This is the accepted version of

Hanel, P. H. P., & Mehler, D. M. A. (accepted). Beyond Reporting Statistical Significance: Identifying Informative Effect Sizes to Improve Scientific Communication. *Public Understanding of Science*.

Beyond Reporting Statistical Significance: Identifying Informative Effect Sizes to Improve Scientific Communication

Paul H. P. Hanel[1], David M. A. Mehler[2,3]

[1] Department of Psychology, University of Bath, Bath, United Kingdom

[2]CUBRIC, School of Psychology, Cardiff University, Cardiff, United Kingdom

[3]Department of Psychiatry, University of Münster, Münster, Germany

Author note. Please address correspondence to Paul Hanel, Department of Psychology, University of Bath, Claverton Down, BA2 7AY Bath, European Union, p.hanel@bath.ac.uk. The author(s) declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

Author contribution. PH generated the idea for the studies. PH and DM jointly designed the studies. PH collected the data. PH and DM analyzed the data and wrote the manuscript. Both authors approved the final submitted version of the manuscript.

Acknowledgements. We want to thank Daniël Lakens, Johannes Algermissen, Bruna Nascimento, and Chris Crandall for valuable comments. DM was supported by a PhD

studentship from Health and Care Research Wales (HS/14/20). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author's biographies. Paul Hanel is interested in similarities between groups of people, human values, as well as in social, personality, political, and cross-cultural psychology.

David Mehler holds a PhD in translational neuroimaging from Cardiff University Brain Research Imaging Centre (CUBRIC) and is currently finishing his medical studies at Muenster University, Germany. His research interests include non-invasive rehabilitation techniques such as neurofeedback training, machine learning, open science methods, and statistics.

Abstract

Transparent communication of research is key to foster understanding within and beyond the scientific community. An increased focus on reporting effect sizes in addition of *p*-value based significance statements or Bayes Factors may improve scientific communication with the general public. Across three studies ($N = 652$), we compared subjective informativeness ratings for five effect sizes, Bayes Factor, and commonly used significance statements. Results showed that Cohen's U3 was rated as most informative. For example, 440 participants (69%) found U3 more informative than Cohen's *d* while 95 (15%) found *d* more informative than U3, with 99 participants (16%) finding both effect sizes equally informative. This effect was not moderated by level of education. We therefore suggest that in general Cohen's U3 is used when scientific findings are communicated. However, the choice of the effect size may vary depending on what a researcher wants to highlight (e.g., differences or similarities).

Keywords: statistical communication, scientific communication, effect size, statistical significance, Cohen's U3, Cohen's *d*

Beyond Reporting Statistical Significance: Identifying Informative Effect Sizes to Improve Scientific Communication

Recently, social and medical sciences have begun to increasingly value open science practices including more transparent documentation of research, open data, and pre-registration (e.g., Allen & Mehler, 2018; Miguel et al., 2014; Morey et al., 2016). While open science practices allow public access to research material, transparency in communicating statistical findings seem equally important: clear communication allows readers a better evaluation of scientific findings in a more educated way. This is especially true when research is funded by the general public and in times when the public is increasingly engaging in scientific projects (e.g., "citizen science", Bonney et al., 2014). Indeed, many funding bodies also require that the findings are disseminated in an open and transparent fashion (National Institutes of Health, 2017; Research Councils UK, n. d.). However, clarity and readability of scientific texts have decreased in the last decades because of an increase in scientific jargon, impacting the accessibilities of research findings (Plavén-Sigray, Matheson, Schiffler, & Thompson, 2017). In the present article, we investigate how scientific findings can be reported in a more transparent and informative way.

Currently, scientific findings are predominantly presented as a variety of "there is a (statistically significant) difference between X and Y" or "X is associated with (or caused by) Y". However, quantifying, for example, the difference between two groups or relations between variables by the means of effect sizes is more informative because it adds a more exact "numerical statement of facts" (Bowley, 1915, p. 1). Empirical effect sizes quantify the estimated strength or magnitude of an effect and hence provide essential, complementary information that may inform decisions. While some statistical analyses such as correlations are almost exclusively reported in terms of effect sizes (i.e., a correlation coefficient that informs

about the strength of a relation between two variables), other analyses such as group difference tests often lack reporting of effect sizes, despite it has repeatedly been advocated across scientific fields (Nakagawa & Cuthill, 2007; Sullivan & Feinn, 2012; Thompson, 2002). For example, effect sizes allow to assess particular questions about group differences and are hence particular suitable to inform evidence-based decisions. For clinical work in particular, (non-standardized) effect sizes can inform about treatment success: Whereas a treatment effects may be statistically significant between the treatment and control group, overall clinical benefit on the outcome may be negligible (Sedgwick, 2014). These examples illustrate that there are statistical and scientific reasons to report effect sizes. Moreover, we argue that effect sizes can also increase the transparency of scientific communication – in particular, but not exclusively with the general public and researchers from other fields – because they can add information about the strength of an effect for an observed phenomenon and hence improve the evaluation of reported findings. However, while various effect size measures exist and their statistical properties have been well investigated (e.g., Cohen, 1988; Cumming, 2014; Ruscio, 2008), it is unclear which are perceived as most informative. With regards to public science communication, perceived informativeness of effect sizes among the general public remains to be empirically investigated.

We propose that science communication can be improved by reporting effect sizes that are perceived as informative. It has been criticized that the common practice in science communication of only reporting findings as (statistically) significant fails to inform readers about the magnitude of an effect (Nakagawa & Cuthill, 2007; Sullivan & Feinn, 2012; Thompson, 2002). More specifically, mere significance statements seem often not very informative and not easy to interpret. For instance, the general public as well as the majority of members of the academic community struggle to define and interpret (statistical) significance statements (Haller & Krauss, 2002; Hoekstra, Morey, Rouder, & Wagenmakers, 2014;

Tromovitch, 2015). In fact, one study found that only a marginal fraction of a representative sample provided a correct definition of the term "significance" (Tromovitch, 2015). Further, another study found that only 3% of researchers correctly rejected six false statements about confidence intervals (Hoekstra et al., 2014). Thus, science communication that merely focuses on statements of statistical inference likely suffers from common misconceptions around p-values and related estimates.

