RESEARCH ARTICLE

EASP WILEY

Empirically derived guidelines for effect size interpretation in social psychology

Andrey Lovakov[1] | Elena R. Agadullina[2]

[1]Center for Institutional Studies, National Research University Higher School of Economics, Moscow, Russia

[2]School of Psychology, National Research University Higher School of Economics, Moscow, Russia

Correspondence
Andrey Lovakov, Center for Institutional Studies, National Research University Higher School of Economics, 20 Myasnitskaya Ulitsa, Moscow 101000, Russia.
Email: lovakov@gmail.com

Abstract

This study estimates empirically derived guidelines for effect size interpretation for research in social psychology overall and sub-disciplines within social psychology, based on analysis of the true distributions of the two types of effect size measures widely used in social psychology (correlation coefficient and standardized mean differences). Analysis of empirically derived distributions of 12,170 correlation coefficients and 6,447 Cohen's d statistics extracted from studies included in 134 published meta-analyses revealed that the 25th, 50th, and 75th percentiles corresponded to correlation coefficient values of 0.12, 0.24, and 0.41 and to Cohen's d values of 0.15, 0.36, and 0.65 respectively. The analysis suggests that the widely used Cohen's guidelines tend to overestimate medium and large effect sizes. Empirically derived effect size distributions in social psychology overall and its sub-disciplines can be used both for effect size interpretation and for sample size planning when other information about effect size is not available.

KEYWORDS
Cohen's d, correlation, effect size, sample size

1 | INTRODUCTION

The calculation of the effect size is necessary for research in psychology today and the American Psychological Association recommends the reporting of effect sizes in publications (Wilkinson, 1999). First, effect size is important because it enables the strength of empirically identified relationships or differences to be estimated, which can help researchers understand if they have a practical meaning. Second, effect size is required for a priori power analysis and calculation of sample sizes of sufficient power in future research.

The most widely used thresholds by which the effect size is interpreted as small, medium, or large are those proposed by Cohen (1988, 1992). In particular, $d = 0.20$ or $r = 0.10$ is interpreted as small effects, $d = 0.50$ or $r = 0.30$ as medium effects, and $d = 0.80$ or $r = 0.50$ as large effects. At the same time, according to Cohen (1988, p. 25), "the terms 'small', 'medium', and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being

employed in any given investigation". In other words, each field has different thresholds; in each case a "medium" effect is the average effect across a field, and "small" and "large" effects are equidistant to this.

Several recent pieces of research confirm this assumption and show that while for some research areas Cohen's guidelines are relatively adequate (Quintana, 2017), for others they are not. Hemphill (2003), based on analysis correlation coefficients from two meta-analyses, suggests revising Cohen's thresholds and categorizing $r < 0.20$ as small, r from 0.20 to 0.30 as medium, and $r > 0.30$ as large effects for treatment/experiments. Hemphill (2003) also noted that "it seems too simplistic to have a single set of empirical guidelines for interpreting the magnitude of correlation coefficients" (p. 79) and different sets of empirical guidelines are needed for different research areas. Gignac and Szodorai (2016) analyzed the empirical distribution of meta-analytically derived correlations in area of differential psychology and revealed that the 25th, 50th, and 75th percentiles corresponded to correlations of 0.11, 0.19, and 0.29 respectively. They

also reported that fewer than 3% of considered correlations were found to be as large as $r = 0.50$ (large effect according to Cohen's guidelines). Bosco et al. (2015) and Paterson et al. (2016) analyzed the empirical distribution of correlational effect sizes from applied psychology and organizational behavior/human resources literature respectively, and also showed that Cohen's benchmark does not fit well. Thus, determining the thresholds for each specific field is important and will contribute to improving the quality of research and providing more accurate interpretation of the results obtained. Empirically derived thresholds allow more accurate a priori power analyses to determine the sample size required to obtain effects with the necessary statistical power. This is especially important for cases when the research question is formulated for the first time and it is not possible to estimate the sample size from the most relevant studies.

The objective of the current study is to estimate effect size benchmarks in research in social psychology based on empirically derived effect size distribution. The first systematic attempt to estimate the effect sizes in different areas of social psychology was made by Richard et al. (2003). The authors analyzed 322 meta-analyses of social-psychological phenomena and showed that the average (most typical) effect size in social psychology corresponds to $r = 0.21$. At the same time, in different areas of social psychology the typical effect size can be either greater (for example, for group processes studies, $r = 0.32$) or less (for social influence studies, $r = 0.13$). These results indicate that the thresholds proposed by Cohen cannot be directly applied to the field of social psychology as Cohen's effect size benchmarks overestimate the boundaries, at least for medium effect.

Our analysis differs from Richard and colleagues' analysis on several key points. First, in contrast to Richard and colleagues, who analyzed publications from both social and personality psychology, we selected meta-analyses that related to social-psychological phenomena only. This decision was due to the fact that exclusion of effect sizes related to personality makes it possible to estimate more accurately the mean effect sizes for social psychological phenomena. Moreover, a similar analysis has already been carried out in differential psychology by Gignac and Szodorai (2016). Second, Richard and colleagues' analysis included only effect sizes obtained prior to 1997. Our analysis significantly expands the time interval and includes effect sizes obtained from 1928 to 2019. Thus, the analysis presented here was carried out on mainly new data.

Third, Richard and colleagues analyzed effect sizes obtained in meta-analyses. In contrast, we analyzed the primary effect sizes extracted from studies included in meta-analyses. This allowed us to construct the distribution of effect sizes obtained in the primary studies, which may not coincide with the distribution of effect sizes obtained from meta-analyses. The latter are often shifted to the middle of the distribution because they are the result of aggregation, which eliminates substantially low and high values. Including the effect sizes from primary studies rather than aggregated means in our analysis allowed us to collect information about the publication year of each effect size, the type of publication from which it was taken

(published or unpublished), the type of study (experimental or not), and sample size. Information about the publication year allowed us to assess whether the effect size values change over time, something previous research has clearly shown occurs with effect sizes relating to some socio-psychological phenomena (e.g. Eagly et al., 2020; Malle, 2006) and with other characteristics of the studies (e.g. sample size, data collection method) (Sassenberg & Ditrich, 2019).

Information about publication status and sample size allowed us to check for signs of publication bias. Where this is present, published effect sizes are higher than unpublished ones because significant and positive results are more likely to be published (Bakker et al., 2012; Fanelli, 2012; John et al., 2012; O'Boyle et al., 2017). This may bias the pooled effect-size estimates based on published research (Friese & Frankenbach, 2020). Another indicator of publication bias is the small-study effect, in which studies with small samples size (and therefore a larger standard error and lower power) report higher effect sizes (Sutton, 2006). Comparing the published and unpublished effect sizes and testing the correlation between effect size value and sample size enabled us to assess the possibility of potential publication bias in various themes of social psychology. Finally, due to the fact that experimental studies have a greater level of control under conditions and variables, the effect sizes obtained from them may be lower than in cross-sectional designs.

Fourth, since study design is important for interpreting effect size, we separately analyzed two types of effect size measures that are widely used in social psychology. The correlation coefficient (r) is typically used to represent the relationship between two continuous variables. Cohen's d (and closely related Hedges' g) is typically used for the difference representation of means between two groups (e.g., experimental/control, gender groups) (Lakens, 2013).

2 | DATA AND METHOD

2.1 | Selecting and excluding meta-analyses

We chose 29 journals indexed in the Social Sciences Citation Index (SSCI, Web of Science, Clarivate Analytics, ex-Thomson Reuters) and included in the "*Psychology, Social*" category. Then we identified all the papers in these journals that had the word "meta-analysis" or its variants in the title. This search was conducted in April 2020 and yielded 285 papers (see full list of meta-analyses: https://osf.io/8cgfe/).

Our aim was to estimate the effect sizes for social psychology only. Since several journals publish papers in both social psychology and differential psychology (e.g., *Journal of Personality and Social Psychology* or *Personality and Social Psychology Review*), each paper was examined by a second author and trained assistant to clarify whether it referred to social or differential psychology. If the meta-analysis was only about personality traits or other personality variables (e.g., self-esteem, locus of control) it was excluded from the analysis. If, however, the meta-analysis related to the relationship between personality variables and socio-psychological variables (for

example, personality traits and prejudices), it was included in the analysis. Additionally, we excluded articles about methodological issues related to meta-analytical procedures and criteria, tutorials, reports about corrigenda, comments, and cross-temporal meta-analyses, and those about group psychotherapy. In the last case, the decision was based on the fact that these meta-analyses are aimed at accessing the effectiveness of the method (which itself is not related to social psychology) and not group phenomena. Using these exclusion criteria, we identified 227 meta-analyses devoted to social psychology topics. The inter-rater agreement between the two coders was substantial, with Cohen's Kappa = 0.856, all discrepancies were discussed, and a joint decision was made on how to classify each article.

In the next step, each of the selected meta-analyses was assigned to one of 16 topics (e.g., antisocial behaviors, attitude, interpersonal relationships, etc.) based on the *Encyclopedia of Social Psychology* by Baumeister and Vohs (2007). Detailed descriptions of the topics are provided in the online Table S1. In 47 cases two topics were assigned to one meta-analysis if its thematic was at the intersection of different themes. For instance, the meta-analysis by Del Giudice (2011) was assigned to two thematic subgroups (Gender differences and Interpersonal relationships) and the meta-analysis by Rise et al. (2010) was assigned to Attitude and Self thematic subgroups. The topics were coded by the second author and research assistant, and the inter-rater agreement between two coders was substantial, with Cohen's Kappa = 0.731. All discrepancies were discussed and following discussion it was decided which topic or topics should be assigned to the meta-analysis.

After assigning the topics to meta-analysis, the text of each meta-analysis was analyzed for the presence of data from the original studies. Meta-analyses were excluded from the analysis if they were non-empirical, did not report raw effect sizes from the studies included, or did not report the *d*, *g*, or *r* values. The remaining 134 meta-analyses were included in the effect size distribution analysis. Seventy-one meta-analyses reported correlation (*r*) as a measure of effect size, four reported Fisher's *z* scores, and 59 reported standardized mean differences (Cohen's *d* or the closely related Hedges' *g*) (see Figure 1).

