. G. (2019). Does gender leave an epigenetic imprint
464   on the brain?. Frontiers in Neuroscience, 13, 173.

465   Cumming, G. (2012) Understanding the new statistics: effect sizes, confidence intervals, and
466   meta-analyses. New York, NY: Routledge.

467   Diester, C. M., Banks, M. L., Neigh, G. N., & Negus, S. S. (2019). Experimental design and
468   analysis for consideration of sex as a biological variable. Neuropsychopharmacology,
469   44(13), 2159-2162.

470   Duchesne, A., Pletzer, B., Pavlova, M. A., Lai, M. C., & Einstein, G. (2020). Bridging Gaps
471   Between Sex and Gender in Neurosciences. Frontiers in Neuroscience, 14, 561.

472   Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not
473   itself statistically significant. The American Statistician, 60(4), 328-331.

474   Holmes, M. M., & Monks, D. A. (2019). Bridging sex and gender in neuroscience by shedding a
475   priori assumptions of causality. Frontiers in Neuroscience, 13, 475.

476   Johnson, J. L., Greaves, L., and Repta, R. (2009). Better science with sex and gender:
477   facilitating the use of a sex and gender-based analysis in health research. International
478   Journal for Equity in Health, 8, 14.

479   Makin, T. R., & de Xivry, J. J. O. (2019). Science Forum: Ten common statistical mistakes to
480   watch out for when writing or reviewing a manuscript. eLife, 8, e48175.

481   Maney, D. L. (2016). Perils and pitfalls of reporting sex differences. Philosophical Transactions
482   of the Royal Society B: Biological Sciences, 371(1688), 20150119.

483   Maney, D. L. (2014). Just like a circus: The public consumption of sex differences. Current
484   Topics in Behavioral Neuroscience, 19, 279-296.

485   Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of
486   interactions in neuroscience: a problem of significance. Nature Neuroscience, 14(9),
487   1105.

488   NIH, Consideration of sex as a biological variable in NIH-funded research. National Institutes of
489   Health Notice Number: NOT-OD-15–102, issued 6-9-2015.

490   Potluri, T., Engle, K., Fink, A. L., Vom Steeg, L. G., & Klein, S. L. (2017). Sex reporting in
491   preclinical microbiological and immunological research. mBio, 8(6), e01868-17.

492   Radke, A. K., Sneddon, E. A., & Monroe, S. C. (2021). Studying Sex Differences in Rodent
493   Models of Addictive Behavior. Current Protocols, 1(4), e119.

494   Shansky, R. M., & Murphy, A. Z. (2021). Considering sex as a biological variable will require a
495   global shift in science culture. Nature Neuroscience, 24(4), 457-464.

496   Sugimoto, C. R., Ahn, Y. Y., Smith, E., Macaluso, B., & Larivière, V. (2019). Factors affecting
497   sex-related reporting in medical research: a cross-disciplinary bibliometric analysis. The
498   Lancet, 393(10171), 550-559.

499   Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. & Schiebinger, L. Sex and gender analysis
500   improves science and engineering. Nature 575, 137–146 (2019).

501   Wallach, J. D., Sullivan, P. G., Trepanowski, J. F., Steyerberg, E. W., & Ioannidis, J. P. (2016).
502   Sex based subgroup differences in randomized controlled trials: empirical evidence from
503   Cochrane meta-analyses. BMJ, 355, i5826.

504   Weiner, J. (2017). Riverplot: Sankey or Ribbon Plots. https://CRAN.R-
505   project.org/package=riverplot.

506   Will, T. R., Proaño, S. B., Thomas, A. M., Kunz, L. M., Thompson, K. C., Ginnari, L. A., ... &
507   Meitzen, J. (2017). Problems and progress regarding sex bias and omission in
508   neuroscience research. eNeuro, 4(6), e0278.

509   Woitowich, N. C., & Woodruff, T. K. (2019). Opinion: Research community needs to better
510   appreciate the value of sex-based research. Proceedings of the National Academy of
511   Sciences, 116(15), 7154-7156.

512   Woitowich, N. C., Beery, A., & Woodruff, T. (2020). Meta-Research: A 10-year follow-up study
513   of sex inclusion in the biological sciences. eLife, 9, e56344.