Furthermore, inaccurate or careless reporting from researchers and universities can also lead to statistical miscommunication. For example, a substantial part of newspaper articles was found to incorrectly report or exaggerate findings in health-related sciences. The origin of this miscommunication were often based on exaggerations of statistical findings in university press releases (Cooper, Lee, Goldacre, & Sanders, 2012; Sumner et al., 2014). Moreover, the way scientific findings are presented can lead to biases. For example, when psychological findings are presented alongside with extraneous neuroscientific information, the findings appear to be more convincing (Baker, Ware, Schweitzer, & Risko, 2017; Im, Varma, & Varma, 2017). Hence, these examples illustrate that science communication can be distorted by potential conflicts of interest including the promotion of an institution or one's own research. Statistical significance statements can lend themselves to be misused in this way because they do not provide a measure of magnitude about an effect and thus prevent a direct comparison to other evidence.

Communicating scientific findings clearly and transparently can prevent misconceptions and thus detrimental consequences across a range of behavior. For instance, patients may reject treatment because of false beliefs (Marshall, Wolfe, & McKevitt, 2012; Nicoll et al., 1993), misperceptions of climate change have been linked to failed behavioral adjustments in reducing carbon emissions (Bain, Hornsey, Bongiorno, & Jeffries, 2012; Spence, Poortinga, Butler, & Pidgeon, 2011). Taken together, risk perceptions of the general public are at particular risk to be

biased in consequence of misleading science communication. That is, the likelihood of incidences are often misestimated which can lead to inadequate decision making and behavior (Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978; Thomson, Önkal, Avcioğlu, & Goodwin, 2004). As a response to those misunderstandings and misperceptions several guidelines have been developed (Brown University Science Center, 2014; Science Media Center, 2012).[1]

These examples illustrate why clear and transparent science communication is at the heart of good science communication. We believe that one important step forward in this direction is communicating results by the means of effect sizes (cf. also Lakens, 2013), which would be beneficial for the communication of scientific findings in press releases, newspapers, and scientific abstract. We therefore suggest that easy-to-understand effect sizes should be reported in addition to a mere statement that differences have been found. We expect that this will help improving the understanding between members of the scientific community as well as the dialog with members of the general public. Transparency and openness is associated with increased trust (cf. Norman, Avolio, & Luthans, 2010; Schindler & Thomas, 1993), which we deem particularly important in times of "fake news" and as skepticism framed science negligence that can put public approval and funding at risk (Nurse, 2006). In three studies, we explore which effect size is rated as more informative, also compared to a significance condition and the Bayes Factor, using within-subject designs.

[1] We believe that adding easy to understand effect sizes increases transparency and offers less ground for a vague and biased interpretation of the findings. However, we do not believe that this will eliminate biases in the presence of ideological polarizations, because misperceptions are often driven by powerful ideological believes which "override" information and education (Kahan, Peters, Dawson, & Slovic, 2013).

The Present Research

In the present research, we compared the perceived informativeness of various ways how scientific findings can be communicated. Specifically, we compared several often-used statistics and effect sizes as well as some less frequently used ones. These were embedded in a text extract of results from a fictitious study and presented to participants. We did not have any *a priori* hypotheses which statements would be perceived as more or less informative and hence employed an explorative design. To investigate whether participants' ratings varied as a function of educational level (Studies 1-3) or statistical experience and knowledge (Study 1), we explored the data differences between both low and high educational levels and low and high statistical experience of participants. We define informative as expressing the magnitude of an effect in a comprehensible fashion.

In Study 1, we used the medium effect size within social psychology of $r = .21$ (Richard, Bond Jr., & Stokes-Zoota, 2003) and transformed it into five other effect sizes: Cohen's *d*, Cohen's U3, probability of superiority (Ruscio, 2008) which is a generalization of the common language effect size (McGraw & Wong, 1992), the overlapping coefficient (Inman & Bradley, 1989), and partial eta square. The R codes for computing the first four effect sizes can be found in the Appendix. All effect sizes were selected based on how popular we perceived them to be in the psychological literature (Cohen's *d* and eta square), whether we judged them as appropriate to express between-subject differences, and whether we were familiar with them. Two effect sizes express the difference (or overlap) in terms of variation (Cohen's *d* and partial eta square), whereas the remaining three expressed the difference in percentages. Moreover, we added the definition of the Bayes Factor with a fictitious value of 20 and, for obtaining a baseline rate, the default way of presenting scientific findings in terms of significance statements (see Table 1 for illustrations). We focused on two independent groups, because this research design is very

common, and most effect sizes have been developed to express the between-group differences. In other words, all such between-group comparisons can be expressed in either of those ways (assuming the assumptions such as normality and variance homogeneity have been met).

Study 2 replicates Study 1 in a different country and also tests whether participants understand research findings that are presented as "different" in a similar and accurate way. That is, whether participants are able to estimate the effect size by only reading that two groups are different from each other; heterogeneous responses would highlight the need to report informative and clear effect sizes. Study 3 replicates Studies 1 and 2 using abstract examples.

We have not conducted an a priori power analysis because the aim was to get first estimates of the informativeness of effect sizes. After data collection and analysis, we conducted a sensitivity analysis to demonstrate that our sample sizes are more than sufficient to detect small effects using G*Power (version 3.1.9.2; Faul, Erdfelder, Buchner, & Lang, 2009) for a within-subject comparison of the informativeness of two effect sizes. Assuming a power of .80 and an alpha of .05 (two-tailed), a sample size of 297 (as used in Study 1) would be enough to detect an effect of $dz = 0.16$, a sample size of 149 (Study 2) an effect size of $dz = 0.23$, and a sample size of 206 (Study 3) an effect size of $dz = 0.20$.[2]

Items for all studies were part of a larger survey that was irrelevant to the present research question. We report how we determined our sample size, all data exclusions, and all manipulations. The data and code to reproduce statistical analyses have been made publicly available on the Open Science Framework: https://osf.io/9vqyn/

[2] Taking the 21 comparisons into account (Sidak-correction, see below), the alpha-level decreases to .00244. With this alpha-level, the effect sizes for the three studies remain in the small-to-medium range: dzs = .23, .32, .27.