2.2 | Coding

For those meta-analyses with correlation as an effect size, the publication year of each effect size, the publication status (published or unpublished), and the sample size were also coded. Where a meta-analysis included this information in the table with descriptions of the studies, we used this information. If there was no information about the publication year of each effect size or the publication status in the table, we restored it from the references. If the necessary information could not be obtained from the tables or the references, it was coded as "not available". For those meta-analyses with Cohen's *d* as an effect size, the publication year of each effect size, the publication status, the sample size, and the type of study (experimental or non-experimental) were coded. This information was taken from descriptive tables or references or restored from the meta-analysis text. In cases where this was impossible, it was also coded as "not available". Four meta-analyses reported Fisher's *z* scores (460 effect sizes), which were converted to *r* by *fisherz2r* function using the *psych* R package (Revelle, 2017). When meta-analysis reported that Hedges' *g* values and sample sizes were available, they were transformed to Cohen's *d* values (Lakens, 2013, Formula 4). When sample sizes were not available, Hedges' *g* was used as the measure of effect size (311 effect sizes). All negative effect sizes were transformed into absolute values.
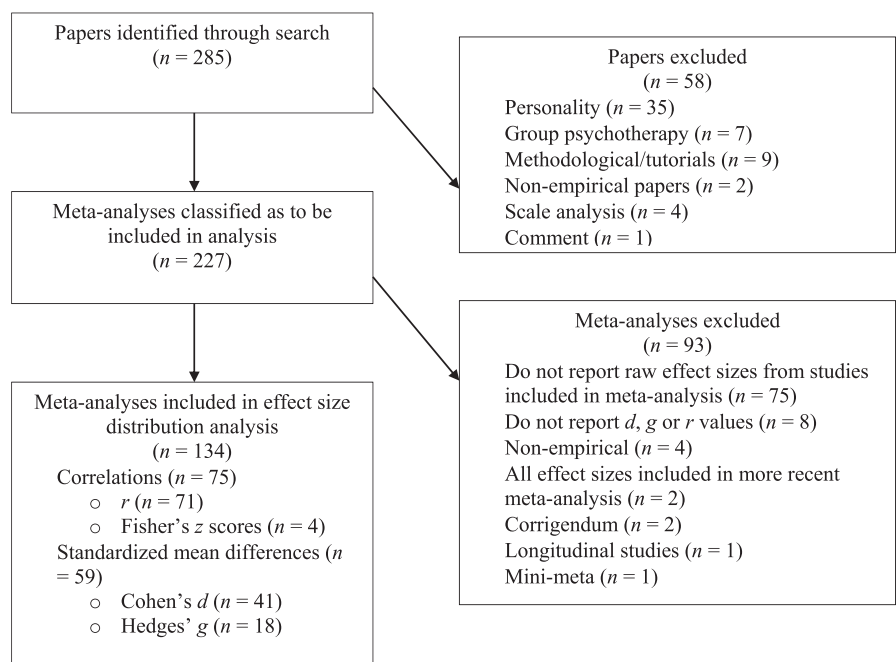

FIGURE 1 Meta-analysis inclusion flow chart for effect size distribution analysis

2.3 | Data analysis

Empirically derived thresholds for small, medium, and large effect sizes were identified as 25th, 50th, and 75th percentiles in empirically derived distributions of correlations and Cohen's *d* values. Relationships between the effect size value and year, sample size, unpublished status, and design were analyzed in two steps. In step one, we analyzed the bivariate relationships between the effect size value and year, sample size, unpublished status, and experimental design. For the relationship between the effect size and continuous variables (year and sample size), we made scatter plots and calculated the correlation coefficients. For the relationship between the effect size and dichotomous variables (unpublished status and experimental design), we plotted effect size distributions for each level of the dichotomous variable, calculated median effect sizes, and compared them using the Wilcoxon rank sum test. We also compared these distributions via the two-sample Kolmogorov-Smirnov test that tests the null hypothesis that two distributions were drawn from the same continuous distribution. This analysis was conducted both for the entire sample of the effect sizes representing social psychology overall and separately for each thematic subgroup. Since the analyzed variables are simultaneous characteristics of the studies in which the effect sizes were obtained, their influence can be confounding. In step two, multiple linear regression was used to test the relationship between the effect size value and year, publication status, sample size, and design (experimental or non-experimental). We entered multiple predictors simultaneously into one model to control for possible confounds between the variables being analyzed. For the entire sample of the effect sizes and for each thematic subgroup, a separate regression was run in which effect size was the dependent variable, whereas the year, sample size, status, and type of design (for Cohen's *d* effect sizes only) were predictors. However, some variables are significantly skewed, meaning that assumptions of the regression model can be violated. One possible way to overcome this is by bootstrapping the regression models. We used the *boot* and *boot.ci* functions from the *boot* R package (Canty & Ripley, 2017; Davison & Hinkley, 1997) to compute the 95% confidence intervals of 10,000 bootstrap estimates for the regression coefficients. If the 95% confidence interval included zero, then we interpreted the coefficient value as not significantly different from zero. If zero was not included in the 95% confidence interval, we interpreted the coefficient value as significantly different from zero.

The analysis was partly guided by Quintana's (2017) recommendations and script. The dataset and script to perform the analyses are freely available at the Open Science Framework: https://osf.io/8cgfe/

3 | RESULTS

3.1 | Correlations thresholds

A total of 12,170 correlation coefficients were extracted from studies or samples covered by 75 meta-analyses, and 8,796 (72.3%) correlations were positive. The 25th (small effect), 50th (medium effect),

and 75th (large effect) percentiles corresponded to correlation values of 0.12, 0.24, and 0.41 respectively (Table 1). Cohen's guideline for small effect was approximately consistent with the empirically derived threshold, whereas the guidelines for medium and large effects slightly overestimated the empirically derived thresholds (Table 2). Some 40.8% of correlations were equal to or greater than 0.30 (Cohen's threshold for medium effect) while only 14.6% were equal to or greater than 0.50 (Cohen's threshold for large effect). Figure 2a demonstrates the empirical distribution of the correlations, which was skewed positively (skew = 0.80; kurtosis = 0.18). The empirically derived distributions for correlations varied substantially between thematic subgroups (Table 3 and Figure 3). The lowest median was observed in the subgroup of effect sizes about Prejudice (median = 0.18), and the highest in the subgroups about Interpersonal relationships (median = 0.30) and Self (median = 0.29). The variation of the effect sizes was also different within thematic subgroups. The smallest variation in the effect sizes was observed in the subgroup of effect sizes about Prejudice (SD = 0.15), and the highest in the subgroups about Groups (SD = 0.25) and Social cognition (SD = 0.24). The standard deviation on the remainder of the thematic subgroups was about 0.20 (see Table 3).

4 | Cohen's *d* thresholds

3,944 Cohen's *d*s and 2,503 Hedges' *g*s (total of 6,447 effect sizes) were extracted from studies covered by 59 meta-analyses, and

TABLE 1 Percentiles associated with correlations (*r*) and Cohen's *d*

Percentile	Correlations (*r*)	Cohen's *d*
5	0.02	0.01
10	0.04	0.04
15	0.07	0.07
20	0.10	0.11
25	**0.12**	**0.15**
30	0.14	0.18
35	0.17	0.23
40	0.20	0.27
45	0.22	0.32
50	**0.24**	**0.36**
55	0.27	0.41
60	0.30	0.46
65	0.33	0.52
70	0.37	0.58
75	**0.41**	**0.65**
80	0.45	0.72
85	0.50	0.84
90	0.57	0.99
95	0.67	1.30

Note.: [a]Values in bold correspond to 25th, 50th, and 75th percentiles and are identified as thresholds for small, medium, and large effect sizes.

TABLE 2 Comparison of Cohen's and empirically derived thresholds for effects size

Thresholds	Correlation			Cohen's *d*		
	Small	Medium	Large	Small	Medium	Large
Cohen's thresholds	0.10	0.30	0.50	0.20	0.50	0.80
Empirically derived thresholds	0.12	0.24	0.41	0.15	0.36	0.65

FIGURE 2 The distribution of correlations (a) and Cohen's *d* (b). The dashed red lines represent the 25th, 50th, and 75th percentiles
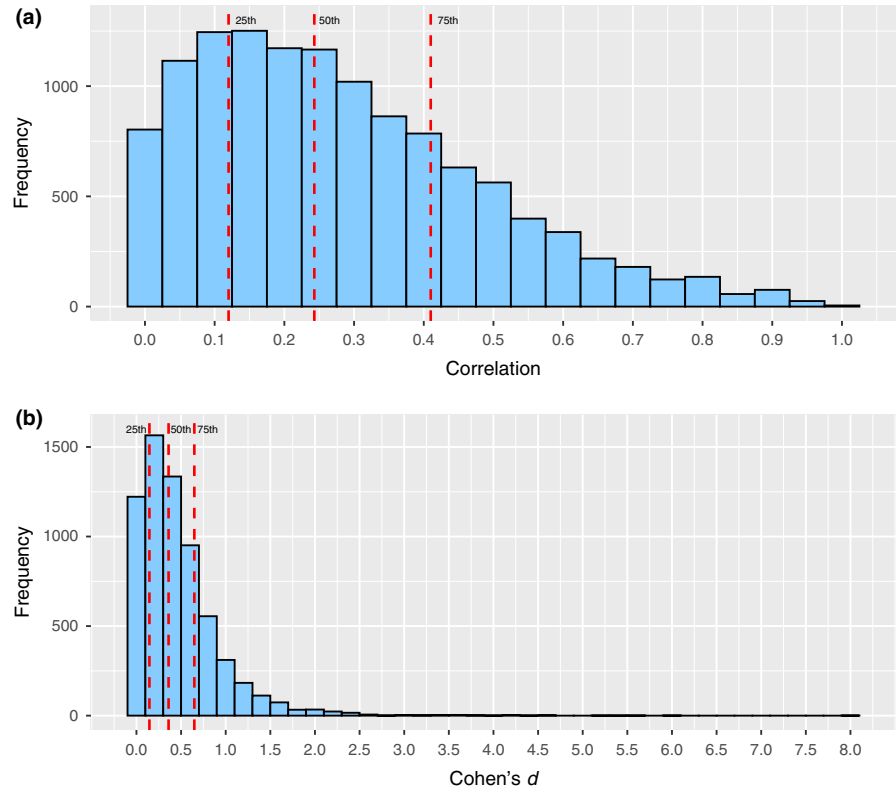
4,247 (65.9%) effect sizes were positive. The 25th, 50th, and 75th percentiles corresponded to Cohen's *d* values of 0.15, 0.36, and 0.65 respectively (Table 1). Cohen's guidelines for all levels of effect overestimated the empirically derived thresholds (Table 2), and the difference increased from small to large levels of effect. Only 36.4% of standardized mean differences were equal to or greater than 0.50 (Cohen's threshold for medium effect) and 16.3% of standardized mean differences were equal to or greater than 0.80 (Cohen's threshold for large effect). Figure 2b demonstrates the empirical distribution of Cohen's *d*s, which skewed positively (skew = 3.34; kurtosis = 24.98). The empirically derived distributions for Cohen's *d* also varied substantially between thematic subgroups (Table 2 and Figure 3). The lowest medians were observed in the subgroups of effect sizes about Gender differences (median = 0.22) and Interpersonal relationships (median = 0.28), and the highest in subgroups about Self (median = 0.48) and Social cognition (median = 0.50). The variation of the effect sizes was also different within thematic subgroups. The smallest variation in effect sizes was observed in the subgroup about Gender differences (SD = 0.31), and the highest in the subgroups about Self

(SD = 0.56) and Social cognition (SD = 0.52). The standard deviation in the remainder of the thematic subgroups was about 0.40 (see Table 3).