Study 1

Method

Participants. Data were collected online on Prolific academic in November 2016 from 300 participants, who reported to live in the United Kingdom. Three participants were excluded from data analysis because they failed both included attention-check items. The mean age of the remaining 297 participants was 39.81 years (*SD* = 12.59, *range* = 18 - 68), with 58.59 percent of the participants being women. One-hundred thirty-one participants had at least graduated from a university, while the remaining 166 had a lower education level. One-hundred ninety-one participants had no statistical training, 93 a bit, and 13 a lot.

Material and Procedure. To measure the informativeness of effect sizes, we created a fictitious and slightly abstract example of a between-subject design. We first explained to the participants that an effect was found in a study which can be expressed in different ways. Participants were asked how informative they found each possibility of expressing the same effect. Hence, our operational definition of informativeness concerns the comprehensibility of different statements. Participants were then told that "a researcher has compared women and men with regard to an important personality characteristic and needs your help to find the clearest and most informative way to report her findings. For this task, it is irrelevant whether women or men score higher." Participants were asked to first read all randomly presented possibilities before rating how informative they find each of them (see Table 1 for all the possibilities). Responses were given on a 7-point scale ranging from 1 (Extremely uninformative) to 7 (Extremely informative). Thus, the informativeness of each item was measured.

Education was measured using a 7-point scale ranging from 1 (Illiterate) to 7 (Profession or honours). Statistical familiarity was measured with 8-items by asking participants how familiar they are with various statistical tests and effect sizes, including two-sample t-tests,

correlation, Bayesian statistics, and multiple regression. Familiarity responses were given on a 5-point scale ranging from 1 (Not familiar at all) to 5 (Extremely familiar). We computed the mean across all 8-items to obtain an overall statistical familiarity score. The overall familiarity was low (M = 1.63, SD = 0.70). Additionally, we asked whether participants "had statistical training (e.g., at school, university, or during job training)?" Responses were given on a 3-point scale: 1 (no), 2 (a little), and 3 (a lot). Statistical familiarity correlated with statistical training, $r(295)$ = .53, $p < .001$.

Data analysis

Data were analyzed with a series of two-way mixed ANOVAs to test for interactions with educational level and statistical familiarity, and a repeated measures ANOVA. In a next step, we performed a series of equivalence tests (Lakens, 2017) to test which way of phrasing scientific findings are practically equivalent, i.e. whether an effect of a smallest effect size of interest can be rejected. As the smallest effect size of interest (SESOI), we used Cohen's d = 0.30. In other words, if the upper or lower tail of a 90%-confidence interval of a paired Cohen's d does not surpass |0.30|, we consider two statements as equivalent, because the standardized mean difference would then be too small to be relevant. We chose a Cohen's d of 0.30 because we wanted to be able to reject small/marginal effects which we judged to be less interesting in this exploratory study. We use the term small/marginal effect in a relative sense: 25% of the effect sizes reported in psychological literature are smaller than d = 0.30 (25th to 75th percentile: 0.29-0.96; Figure 3a in Szucs & Ioannidis, 2017).

In other words, if a Cohen's d of 0.30 was found for a difference in informativeness ratings between two statements, it would mean that only 61% of the people would rate one effect size to be more important than the other, whereas 39% would still rate the other one to be more important. Although we acknowledge that small effect sizes should not be neglected, in the

present exploratory work we were more interested in larger effects. The assumptions of the two-way mixed model and repeated measures ANOVAs (e.g., sphericity) were met in all studies.

Results

Moderation analyses. First, we tested whether educational level, statistical familiarity, and statistical training would interact with how informative the various response options are perceived. Level of education and statistical training did not result in a significant interaction in a two-way mixed ANOVA with education or statistical training as between-subject factors and the seven statements as within-subject factor (Fs < 1.60, ps > .14). However, statistical familiarity did interact with the within-subject-factor informativeness, $F(6, 1716) = 3.80$, $p = .001$, $\eta_p^2 = .01$. As expected, the within-subject factor statements was also significant, $F(6, 1716) = 56.43$, $p < .001$, $\eta_p^2 = .16$. Cohen's d and the Bayes Factor were judged to be more informative by people with more statistical experience, but the informativeness-order of the seven items still remained the same for both levels of statistical experience. Hence, we collapsed across all groups and focus on the overall statistics.[3]

Informativeness of statements. Table 1 shows all statements ranked by their informativeness, including pairwise comparisons (Sidak corrected). Of interest, the default statement was rated as most informative, although it did not differ significantly from Cohen's U3 and the probability of superiority. The two least informative items were the Bayes Factor and Cohen's *d*.

Table 1

Descriptive and inferential statistics, alongside brief definitions of the effect sizes

[3] We also performed a series of equivalence tests, comparing pairwise the participants scoring higher on each of the three moderators with those scoring lower on them, while correcting for multiple comparison (21 comparisons, Sidak correction). None of the comparisons was equivalent.

Statistic/Effect size	Definition	*M (SD)*
Default	The difference between men and women was statistically significant.[a]	4.71 (1.62)
Cohen's U3	67% of the members of one group (i.e., either women or men) score higher than the mean of the other group (i.e., either women or men).[b]	4.50 (1.52)
Probability of Superiority	There is a 62% chance that a person selected at random from one group (i.e., either men or women) will have a higher score than a person selected at random from the other group.[c]	4.39 (1.55)
Overlapping coefficient	The overlap of the responses given by men and women was 83%. [a,d]	4.18 (1.65)
Partial eta square	The proportion of variance explained by group membership is 4%. This means that whether a person is male or female explains 4% of the individual differences on the personality measure. [a,b,c,d,e]	3.76 (1.66)
Bayes Factor	The Bayes factor is 20, indicating that the data are at least than 20 times more likely under the assumption that men and women are different than under the assumption that they are equal. [a,b,c,d,e]	3.26 (1.77)
Cohen's *d*	The difference between men and women was Cohen's d = .43, with Cohen's d being the difference in the two groups' means divided by the average of their standard deviations. [a,b,c,d,e]	3.03 (1.84)

Note. Same superscript indicates significant differences between two statements at $p < .05$ (Sidak corrected). For example, all statements with the superscript *a* are significantly different from each other at $p < .05$.

Effect sizes and equivalence tests. Table 2 shows pairwise comparisons of the informativeness of all statements. As effect sizes for the pairwise comparisons we report Cohen's U3s and Cohen's *d*s, including Sidak-corrected confidence intervals. We also report Cohen's *d* for the respective effects because of its convenience in computing equivalence tests (Lakens, 2017) and because of its widespread use in the scientific literature. We found effect sizes that ranged from negligible to medium-to-large ($0.07 < d < 0.71$). For example, for the comparison between the default way of reporting findings as "statistically significant" and Cohen's U3 we found Cohen's $dz = .10$ (99.5%-CI [-.13, .33], Cohen's U3 = 54). Next, we tested whether the three statements that were rated as most informative can also be considered as

statistically equivalent. We therefore computed the Sidak corrected confidence intervals that were corrected for three comparisons. The obtained confidence interval for the comparison of the top two rated statements, the default statement and Cohen's U3, was statistically equivalent, 96.5%-CI [-.07, .28]. Here the numbers in the squared brackets express the lower and upper equivalence bounds in units of Cohen's *d*. Also, Cohen's U3 and the probability of superiority were equivalent, 96.5%-CI [-.10, .25]. The confidence interval for the default statement and probability of superiority, however, remained undetermined because its CI extended slightly beyond the upper equivalence bound of 0.30, 96.5%-CI [-.01, .34].

Table 2

Cohen's U3s and Cohen's ds of pairwise comparisons with confidence intervals

	1	2	3	4	5	6
1. Default						
2. Cohen's U3	54, .10 [-.13, .33]					
3. Probability of Superiority	57, .17 [-.07, .40]	53, .07 [-.16, .31]				
4. Overlapping coefficient	60, .24 [.01, .48]	57, .17 [-.07, .40]	54, .10 [-.13. .33]			
5. Partial eta square	66, .41 [.18, .64]	66, .43 [.19, .66]	63, .33 [.09, .56]	59, .22 [-.02, .45]		
6. Bayes Factor	73, .62 [.38, .86]	73, .61 [.37, .85]	71, .54 [.30, .78]	67, .43 [.20, .67]	60, .25 [.01, .48]	
7. Cohen's d	76, .71 [.47, .95]	76, .70 [.46, .94]	73, .62 [.38, .86]	71, .56 [.32, .79]	64, .35 [.12, .59]	56, .14 [-.09, .38]

Note. The first number in each cell is Cohen's U3 in percent, followed by Cohen's d. Numbers in brackets refer to lower and upper limit of 99.5%-confidence interval (Sidak corrected Cis of Cohen's *d* for 21 comparisons).

Study 2

The results of Study 1 revealed that some ways of expressing scientific findings are perceived as more informative than others (see Table 1). For example, Cohen's U3 was judged

to be more informative than Cohen's *d*. However, it was surprising that the default version was rated as more informative than many effect sizes. Given that the default version is likely to be more familiar to most participants, this finding can be due to a variant of the mere exposure effect (Montoya, Horton, Vevea, Citkowicz, & Lauber, 2017; Zajonc, 1968). We therefore tested in Study 2 whether the default version would still be rated as more informative when aligned with an effect size. To test whether our findings would conceptually replicate in another English-speaking country, this sample was limited to participants who reported to live in the United States of America (USA).

Additionally, we asked participants how large they estimate gender differences for anxiousness and care orientation. We were mainly interested whether the effect estimates are more homo- or more heterogeneous, irrespective of whether the effect estimate was accurate or not. A heterogeneous interpretation of effect sizes would indicate that people understand the results differently and thus emphasize the need to report informative effect sizes to ensure that people can interpret the findings similarly and do not over- or underestimate the effect size.

Method

Participants. Data were collected online in July 2017 from 156 participants who reported living in the USA. Participants were recruited via Mturk. Seven participants were excluded from data analysis because they responded to all items too fast (see below for a more detailed justification). The mean age of the remaining 149 participants was 33.70 years (*SD* = 11.56, *range* = 18 - 79), with 62 being women (41.61% with one missing value). One-hundred eight participants (72.48%) had at least graduated from a university (55.36% with a Bachelor and 18.12% with a Master or doctorate degree), while the remaining 40 (26.85%) reported to have a lower education (one missing response).

Material. Initially, participants were informed that they would be presented with findings of several studies. In the part of the survey relevant to the present study, participants responded to three items, which were presented on separate screens. The first two items aimed to estimate how homogeneous and accurately participants understood two statements from scientific studies that were extracted from meta-analyses which were presented as large studies, whereas the third item was an extension of the item with seven response options used in Study 1. The first item informed participants that a large study which "included around 40,000 people" has "found that *women are more anxious than men.*" Participants were asked to rate on a slider measure from 50 to 100 percent how they understand the finding, using Cohen's U3 as an effect size: "*Specifically, what percentage of women is more anxious than the average men?* Choosing 50% would suggest that both women and men are equally anxious (which isn't what the study has found). In contrast, choosing 100% would mean that all women are more anxious than the average men." This item is based on the meta-analysis of Feingold (1994), who has found a sex difference of around U3 = 61% (d = .29). The second item was the same as the first item, except that we used an example from the meta-analysis of Jaffee and Hyde (2000), who found that women had a stronger care orientation than men (U3 = 61%, d = .28). The wording was the same, except that we replaced anxious with care and added the definition of the care orientation of Jaffee and Hyde: "maintaining relationships, responding to the needs of others, and a responsibility not to cause hurt" (p. 703).