5 | Relationship between the effect size value and year, sample size, unpublished status, and design

We analyzed the relationship between the effect size value and the year it was received or published, sample size, unpublished status, and experimental or non-experimental design. Figure 4 shows scatter plots of correlational effect size and year for the entire sample of the effect sizes representing social psychology overall and seven thematic subgroups. Visual inspection of the scatter plots shows that over the entire sample the average effect size remains relatively stable, although in seven thematic subgroups several patterns of effect size dynamics can be observed. In two thematic subgroups (Groups and Interpersonal relationship) there were no pronounced consistent trends; the value of the effect sizes, on average, remained approximately the same throughout the period analyzed. Two other thematic subgroups (Prejudice and Attitude) showed a downward trend. The

Subgroup	Number of meta-analysis	Number of effect sizes	Median	Mean	SD
Correlation					
Groups	15	998	0.26	0.31	0.25
Interpersonal relationships	12	2,323	0.30	0.32	0.19
Prejudice	10	2,639	0.18	0.21	0.15
Self	10	1991	0.29	0.31	0.20
Attitude	9	2,352	0.26	0.29	0.20
Social cognition	9	1,248	0.27	0.33	0.24
Gender differences	5	585	0.23	0.27	0.20
Cohen's d					
Gender differences	12	1,261	0.22	0.30	0.31
Prejudice	10	1,370	0.34	0.44	0.40
Self	10	884	0.48	0.59	0.56
Interpersonal relationships	9	1,075	0.28	0.39	0.41
Social cognition	9	750	0.50	0.58	0.52
Attitude	5	428	0.39	0.47	0.38

Note.: Table shows the descriptive statistics only for those topics for which there is a sufficient number of meta-analyses and the effect sizes (for correlation: >= 5 meta-analysis & > 500 effect sizes; for Cohen's *d*: >= 5 meta-analysis & > 300 effect sizes).

average effect size decreased steadily over time (in the case of the Attitude subgroup, we are talking about the period since the mid-1970s, for which there is a sufficient number of observations in our sample). The three remaining thematic categories showed mixed patterns. In the Social cognition subgroup, the average effect size initially increased until the late 1980s, then declined. In the Self and Gender differences categories, meanwhile, the average effect size initially decreased before increasing. However, whereas in the case of the Self subgroup the changes were smooth and small, in the Gender differences category there was a rapid and significant increase.

Figure 5 shows scatter plots for Cohen's *d* and year. Again, over the entire sample, the average effect size remains relatively stable, but several different patterns can also be observed in different thematic subgroups. In the two thematic categories (Gender differences and Self), the average effect size remained approximately the same throughout the period analyzed. In the Prejudice category, a small but constant increase is observed. There are mixed patterns in the other three subgroups. In the Interpersonal relationship category, fluctuations can be observed in the early years, which, starting in the 1990s, turn into a decline before levelling out in the mid-2000s. In the two remaining subgroups (Social cognition and Attitude), an increase was seen at first, followed by a decrease in the average effect size. Thus, we can conclude that in most thematic subgroups, there is a dynamic of the average effect size over time. However, these dynamics can be different.

Figure 6 shows the scatter plots of correlational effect size and sample size. Across the entire sample, there is a weak negative correlation between sample size and effect size ($r = -0.07$, $p < .001$), that is, the larger the sample size, the lower the effect size. Visual inspection of the scatter plots for thematic subgroups shows that

several patterns of effect size dynamics can be observed. In three of the seven thematic subgroups, there is also a negative relationship (Prejudice: $r = -0.10$, $p < .001$; Social cognition: $r = -0.36$, $p < .001$; Gender differences: $r = -0.16$, $p < .001$). In three other subgroups there is no relationship (Groups: $r = -0.01$, $p = .651$; Self: $r = 0.01$, $p = .735$; Attitude: $r = 0.02$, $p = .455$). Only in one subgroup is there a positive relationship between the value of the effect size and the sample size (Interpersonal relationships: $r = 0.08$, $p < .001$). More consistent results are observed regarding the relationship between the Cohen's *d* and sample size. In all six thematic subgroups, Cohen's *d* is negatively associated with sample size (Interpersonal relationships: $r = -0.25$, $p < .001$; Prejudice: $r = -0.15$, $p < .001$; Self: $r = -0.20$, $p < .001$; Attitude: $r = -0.48$, $p < .001$; Social cognition: $r = -0.18$, $p < .001$; Gender differences: $r = -0.26$, $p < .001$; see Figure 7). Thus, it can be concluded that in the overwhelming majority of cases, large studies report smaller effect sizes.

Figures 8 and 9 show the distributions of published and unpublished effect sizes. In the case of correlational effect size, there was significant difference between distributions of published and unpublished effect sizes in the entire sample ($D = 0.053$, $p < .001$), but difference between medians is not very large in size (median [pub.] = 0.23, median [unpub.] = 0.25, $W = 8,271,646.5$, $p < .001$). There were also significant differences between distributions of published and unpublished effect sizes in four of the seven thematic subgroups (see Figure 8). In the Prejudice and Attitudes subgroups, the median of the published effect sizes was higher than the median of the unpublished ones (Prejudice: median [pub.] = 0.18, median [unpub.] = 0.16, $W = 623,802$, $p < .001$; and Attitude: median [pub.] = 0.27, median [unpub.] = 0.21, $W = 436,490.5$, $p < .001$), which may indicate the presence of a potential publication bias. However, in
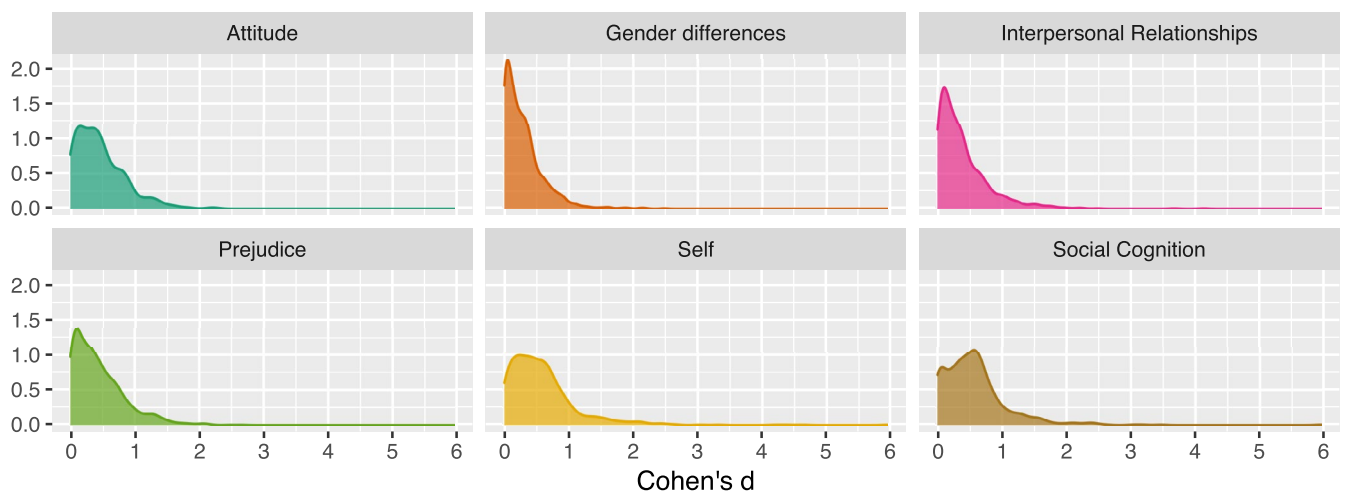
(a) Correlation
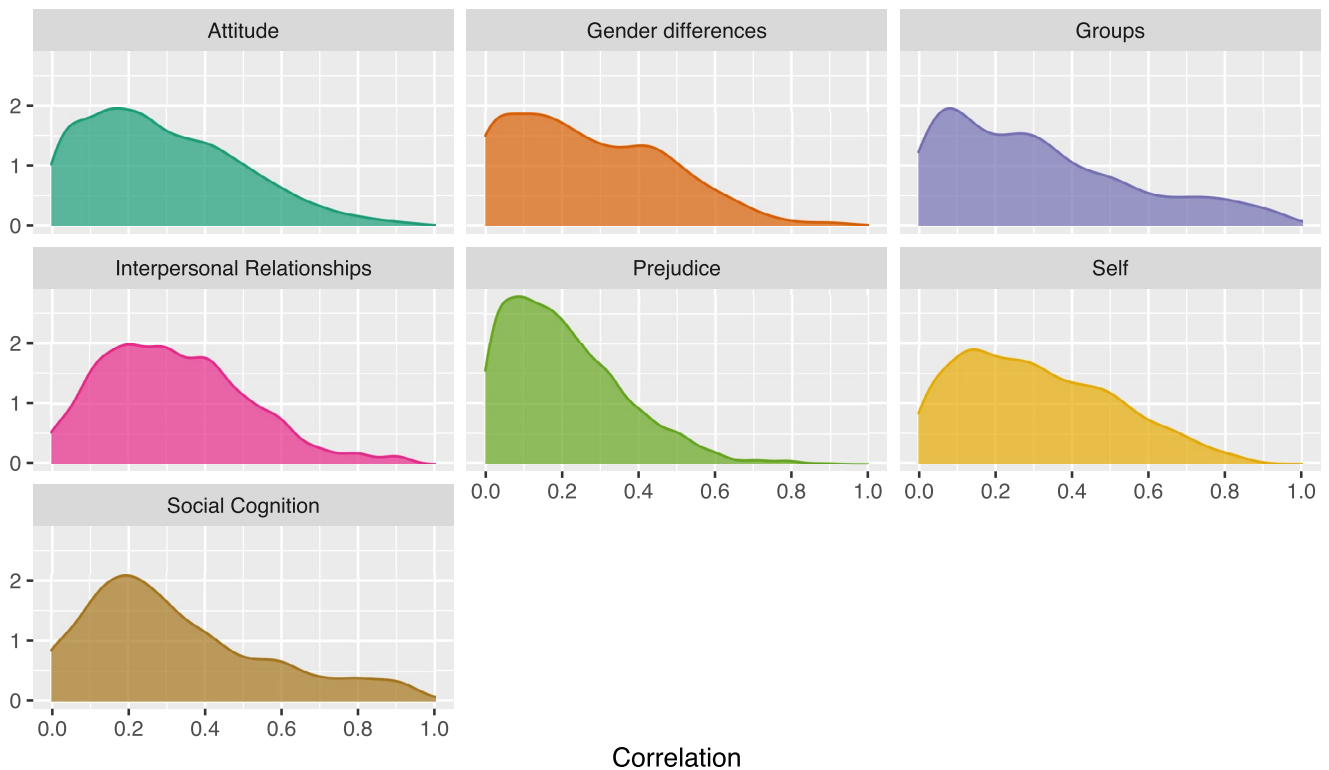
(b) Cohen's d

FIGURE 3 Density plots illustrating the distribution of correlations (a) and Cohen's *d* (b) in thematic subgroups. Figure shows the distributions only for those topics for which there are a sufficient number of meta-analyses and the effect sizes (for correlation: >=5 meta-analysis & >500 effect sizes; for Cohen's *d*: >=5 meta-analysis & >300 effect sizes (see Table 3))

the Interpersonal relationships and Gender differences subgroups, the median of the published effect sizes was lower than the median of the unpublished ones (Interpersonal relationships: median [pub.] = 0.29, median [unpub.] = 0.35, $W = 288,348$, $p < .001$; and Gender differences: median [pub.] = 0.17, median [unpub.] = 0.43, $W = 11,373$, $p < .001$). In the three remaining subgroups, the median of the published and unpublished effect sizes did not differ significantly (see Figure 8). In the case of Cohen's *d*s, there was also

significant difference between the distributions of the published and unpublished effect sizes ($D = 0.134$, $p < .001$) and the median of the published effect sizes was higher than for the unpublished ones (median [pub.] = 0.37, median [unpub.] = 0.27, $W = 2,704,325.5$, $p < .001$). There were also differences between the distributions of the published and unpublished effect sizes in three of the six thematic subgroups, and in all three of those subgroups, the median of the published effect sizes again was higher than for the
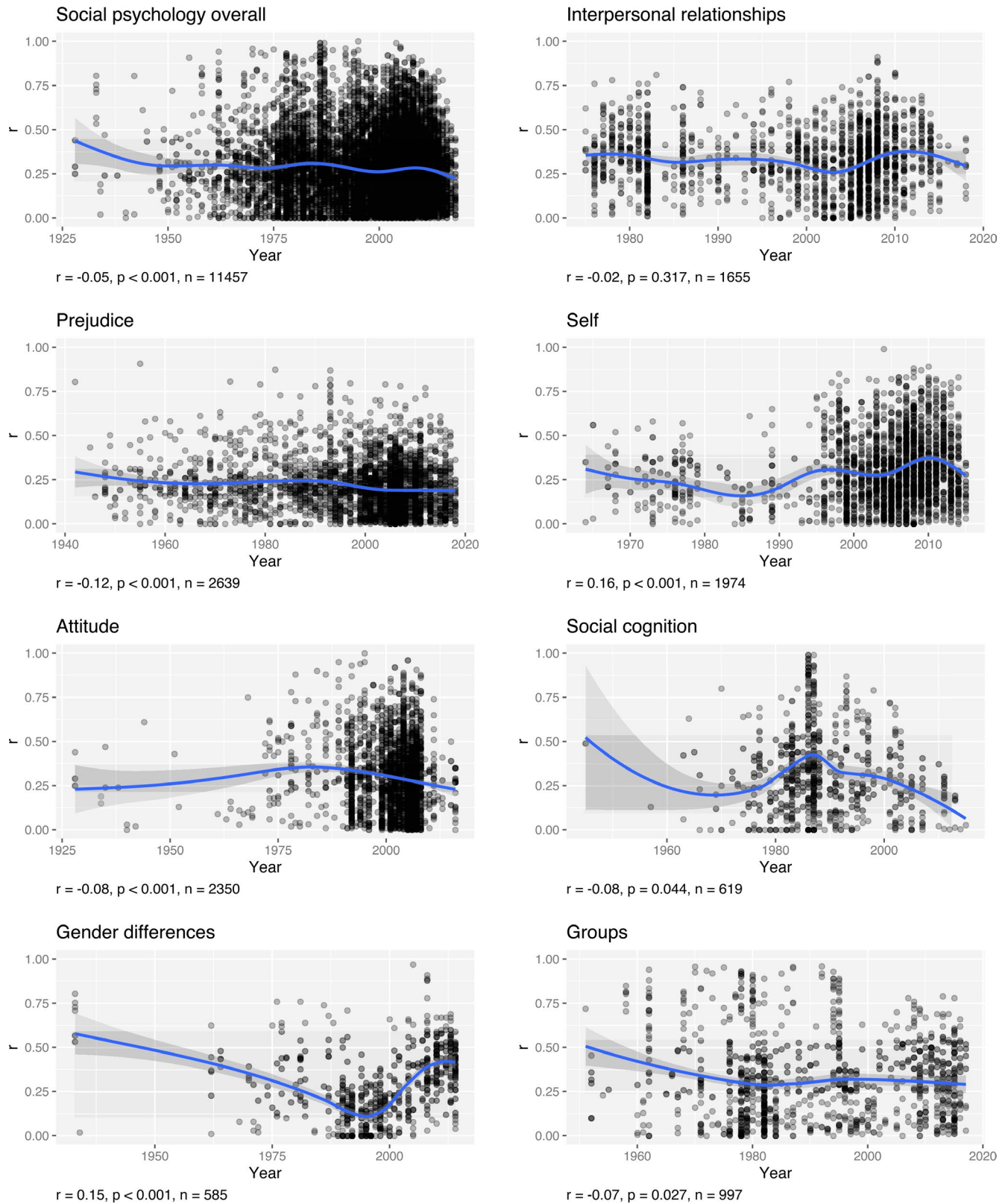
FIGURE 4 Relationship between correlational effect size and year (r—Pearson's product-moment correlation coefficient, p—p-value, n—sample size)

unpublished ones (Interpersonal relationships: median [pub.] = 0.28, median [unpub.] = 0.23, W = 49,787.5, p =.009; Prejudice: median [pub.] = 0.36, median [unpub.] = 0.30, W = 115,858.5, p =.021; Social cognition: median [pub.] = 0.53, median [unpub.] = 0.23, W = 27,847.5, p <.001). In the other three, the medians did not differ significantly (see Figure 9).
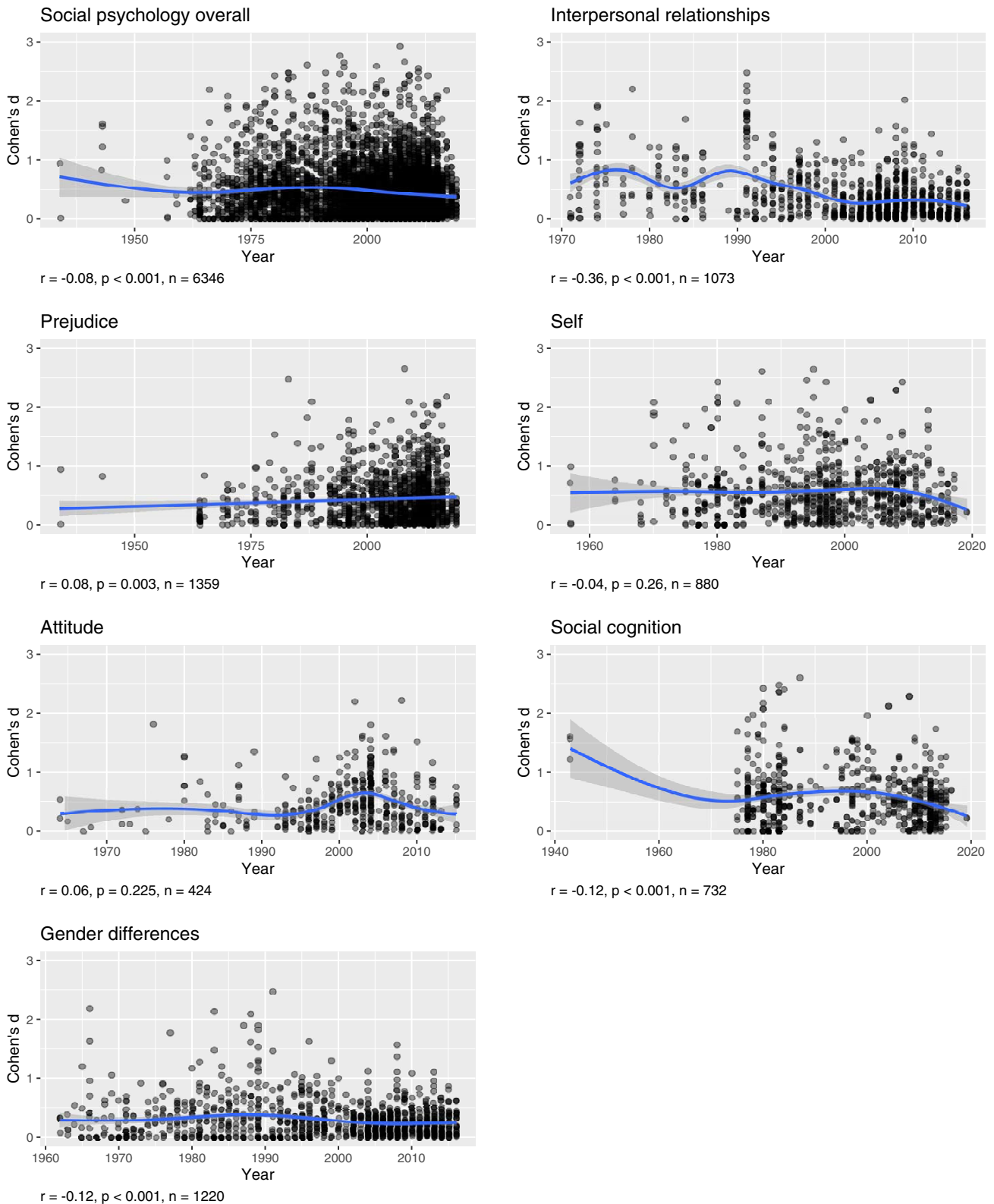
FIGURE 5 Relationship between Cohen's *d* and year (Cohen's *d*s higher than 3.0 are not shown for easier interpretation, *r*—Pearson's product-moment correlation coefficient, *p*—*p*-value, *n*—sample size)