Item 3 was the same as in Study 1, except that each of the seven statements that contained fictitious examples started with the default version (e.g., "The difference between men and women was statistically significant. Specifically, the overlap of the responses given by men and women was 83%"). The default version was the same as in Study 1. The order of the seven

statements was randomized. Education level was measured with a one-item scale ranging from 1 (Did not attend school) to 6 (Master's degree or higher).

To improve the data quality, unrealistic fast responders were excluded. Given that the text for the first item consisted of 87 words, the text for item 2 of 113 words, and the total text for item 3 of 352 words, we felt that the responses of participants who answered within a few seconds are not interpretable. Specifically, data of 15 participants who responded in less than 15 seconds on the first, 14 participants on the second item were excluded. Further, data of 22 participants who responded in less than 30 seconds on all 7 sub-items were excluded. That is, the means across all 7 response options for the 22 participants were more similar than for the remaining participants.[4] This procedure let to the exclusion of all responses of 7 participants, and a partial exclusion of another 15 participants. The data analyses followed the same pattern as Study 1.

Results

Moderation analyses. First, we tested whether there were differences in educational level for all three items separately. However, we neither found significant differences in how large participants estimated the difference between men and women for anxiousness and care (*t*s < 1.53, *ps* > .13), nor did this response interact with the seven response options of item 3,$F(4.51, 595.02) = 0.56$, $p = .71$.[5]

Heterogeneity in interpretation of findings. Responses to item 1 and 2 suggest that the responses were quite heterogeneous. The dispersion as quantified by the standard deviation and

[4] Also, because our cut-off criteria are somewhat arbitrary, we have also uploaded the dataset containing the responses of all participants prior to any exclusion (see above for the link). The pattern of results barely changed without the exclusion. However, the answers of fast responders are not interpretable.

[5] Also, for this analysis we performed a series of equivalence tests, comparing pairwise lower with higher educated participants, while correcting for multiple comparison (9 comparisons, Sidak correction). None of the comparisons was equivalent, indicating that the sample was likely underpowered to provide evidence for statistical equivalence for a range of small effects (Cohen's $d = -0.3 – 0.3$).

interquartile range (IQR) was substantial for both items, $SD_1 = 11.11$, $IQR_1 = 60 - 78.5$, $SD_2 = 12.07$, $IQR_2 = 60 - 81.5$, indicating that responses varied largely across participants. Further, participants overestimated the gender differences. Two one-sample t-tests against the correct response of U3 = 61 revealed that participants overestimated the gender differences for anxiety, $M = 67.94$, $t(140) = 7.41$, $p < .001$, and care, $M = 72.31$, $t(140) = 11.13$, $p < .001$.

Informativeness of statements. Table 3 show that Cohen's U3 was rated as most informative, followed by the probability of superiority and the overlapping coefficient. Cohen's *d* and the default version were rated as least informative. Table 3 also shows pairwise comparisons of the informativeness of all statements. As effect sizes for the pairwise comparisons we report again Cohen's U3s and Cohen's *d*s, including Sidak-corrected confidence intervals. The effect sizes range from negligible to large ($0.01 < d < 0.80$). Figure 1 displays the distribution of the responses to all seven statements including boxplots.

Table 3

Cohen's U3s and Cohen's ds of pairwise comparisons with confidence intervals

	M (SD)	1	2	3	4	5	6
Cohen's U3	5.22 (1.35)[a]						
Probability of	4.82	61, .27					
Superiority	(1.49)[a,b]	[-.07, .62]					
Overlapping	4.40 (1.62)[a,c]	66, .41,	59, .22				
coefficient		[.06, .76]	[-.13, .56]				
Partial eta square	4.25	70, .52	61, .29	53, .08			
	(1.55)[a,b,d]	[.17, .88]	[-.06, .64]	[-.27, .43]			
Bayes Factor	3.72	76, .72	69, .50	63, .33	62, .30		
	(1.80)[a,b,c,d]	[.36, 1.07	[.15, .86]	[-.02, .67]	[-.05, .65]		
Cohen's d	3.51	79, .80	72, .59	65, .38	66, .41	55, .13	
	(1.85)[a,b,c,d]	[.44, 1.16]	[.24, .94]	[.02, .72]	[.06, .76]	[-.21, .48]	
Default	3.50	78, .76	72, .59	66, .41	62, .29	54, .09	50, .01
	(1.77)[a,b,c,d]	[.41, 1.12]	[.23, .94]	[.06, .76]	[-.05, .64]	[-.25, .44]	[-.34, .35]

Note. Same superscript indicates significant differences between two statements at $p < .05$ (Sidak corrected). First number in each cell from column 3 onwards is Cohen's U3 in percent, followed by Cohen's *d*. Numbers in brackets refer to lower and upper limit of 99.5%-confidence interval (Sidak corrected CI of Cohen's *d* for 21 comparisons).