Analysis of the relationship between the value of the effect size and experimental design was conducted only for Cohen's *d* effect sizes. Separating effect sizes by the study design from which they were obtained also led to differences between the distributions of the effect sizes ($D = 0.107$, $p <.001$). The median of the effect sizes obtained in the experimental studies was higher than that obtained in the non-experimental versions (median [exp.] = 0.40, median [non-exp.] = 0.31, $W = 4,843,858.5$, $p <.001$). In the Interpersonal relationships and Attitude
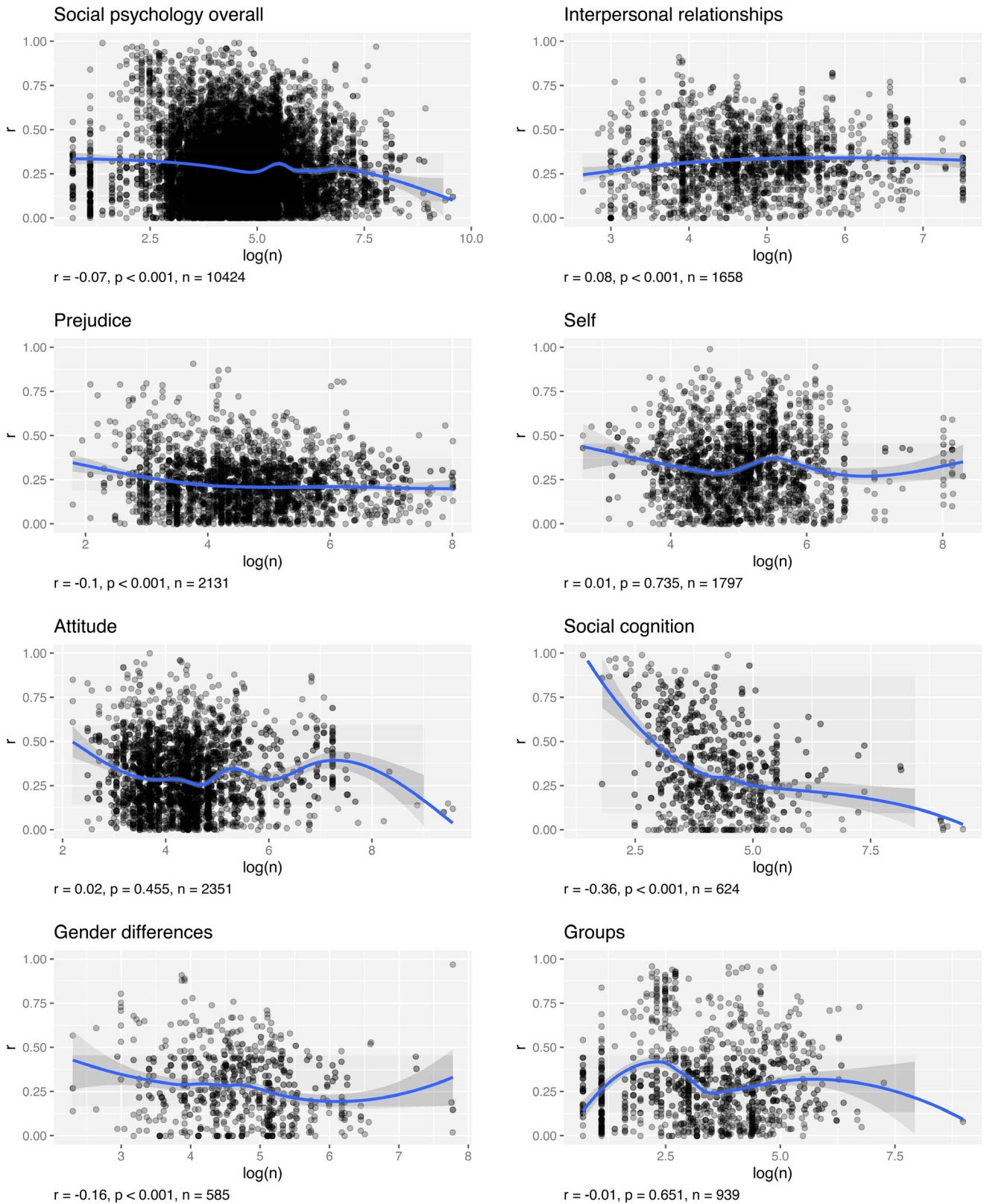
FIGURE 6 Relationship between value of correlational effect size and sample size (logarithm) (*r*—Pearson's product-moment correlation coefficient, *p*—*p*-value, *n*—sample size)

subgroups, the median of the effect sizes obtained in the experimental studies was also higher than that obtained in the non-experimental versions (Interpersonal relationships: median [exp.] = 0.67, median [non-exp.]

= 0.19, *W* = 119,348.5, *p* <.001; Attitude: median [exp.] = 0.35, median [non-exp.] = 0.24, *W* = 10,979, *p* =.001). In the Prejudice subgroup, however, the median of the experimental effect size was lower (median
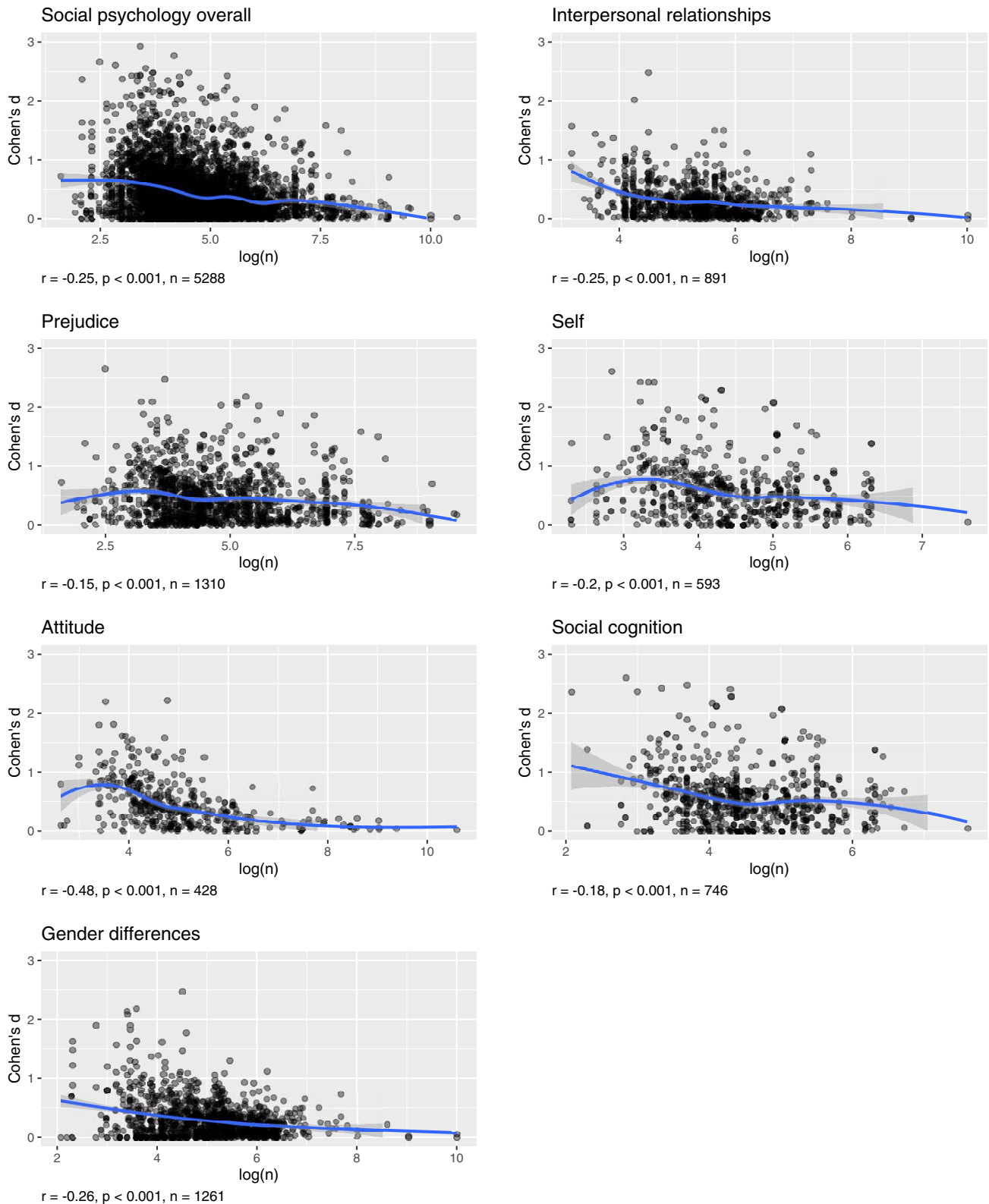
FIGURE 7 Relationship between Cohen's *d* and sample size (logarithm) (Cohen's *d*s higher than 3.0 are not shown for easier interpretation, *r*—Pearson's product-moment correlation coefficient, *p*—*p*-value, *n*—sample size)

[exp.] = 0.33, median [non-exp.] = 0.53, $W = 107,096$, $p < .001$). In the remaining two subgroups, there were no differences in medians of the experimental and non-experimental effect sizes (see Figure 10).

To control for possible confounds between the variables analyzed, we entered multiple predictors simultaneously into the same regression model. The bivariate analyses clearly showed
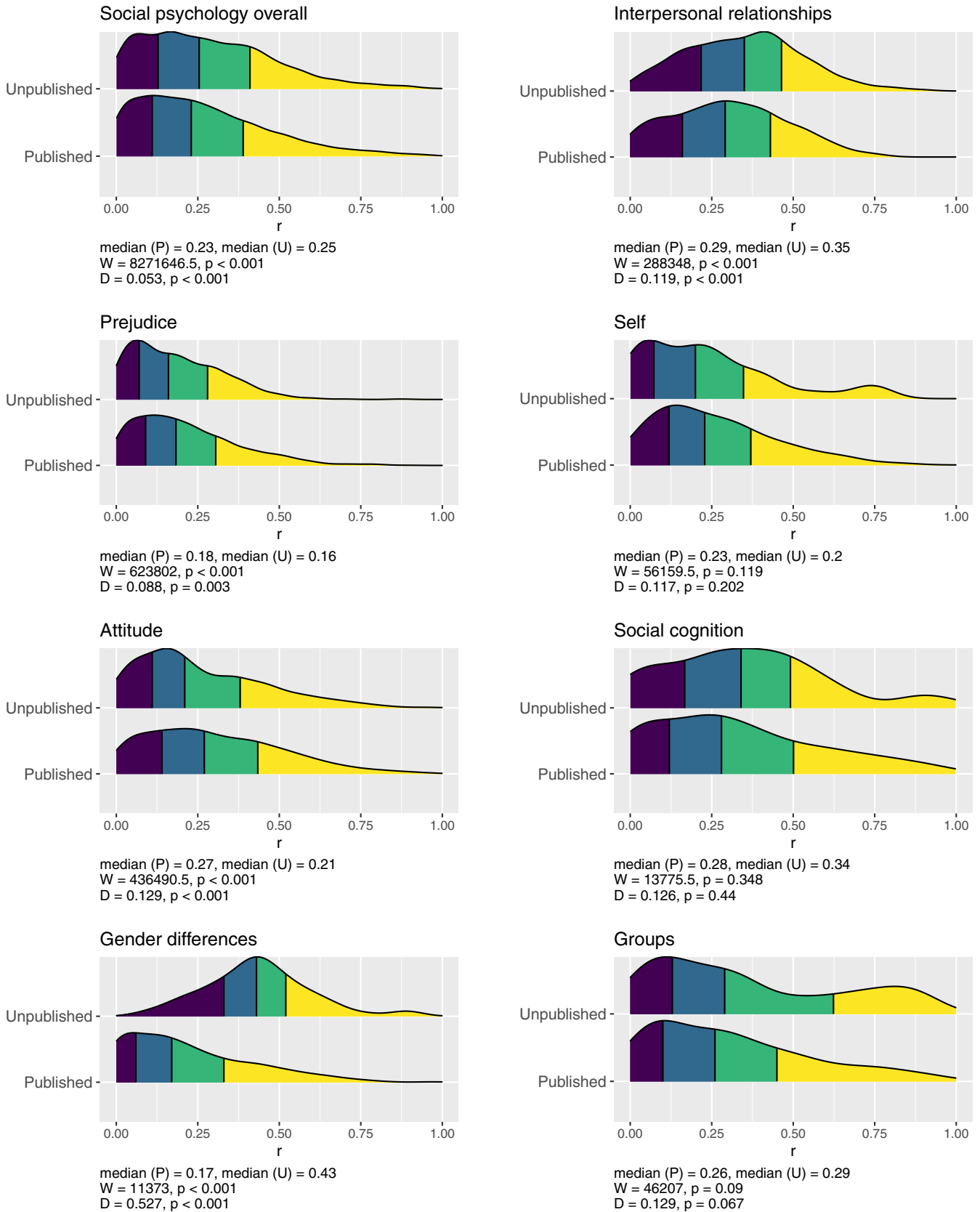
FIGURE 8 The distribution of published and unpublished correlations (colors show quartiles, *W*—Wilcoxon rank sum test with continuity correction, *D*—Two-sample Kolmogorov-Smirnov test, *p*—*p*-value)
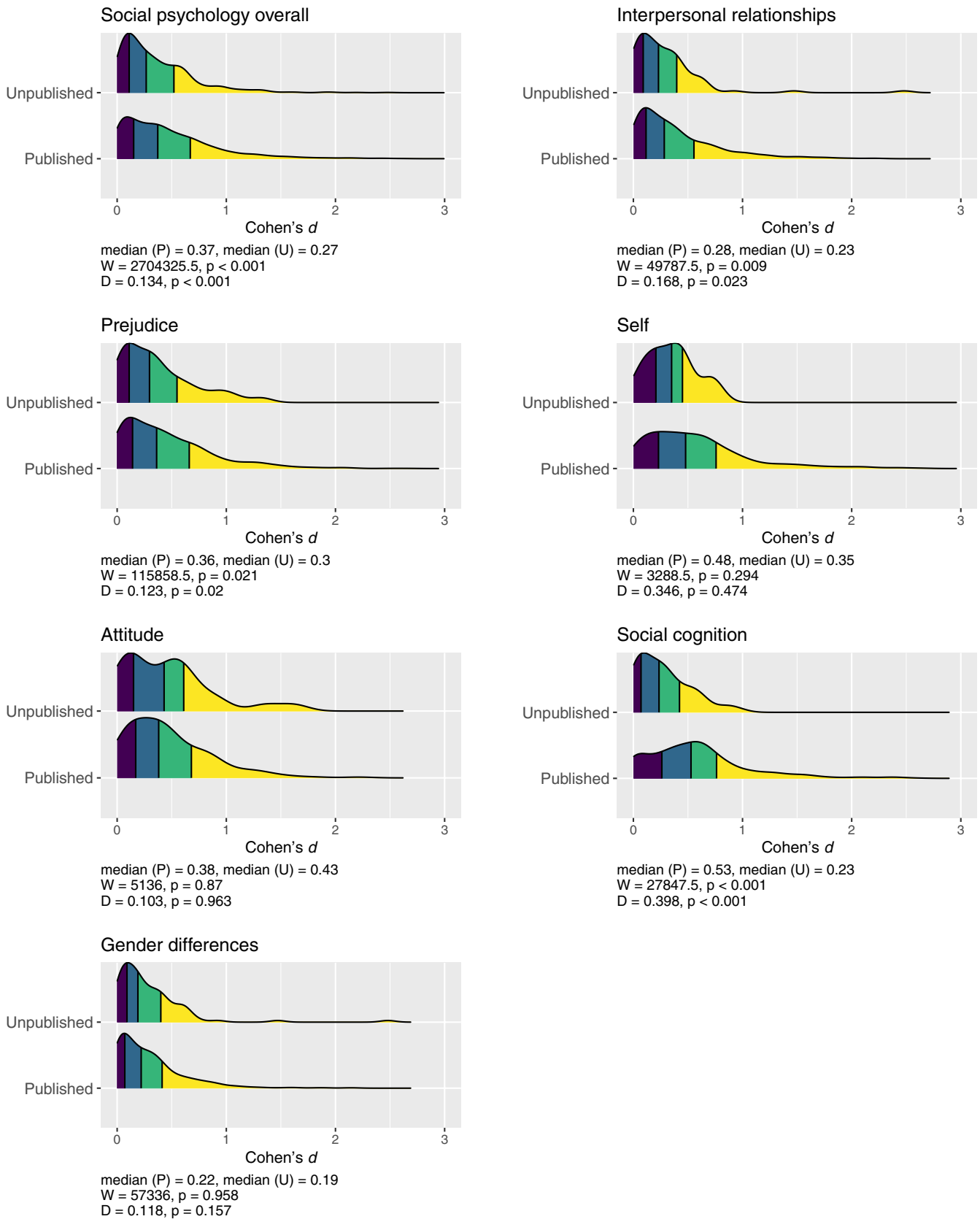
FIGURE 9 The distribution of published and unpublished Cohen's *d*s (Cohen's *d*s higher than 3.0 are not shown for easier interpretation, colors show quartiles, *W*—Wilcoxon rank sum test with continuity correction, *D*—Two-sample Kolmogorov-Smirnov test, *p*—*p*-value)

that the relationships between the effect size value and year are mostly non-linear. To control these non-linear relationships, we added quadratic and cubic trends for the year to the model. Multiple regression models mostly replicate results of the analysis of the bivariate relationship between effect size values and studies' characteristics. There are only a few differences that could arise both due to the confound of the influence of different predictors and because, in some cases, the sample size in the bivariate analysis and regression analysis was different due to the missing values in the variables being analyzed. The results of the multivariate analysis of the correlational effect sizes (Table 4) showed that in the entire sample and in two of the seven thematic

categories (Self and Gender differences) the publication year was related to the value of the effect size and this relationship is non-linear. Multivariate analysis of Cohen's *d* effect sizes (Table 5) also demonstrated the non-linear relationships between the effect size value and year in three of the six thematic categories (Interpersonal relationships, Self, and Social cognition). The observed patterns are very different. Since the patterns for each thematic subgroup are clearly visible in the scatterplots (Figures 4 and 5), we decided not to try to model them in regression analysis. We use the linear, quadratic, and cubic trends for a year only as a control variable and do not meaningfully discuss the values of the coefficients for these predictors.
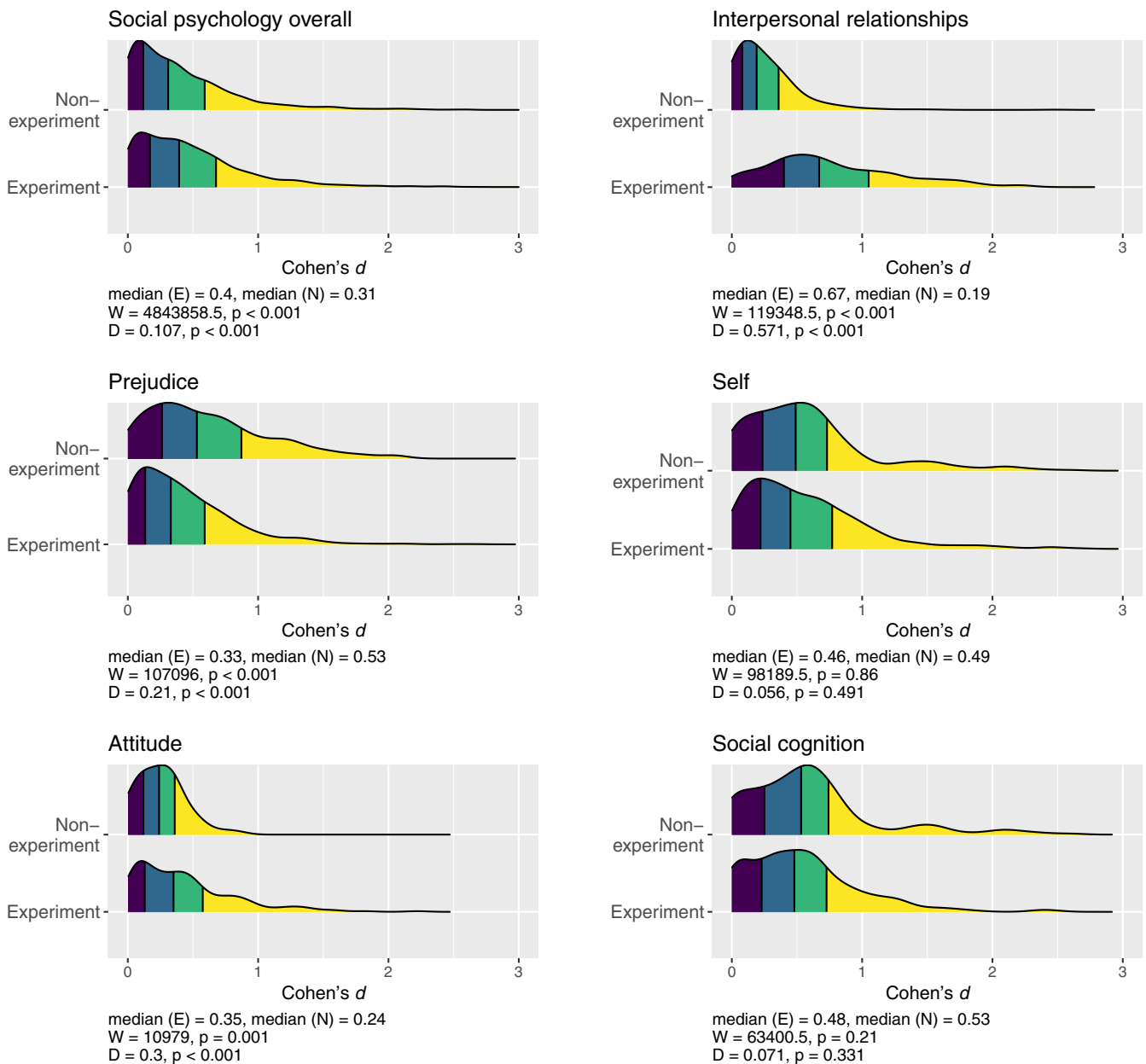