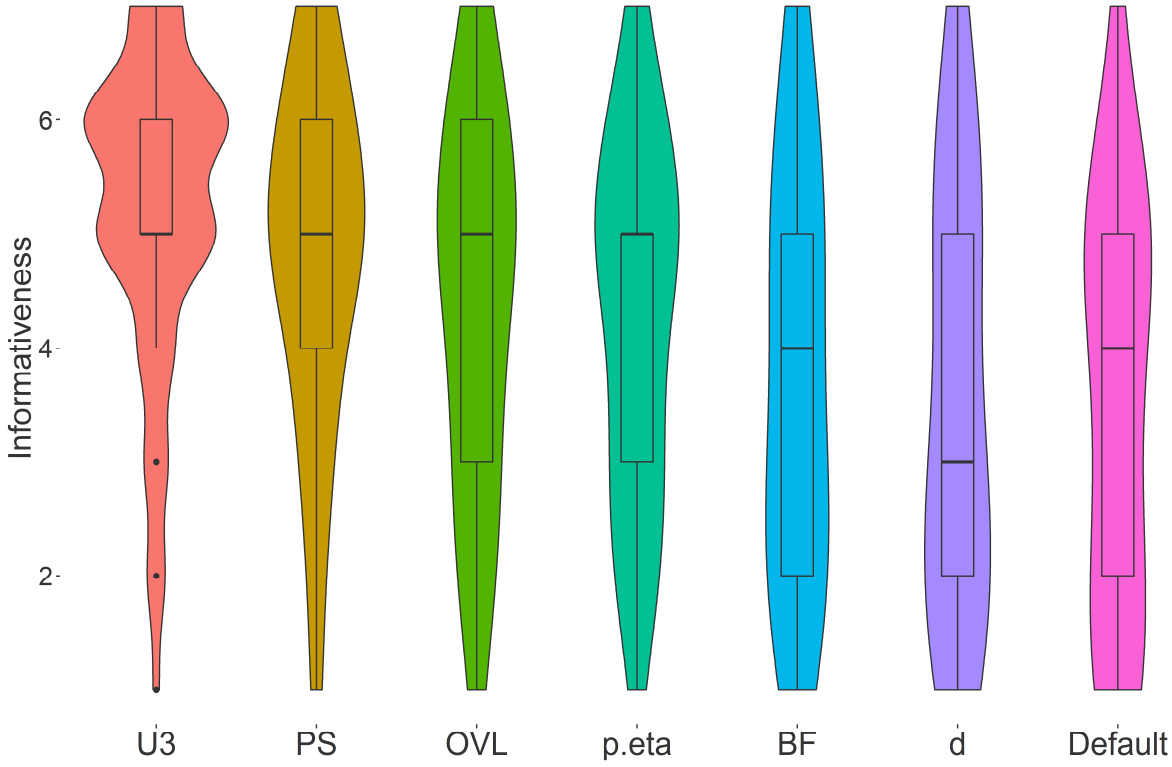

Figure 1. Violin plots with boxplots (bold line in each violin is the median).

Note. The wider a violin plot is on a specific point of the y-axis, the more people have chosen this response. For example, most participants rated the informativeness of Cohen's U3 with 5 or 6 and very few with 1, 2, or 3. The box of the boxplots shows in which range 50% of the data falls; below the box are 25% of the responses and above the box another 25%. The line in bold print represents the median. U3: Cohen's U3, PS: Probability of Superiority, OVL: Overlapping Coefficient, p.eta: partial eta square, BF: Bayes Factor, d: Cohen's *d*.

Study 3

Studies 1 and 2 used gender differences as an example. This might have caused participants to rely on their gender stereotypes and rate numerically larger effect sizes as more informative than those with smaller ones, because they were expecting larger gender differences.

Thus, we decided to replicate our findings using an abstract example where it was unlikely that we would tap into social stereotypes.

Method

Participants. Data were collected online on Prolific academic in November 2018 from 231 participants, who reported to live in the United Kingdom. Twenty-five participants were excluded from data analysis because they failed a simple attention-check item (see below). The mean age of the remaining 206 participants was 38.49 years (*SD* = 13.25, *range* = 20 - 73), with 64.08 percent of the participants being women. One-hundred seven participants had at least graduated from a university, while the remaining 97 had a lower education level. The present study was collected together with another study in which the persuasiveness of arguments either in favor or against Brexit were investigated. The median completion time was 15 minutes (911 seconds).

Material. We presented the same items in a randomized order we used in Study 2 with the only difference that we replaced women and men with Group A and Group B. For example, the instructions now stated: "Imagine, a researcher has compared two groups with regard to an important personality characteristic and needs your help to find the clearest and most informative way to report her findings. Let us call the groups Group A and Group B. Let us assume that the members of Group A score on average higher than those of Group B. Below are seven possibilities…". The item for Cohen's U3 stated: "The difference between Group A and Group B is statistically significant. 67% of the members of Group A score higher than the mean of Group B."

To test whether participants paid sufficient attention to the items we asked on the next page of the online survey: "Which of the two groups scored on average higher in the preceding example?" Twenty-five participants incorrectly said Group B and were therefore excluded.

Additionally, we asked participants whether they managed to understand the seven items. Responses were given on a 7-point scale ranging from 1 ("understood none of the seven possibilities") to 7 ("understood all seven possibilities"). We forgot to include the response option "understood two of the seven possibilities".

Results

Moderation analyses. First, we tested whether there were differences in educational level and how many items participants understood. This was done again using a mixed ANOVA with the seven statements as the within-subject factor and one of the two moderators as between-subject factor. We neither found a significant interaction for education, $F(18.66, 914.30) = 1.46$, $p = .10$[6] nor how well they understood the items in general, $F(27.60, 896.42) = 1.02$, $p = .44$. Thus, we collapsed across all groups.

Informativeness of statements. Table 4 show that Cohen's U3 was rated as most informative, followed by the probability of superiority and the default option. The Bayes Factor and Cohen's *d* were rated as least informative. Table 4 also shows pairwise comparisons of the informativeness of all statements. As effect sizes for the pairwise comparisons we report again Cohen's U3s and Cohen's *dz*s, including Sidak-corrected confidence intervals. The effect sizes range from negligible to large ($0.00 < dz < 0.80$). Overall, 150 participants rated Cohen's U3 as more informative than Cohen's d, whereas only 32 participants rated d as more informative than U3 (the remaining 24 participants rated both effect sizes as equally informative).

[6] Again, we performed a series of equivalence tests, comparing pairwise lower with higher educated participants, while correcting for multiple comparison (7 comparisons, Sidak correction). None of the comparisons was equivalent, indicating that the sample was likely underpowered to provide evidence for statistical equivalence for a range of small effects (Cohen's $d = -0.3 – 0.3$).