FIGURE 10 The distribution of experimental and non-experimental of Cohen's *d*s (Cohen's *d*s higher than 3.0 are not shown for easier interpretation, colors show quartiles, the x-axis is limited by value 3 for better visualization, *W*—Wilcoxon rank sum test with continuity correction, *D*—Two-sample Kolmogorov-Smirnov test, *p*—*p*-value)

TABLE 4 Relationship between the value of the correlational effect size and year, sample size, and unpublished status

	Social psychology overall	Interpersonal relationships	Prejudice	Self	Attitude	Social cognition	Gender differences	Groups
(Intercept)	5,854.82* [516.84, 11,505.18]	-28,393.62 [-88,095.50, 24,069.69]	7,437.62 [-2,458.35, 20,318.23]	74,584.17* [27,439.15, 117,186.85]	8,082.59 [-5,168.43, 20,775.66]	-6,789.04 [-85,189.80, 61,523.23]	-38,298.76* [-54,255.93, -22,837.64]	2,391.82 [-26,279.57, 31,238.10]
Year	-8.88* [-17.41, -0.81]	42.98 [-35.95, 132.70]	-11.30 [-30.75, 3.65]	-112.20* [-176.33, -41.27]	-12.43 [-31.67, 7.67]	9.40 [-93.72, 127.42]	58.50* [35.08, 82.66]	-3.36 [-46.96, 39.89]
Year (quadratic)	0.00* [0.00, 0.01]	-0.02 [-0.07, 0.02]	0.01 [-0.00, 0.02]	0.06* [0.02, 0.09]	0.01 [-0.00, 0.02]	-0.00 [-0.06, 0.05]	-0.03* [-0.04, -0.02]	0.00 [-0.02, 0.02]
Year (cubic)	-0.00* [-0.00, -0.00]	0.00 [-0.00, 0.00]	-0.00 [-0.00, 0.00]	-0.00* [-0.00, -0.00]	-0.00 [-0.00, 0.00]	0.00 [-0.00, 0.00]	0.00* [0.00, 0.00]	-0.00 [-0.00, 0.00]
Sample size (log)	-0.02* [-0.02, -0.01]	0.01* [0.00, 0.02]	-0.01* [-0.02, -0.01]	-0.02* [-0.04, -0.01]	-0.00 [-0.01, 0.01]	-0.08* [-0.10, -0.07]	-0.04* [-0.06, -0.02]	0.00 [-0.01, 0.01]
Unpublished	0.03* [0.02, 0.04]	0.04* [0.02, 0.06]	-0.02* [-0.03, -0.00]	-0.04 [-0.09, 0.01]	-0.04* [-0.06, -0.02]	-0.01 [-0.08, 0.06]	0.10* [0.06, 0.15]	0.10* [0.04, 0.16]
Observations	9,617	1,652	2,131	1,022	2,349	616	585	938
R^2	0.019	0.034	0.029	0.033	0.021	0.189	0.379	0.039

The bounds of 95% bootstrap confidence intervals are shown in square brackets (10,000 bootstrap samples). Unpublished is a dummy variable (1—if effect size is unpublished; 0—if effect size is published).

*95% confidence interval does not include zero.

TABLE 5 Relationship between the value of the Cohen's d and year, sample size, unpublished status, and experimental design

	Social psychology overall	Interpersonal relationships	Prejudice	Self	Attitude	Social cognition	Gender differences
(Intercept)	16,716.73 [−11,114.10, 44,262.95]	−172,024.52* [−330,212.92, −58,043.46]	16,575.85 [−17,006.13, 45,509.33]	302,474.82* [49,058.89, 523,397.80]	75,681.56 [−60,637.22, 201,045.13]	79,513.53* [4,800.57, 130,534.74]	−30,877.92 [−81,362.16, 26,512.17]
Year	−25.31 [−66.76, 16.54]	258.24* [86.97, 495.96]	−25.02 [−68.71, 25.65]	−454.74* [−786.73, −74.17]	−113.90 [−302.64, 91.62]	−119.92* [−197.37, −8.33]	46.36 [−40.12, 122.33]
Year (quadratic)	0.01 [−0.01, 0.03]	−0.13* [−0.25, −0.04]	0.01 [−0.01, 0.03]	0.23* [0.04, 0.39]	0.06 [−0.05, 0.15]	0.06* [0.00, 0.10]	−0.02 [−0.06, 0.02]
Year (cubic)	−0.00 [−0.00, 0.00]	0.00* [0.00, 0.00]	−0.00 [−0.00, 0.00]	−0.00* [−0.00, −0.00]	−0.00 [−0.00, 0.00]	−0.00* [−0.00, −0.00]	0.00 [−0.00, 0.00]
Sample size (log)	−0.10* [−0.12, −0.09]	−0.07* [−0.10, −0.04]	−0.06* [−0.08, −0.04]	−0.12* [−0.17, −0.07]	−0.13* [−0.15, −0.10]	−0.12* [−0.18, −0.07]	−0.07* [−0.09, −0.05]
Unpublished	−0.10* [−0.13, −0.06]	0.07* [0.01, 0.20]	−0.06* [−0.12, −0.01]	–	−0.25* [−0.36, −0.13]	−0.24* [−0.32, −0.15]	0.02 [−0.03, 0.12]
Experiment	0.01 [−0.02, 0.04]	0.11 [−0.03, 0.28]	−0.21* [−0.27, −0.16]	−0.06 [−0.17, 0.05]	0.13* [0.05, 0.23]	−0.06 [−0.15, 0.03]	–
Observations	4,876	702	1,299	589	334	729	1,220
R²	0.070	0.100	0.088	0.052	0.247	0.080	0.073

*95% confidence interval does not include zero. The bounds of 95% bootstrap confidence intervals are shown in square brackets (10,000 bootstrap samples). Unpublished is a dummy variable (1—where the effect size is unpublished; 0—where the effect size is published; not included in regression for the Self thematic subgroup because of the small number of unpublished effect sizes [n = 6]). Experiment is a dummy variable (1—where the effect size is from an experimental study; 0—where the effect size is from a non-experimental study; not included in regression for the Gender differences thematic subgroup because of the small number of experimental effect sizes [n = 38]).

In the entire samples and in most of the thematic subgroups, there was a negative relationship between the effect size value and the sample size (see Tables 4 and 5). The larger the sample in the study, the lower the effect size that was obtained. Only two regressions out of 15 did not show a relationship between effect size and sample size, and in one case the relationship was positive.

Unpublished status was significantly related to the value of the correlational effect size in the entire sample and in five of the seven thematic subgroups (Table 4). However, this relationship was negative (i.e., the unpublished effect sizes were on average lower than the published effect sizes) in only two subgroups (Prejudice and Attitude). This relationship can be interpreted as a sign of the existence of potential publication bias. However, in three other subgroups (Interpersonal relationships, Gender differences, and Groups) the unpublished effect sizes were, in contrast, higher than in the published ones, which is the opposite of what would be expected if publication bias was present. In the analysis of Cohen's *d* effect sizes (see Table 5) a sign of the existence of publication bias was observed both in the entire sample and in three thematic subgroups (Prejudice, Attitude, and Social cognition). In studies on these topics, unpublished effects sizes were lower than for published ones. However, in the Interpersonal relationships thematic subgroup, the unpublished effect sizes were again higher than in the published ones. In the Gender differences thematic subgroups, the status of the effect size did not relate to its value.

The effect sizes obtained in the experimental studies were different from those obtained in the non-experimental ones in two of the five thematic subgroups. However, the direction of this relationship is mixed: the effect sizes from experimental studies were higher in the Attitude subgroup, but lower in the Prejudice subgroup compared to the effect sizes from non-experimental studies (Table 5). These differences are reflected in the 25th, 50th, and 75th percentiles (see Figure 10); thus, the thresholds of the effect size interpretation were also different.

6 | DISCUSSION

The results of this study suggest that the interpretation guidelines for empirically derived effect sizes are different from the commonly cited and widely used guidelines suggested by Cohen (1988, 1992). Cohen's guidelines tend to overestimate effect sizes, especially medium and large effect sizes in social psychology. On the basis of an empirically derived effect size distribution, it is recommended that the correlation coefficients of 0.12, 0.24, and 0.41 and Cohen's *d*s of 0.15, 0.36, and 0.65 should be interpreted as small, medium, and large effects for studies in social psychology. The differences do not seem very large (see Table 2), but they may create large difference in the sample sizes required to achieve appropriate statistical power. It is well-known that in psychology in general, most studies are underpowered (Maxwell, 2004) and the findings by Fraley and Vazire (2014) confirm that in journals on social-personality psychology, the majority of empirical studies published have a low statistical power. Empirically derived effect size distribution can help not only for effect size interpretation but also for sample size planning. If no other

information is available (e.g., there is no relevant previous research), a researcher might assume that the effect size will be similar to what studies in the relevant sub-discipline of psychology typically find (Giner-Sorolla et al., 2019) and can use the median effect size for a thematic subfield (Table 3) for a priori power analysis. If a thematic subgroup is not covered in this study, a researcher can use the median effect size for social psychological studies in general (median for correlations = 0.24, median for Cohen's *d*s = 0.36).