Table 4

Descriptive statistics and pairwise comparisons with Cohen's U3s and Cohen's ds of pairwise comparisons with confidence intervals

	M (SD)	1	2	3	4	5	6
Cohen's U3	5.50 (1.64)[a]						
Probability of Superiority	5.13 (1.65)[b]	57, .18 [-.10, .46]					
Default	5.06 (1.58)[c]	58, .21 [-.07, .49]	51, .03 [-.25, .31]				
Overlapping coefficient	3.77 (1.98)[a,b,c,d]	76, .72 [.43, 1.01]	72, .59 [.31, .88]	69, .49 [.21, .77]			
Partial eta square	3.57 (2.18)[a,b,c]	76, .69 [.41, .98]	72, .60 [.31, .88]	69, .51 [.22, .79]	54, .10 [-.18, .38]		
Bayes Factor	3.17 (2.13)[a,b,c,d]	80, .84 [.55, 1.13]	78. .77 [.48, 1.06]	75, .68 [.39, .97]	60, .25 [-.03, .53]	58, .19 [-.09, .47]	
Cohen's d	3.13 (2.20)[a,b,c,d]	80, .86 [.57, 1.15]	79, .79 [.50, 1.08]	76, .70 [.41, .99]	61, .27 [-.01, .55]	58, .19 [-.09, .47]	51, .03 [-.25, .30]

Note. Same superscript indicates significant differences between two statements at $p < .05$ (Sidak corrected). First number in each cell from column 3 onwards is Cohen's U3 in percent, followed by Cohen's *d*. Numbers in brackets refer to lower and upper limit of 99.5%-confidence interval (Sidak corrected Cis of Cohen's *d* for 21 comparisons).

General Discussion

The present studies compared the informativeness of various ways in which statistical findings can be communicated. First, we found in all studies that Cohen's U3 was rated as one of the most informative effect sizes. In Study 1 and 3, the informativeness of Cohen's U3 did not statistically differ from the effect size probability of superiority, as well as the default statement. However, given that most science communication is using the default way, that is reporting mere significance statements, informativeness ratings of the default statement can partly be explained

by an exposure effect (Montoya et al., 2017; Zajonc, 1968) or a form of status quo bias (Samuelson & Zeckhauser, 1988). An alternative explanation may be that the default statement contained less statistical jargon compared to most other effect size measures definitions that were provided to participants (see Table 1). However, we argue that this explanation is not supported by the data from informative ratings in Study 2, in which we controlled for this possible confound by presenting all effect size statements in combination with the default statement and also presented the default statement alone. Results corroborated that Cohen's U3 (in combination with the default statement) was rated as most informative. In contrast, the default statement alone was rated as least informative.

Three additional observations were made in Study 2: First, we found that also Probability of Superiority and Overlapping Coefficient were rated as informative, although significantly less informative than Cohen's U3. Second, Cohen's *d* was considered the least informative effect size in both studies. This finding may be explained by the fact that the definition of Cohen's d indeed contained more statistical jargon ("mean" and standard deviation") than the definitions of most other measures. For comparison, across all studies, results showed that 56% of the participants found Cohen's U3 more informative than the second most informative effect size, the probability of superiority, 57% found U3 more informative than d and 77% found U3 more informative than the least informative effect size, Cohen's *d*. To express it in terms of absolute frequencies – which are easily accessible to people (Hoffrage & Gigerenzer, 1998) – we found that 243 participants rated Cohen's U3 more informative than the second most informative effect size (probability of superiority), whereas 159 rated the latter as more informative. Further, 296 rated U3 as more informative than the default statement, whereas 188 rated the latter as more informative. Finally, 440 participants rated U3 as more informative than d, whereas 95 felt the other way. Ninety-nine participants rated both effect sizes as equally informative (18 participants

did not respond to either the U3 or d items, or both). Lastly, findings from Study 2 further suggested that participants interpreted the findings heterogeneously, highlighting the need for reporting informative effect sizes to avoid that findings are over- or underestimated (cf. Posavac & Sinacore, 1984).

Furthermore, summarizing research findings with Cohen's U3 may also make the findings more accessible to some individuals who dismiss scientific findings based on anecdotal evidence. In our experience, some lay-people would, for example, dismiss the meta-analytical finding that women are more anxious than men (Feingold, 1994) if they can think about one man who is more anxious than one woman. Better known as the availability heuristic in cognitive psychology and behavioral economics, immediate information that can be recalled may override other evidence that seems less accessible (Tversky & Kahneman, 1973). We suggest that reporting scientific evidence in a more transparent fashion, for instance by stating that "61% of the women are more anxious than the average man" in the current example, makes clear that gender differences refer to the group averages and provides a statistical magnitude for an effect that may be recalled more easily. However, which effect size a researcher wants to report depends on the context: while Cohen's U3 or the probability of superiority emphasize more potential differences between groups of people, the overlapping coefficient highlights similarities between groups of people. For example, if researchers compare polarized groups, highlighting similarities can improve intergroup attitudes (Hanel, Maio, & Manstead, in press).

We note that it may seem surprising that one of the most frequently used effect sizes, Cohen's *d*, was consistently considered to be least informative. However, the definition of Cohen's *d* contains the term standard deviation, a statistical term that rarely appears in everyday language. Also, partial eta square was presumably rated as less informative because it contained the statistical term "variance" with which many participants were likely unfamiliar with.

Conversely, the three effect sizes that were rated as most informative (Cohen's U3, the probability of superiority, and the overlapping coefficient) used percentage, which is one of the most frequently statistics in everyday language. Thus, when communicating scientific findings for a wider audience, commonly understood statistical concepts such as proportions are preferable. As a more general point with regards to standardized effect sizes, it should be noted that these do not require domain knowledge (in contrast to non-standardized effect sizes that operate in respective measurement units), and that they are hence in particular useful for communication with the public and researchers from other fields. Based on the current evidence, we suggest that besides statistical properties that may guide researchers in reporting certain effect sizes (Cumming, 2014), perceived informativeness by the general public may also be considered, because it can potentially increase their public outreach and impact (Bonney et al., 2014).

The current studies were based on a convenience samples and thus only provided post-hoc control for educational levels of participants and not for the exact profession. Hence, the generalizability of the current findings is limited. We note, however, that we found no evidence that our effects were moderated by educational level and statistical training. Although this is speculative and remains to be tested in future studies, it is possible that our findings would replicate within a sample of quantitative researchers. Also, future studies could test whether reports of scientific findings presented with Cohen's U3 are processed more accurately by lay people, than for instance findings that are expressed with Cohen's *d*, or without any effect size. This would provide further rationale using effect sizes in scientific communication.