The results obtained in this study make it possible to better understand the research features and conditions on which the effect size may depend. In most of the thematic subgroups there is a dynamic of the average effect size over time and these dynamics are different. The reasons for this can be both changes in the population effects being studied (e.g. Eagly et al., 2020) and changes in the characteristics of the studies and research practices by which these population effects are studied (Sassenberg & Ditrich, 2019). A rapid and significant increase in average correlation coefficient in the Gender differences category may be a consequence of the fact that the effect sizes obtained after 2000 were mainly drawn from one meta-analysis on cross-cultural differences in parental acceptance and the psychological adjustment of men and women. Previous research showed that studies focusing on differences in socialization between men and women may have stronger effects than studies based on the search for sex differences (Eagly & Wood, 2013). Curvilinear dynamics in the effect size value in the Social cognition subgroup may be related to the growing interest in social cognition research by the mid-1980s (Carlston & Schneid, 2014). It is likely that a large number of studies carried out on small samples resulted in an increase in the average effect size, and the subsequent growing demands on experimental procedures and interest in implicit measures (especially in the field of social cognition) led to a decrease in the effect sizes obtained. Carlston and Schneid (2014) showed that, since the 1990s, the number of articles referring to implicit, automatic processes has been continuously increasing, especially in the field of research into social cognition, attitudes, and prejudices. As the effect sizes associated with explicit measures are higher than those with implicit ones (Phipps et al., 2019), the decrease in the effect sizes in these areas seems to be expected and understandable. This result has practical importance for guidelines for effect size interpretation and sample size planning. When choosing the thresholds for interpretation and the exact values of the effect size for power analysis, one should take into account changes both in the true population effect and in methods and research practices in the relevant area of research, and choose more recent estimations for the effect.

The design of studies from which the effect size is obtained also matters, since at least in two of the five subgroups, the experimental effect sizes were on average different from the non-experimental ones. In particular, in the Attitude subcategory, the effect sizes obtained in the experimental studies were higher than in the non-experimental study, while in the Prejudice subcategory the direction of this relationship was the opposite. It is likely that these differences can also be associated with the characteristics of the relationships being studied. Experimental studies of prejudice most often

investigate methods for reducing bias or actualizing stereotypes. In this case, the effect sizes themselves may be weak because people find it difficult to give up their beliefs (e.g., confirmation bias). Moreover, in the experimental studies, the effect size can become even smaller since it largely depends on the success in controlling social desirability and the effectiveness of the manipulation of the independent variable. In the case of Attitude studies, experimental research quite often tends to actualize intentions and check their link with behavior. Such an effect might itself be strong since it is not associated with ideological beliefs and even social desirability will contribute to its increase. Thus, the specific features of research, even in thematically close areas of social psychology, can influence the distribution of effect size. These results empirically demonstrate the importance of considering the full research context in interpreting the effect size, which is recommended in the methodological literature (Stukas & Cumming, 2014). A researcher should use effect sizes from studies that use comparable designs, type of manipulation, and ways of measuring variables when choosing thresholds for interpreting effect sizes and planning sample size.

Due to the potential publication bias, the empirically derived thresholds for effect size interpretation may be overestimated. It is well known that studies with smaller effect sizes are less likely to be reported and meta-analyses often include only published results and ignore "gray" literature (unpublished data, dissertations, etc.) (Bakker et al., 2012). This increases the likelihood that the published effect sizes will be overstated. Unfortunately, we cannot conduct a full-fledged evaluation of the impact of publication bias and its correction, because existing methods do not allow this to be done for a sample that combines meaningfully different effect sizes. However, testing the relationship between effect size and its published/unpublished status and sample size can shed light on the existence of potential publication bias.

Analysis of the relationship between the effect size value and its status showed that in some thematic subgroups there were signs of the existence of potential publication bias. In six of the 15 regressions, unpublished effect sizes were on average lower than published ones. At the same time, in other thematic subgroups no difference was found between published and unpublished effect sizes. Moreover, in the three thematic subgroups, unpublished effect sizes were higher than published ones, which is the opposite of what would be expected in the case of publication bias. The most pronounced differences between the effect sizes in published and unpublished studies are in the Gender differences subcategory. This result may also be a sign of a potential publication bias but with reversed direction (large effect sizes may be published less frequently than small ones). Eagly and Wood (2013) noted that some psychologists were concerned that the findings of the magnitude and variability of male–female comparisons could be used to justify social inequities. It is possible to imagine that the researchers who have identified strong gender differences in their studies may be less willing to publish such results because they may not want their findings to be used to justify social and gender inequities. However, we must note that this potential mechanism of the publication bias

is just our speculation, which needs to be empirically investigated and tested.

Analysis of the relationship between effect size and sample size revealed a clearer pattern. In 12 regressions, studies with large samples report smaller effect sizes. One of the explanations for this correlation may be the presence of a potential publication bias and small study effect when research with larger samples more likely can provide "statistically significant" results, even for weak effects. However, the correlation between the effect size value and the sample size or the standard error (both characteristics are related to the precision of estimating the size effect) is also observed in the sample of pre-registered studies in which publication bias is absent (Open Science Collaboration, 2015; Schäfer & Schwarz, 2019). Other possible explanations for this correlation could be the use of power analysis to calculate the sample size in the studies analyzed when authors intentionally used larger sample sizes to test weaker effects, or what Schäfer and Schwarz (2019, p. 9) call "a learning effect": "in research areas with larger effects, scientists may have learned that small samples are enough while in research areas with smaller effects, they know that larger samples are needed".

This analysis and its results are, of course, not a rigorous and detailed test for the existence of publication bias. There are other possible explanations for the relationship between effect size and sample size and status (published/unpublished). However, we cannot rule out the impact of publication bias on the thresholds for interpretation of effect sizes empirically identified in this study. If the thresholds are overestimated, this means that the widely used Cohen's thresholds are even more overestimated and should not be used in studies in social psychology. Using Cohen's thresholds for sample size planning will result in many studies being underpowered.

The study has other limitations, which should be taken into account when using the empirically derived effect size distributions and the thresholds for their interpretation. First, the approach used to collect data could lead to an unbalanced coverage of topics within social psychology, because of a lack of published meta-analyses on some topics. Second, a substantial section of the published meta-analyses was not included in the analysis of the true empirical distribution of effect sizes as they did not report effects sizes from studies included in meta-analyses. However, we believe that a sufficiently large number of the collected effect sizes (12,170 correlations and 6,447 standardized mean differences) is very likely to be representative and reflects the features of the distributions of the correlation coefficient and Cohen's *d* statistic. Third, the studies may vary not only in experimental/non-experimental design but also in different characteristics (e.g., type of manipulation, type of data collection, etc.). It is likely that the effect sizes can vary not only between subcategories, but also between subgroups of studies with different characteristics. Further studies of the effect size distributions are needed to take into account other characteristics of the studies.

The analysis presented in this article is one step toward a deeper understanding and interpretation of effect sizes in social psychology. Many of the results obtained in our research raise questions about the features and conditions on which the effect size may depend,

and the interpretation of the differences obtained. Additional research may help us to better understand this important topic for social psychology.

CONFLICT OF INTEREST

The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS APPROVAL STATEMENT

The study did not involve human participants or animals. Ethical approval was not required for this secondary research study.

DATA AVAILABILITY STATEMENT

The dataset and script to perform the analyses are freely available at Open Science Framework https://osf.io/8cgfe/

ORCID

Andrey Lovakov https://orcid.org/0000-0001-8644-9236
Elena R. Agadullina https://orcid.org/0000-0002-1505-1412

REFERENCES

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Baumeister, R., & Vohs, K. (Eds.). (2007). *Encyclopedia of social psychology*. SAGE Publications. https://doi.org/10.4135/9781412956253

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Canty, A., & Ripley, B. (2017). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20.

Carlston, D., & Schneid, E. (2014). Trends in social cognition research. In D. E. Carlston (Ed.), *The oxford handbook of social cognition* (pp. 894–916). Oxford University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

Del Giudice, M. (2011). Sex differences in romantic attachment: A meta-analysis. *Personality and Social Psychology Bulletin*, *37*(2), 193–214. https://doi.org/10.1177/0146167210392789

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*(3), 301–315. https://doi.org/10.1037/amp0000494

Eagly, A. H., & Wood, W. (2013). The nature-nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, *8*(3), 340–357. https://doi.org/10.1177/1745691613484767

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Friese, M., & Frankenbach, J. (2020). p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Giner-Sorolla, R., Aberson, C. L., Bostyn, D. H., Carpenter, T., Conrique, B. G., Lewis, N. A. J., & Soderberg, C. (2019). *Power to detect what? Considerations for planning and evaluating sample size*. Retrieved from. https://osf.io/jnmya/

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79. https://doi.org/10.1037/0003-066X.58.1.78

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*(6), 895–919. https://doi.org/10.1037/0033-2909.132.6.895

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, *43*(2), 376–399. https://doi.org/10.1177/0149206314527133

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Paterson, T. A., Harms, P. D., Steel, P., & Crede, M. (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership & Organizational Studies*, *23*(1), 66–81. https://doi.org/10.1177/1548051815614321

Phipps, D. J., Hagger, M. S., & Hamilton, K. (2019). *A meta-analysis of implicit and explicit attitudes in children and adolescents*. https://doi.org/10.31234/osf.io/52jrs

Quintana, D. S. (2017). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology*, *54*(3), 344–349. https://doi.org/10.1111/psyp.12798

Revelle, W. (2017). psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.7.5. Retrieved from https://cran.r-project.org/package=psych

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rise, J., Sheeran, P., & Hukkelberg, S. (2010). The role of self-identity in the theory of planned behavior: A meta-analysis. *Journal of Applied Social Psychology*, *40*(5), 1085–1105. https://doi.org/10.1111/j.1559-1816.2010.00611.x

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, https://doi.org/10.3389/fpsyg.2019.00813

Stukas, A. A., & Cumming, G. (2014). Interpreting effect sizes: Toward a quantitative cumulative social psychology. *European Journal of Social Psychology*, *44*(7), 711–722. https://doi.org/10.1002/ejsp.2019

Sutton, A. J. (2006). Chapter 10: Evidence concerning the consequences of publication and related biases. In H. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments*. (pp. 175–192). John Wiley & Sons, Ltd.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.