Another limitation pertains the example we used. While we selected gender differences because they are accessible to everyone and concrete (Studies 1-2), and an abstract example (Study 3), we acknowledge that it is not clear whether our findings would generalize to other

contexts (e.g., mean differences between conservatives and liberals, between Christians and Muslims).

However, despite largely consistent findings across both studies regarding the order of informativeness ratings between effect sizes, we also want to highlight that even the most informative effect size (Cohen's U3) was on average rated as only "slightly informative" (a "5" on the 7-point scale ranging from 1 to 7). Future research could thus test whether embedding them in a larger context such as a newspaper article would further increase the perceived informativeness and whether more informative effect sizes result in more behavioral consequences. For example, are students more willing to try a new learning method if the findings of a study advertising this method were presented with Cohen's U3 rather than Cohen's *d* or no effect size.

Informativeness could also be enhanced through visualizations in form of statistical graphs, infographics, and graphical abstracts can likely improve science communication. However, the comparability between graphs can be limited, for example because of differences in graph type (e.g. bar plot vs a cake diagram) and 2) graphs tend to consume more space in publications. In contrast, standardized effect sizes are not affected by different measurement scales, can be converted into each other, and communicate essential information very concisely. Although empirical work on this topic is still scarce, visual aids in combination with effect sizes might help making abstract information more accessible and intuitive to a wider audience (e.g., Gardiner, Sullivan, & Grand, 2018; Lazard & Atkinson, 2015). Such visual aids in combination with effect sizes could be reported in press releases or lay summaries of journals.

Conclusion. To further improve the informativeness and transparency of scientific communication, quantifying differences through effect sizes is important. Across three studies,

we compared several effect sizes relevant for comparisons of two groups. Based on our findings, we suggest reporting Cohen's U3 along with a statistical significance statement.

References

Allen, C. P. G., & Mehler, D. M. A. (2018). Open Science challenges, benefits and tips in early career and beyond. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/3czyt

Bain, P. G., Hornsey, M. J., Bongiorno, R., & Jeffries, C. (2012). Promoting pro-environmental action in climate change deniers. *Nature Climate Change*, *2*(8), 600–603. https://doi.org/10.1038/nclimate1532

Baker, D. A., Ware, J. M., Schweitzer, N. J., & Risko, E. F. (2017). Making sense of research on the neuroimage bias. *Public Understanding of Science*, *26*(2), 251–258. https://doi.org/10.1177/0963662515604975

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. *Science*, *343*(6178), 1436–1437. https://doi.org/10.1126/science.1251554

Bowley, S. A. L. (1915). *An Elementary Manual of Statistics*. P.S. King & son, Limited.

Brown University Science Center. (2014). Quick guide to science communication. Retrieved from https://www.brown.edu/academics/science-center/sites/brown.edu.academics.science-center/files/uploads/Quick_Guide_to_Science_Communication_0.pdf

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

Cooper, B. E. J., Lee, W. E., Goldacre, B. M., & Sanders, T. A. B. (2012). The quality of the evidence for dietary advice given in UK national newspapers. *Public Understanding of Science*, *21*(6), 664–673. https://doi.org/10.1177/0963662511401782

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429–456. https://doi.org/10.1037/0033-2909.116.3.429

Gardiner, A., Sullivan, M., & Grand, A. (2018). Who are you writing for? Differences in response to blog design between scientists and nonscientists. *Science Communication*, *40*(1), 109–123. https://doi.org/10.1177/1075547017747608

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online*, *7*(12). Retrieved from http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf

Hanel, P. H. P., Maio, G. R., & Manstead, A. S. R. (in press). A new way to look at the data: Similarities between groups of people are large and important. *Journal of Personality and Social Psychology*.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine: Journal of the Association of American Medical Colleges*, *73*(5), 538–540.

Im, S., Varma, K., & Varma, S. (2017). Extending the seductive allure of neuroscience explanations effect to popular articles about educational topics. *British Journal of Educational Psychology*. https://doi.org/10.1111/bjep.12162

Inman, H. F., & Bradley, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods*, *18*(10), 3851–3874. https://doi.org/10.1080/03610928908830127

Jaffee, S., & Hyde, J. S. (2000). Gender differences in moral orientation: A meta-analysis. *Psychological Bulletin*, *126*(5), 703–726. https://doi.org/10.1037/0033-2909.126.5.703

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2013). Motivated numeracy and enlightened self -government. *Cultural Cognition Project Working Paper No. 116*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2319992

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

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 1948550617697177. https://doi.org/10.1177/1948550617697177

Lazard, A., & Atkinson, L. (2015). Putting environmental infographics center stage: The role of visuals at the elaboration likelihood model's critical point of persuasion. *Science Communication*, *37*(1), 6–33. https://doi.org/10.1177/1075547014555997

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551. https://doi.org/10.1037/0278-7393.4.6.551

Marshall, I. J., Wolfe, C. D. A., & McKevitt, C. (2012). Lay perspectives on hypertension and drug adherence: systematic review of qualitative research. *BMJ*, *345*, e3953. https://doi.org/10.1136/bmj.e3953

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361–365. https://doi.org/10.1037/0033-2909.111.2.361

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van der Laan, M. (2014). Promoting Transparency in Social Science Research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

Montoya, R. M., Horton, R. S., Vevea, J. L., Citkowicz, M., & Lauber, E. A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin*, *143*(5), 459–498. https://doi.org/10.1037/bul0000085

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … Zwaan, R. A. (2016). The peer reviewers' openness initiative. *Royal Society Open Science*. Retrieved from http://www.research.ed.ac.uk/portal/en/publications/the-peer-reviewers-openness-initiative(4c5dfd3a-fa82-4fb9-994b-bfadb09f711b).html

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, *82*(4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x

National Institutes of Health. (2017). A checklist for communicating science and health research to the public. Retrieved from https://www.nih.gov/institutes-nih/nih-office-

director/office-communications-public-liaison/clear-communication/science-health-public-trust/checklist-communicating-science-health-research-public

Nicoll, A., Laukamm-josten, U., Mwizarubi, B., Mayala, C., Mkuye, M., Nyembela, G., & Grosskurth, H. (1993). Lay health beliefs concerning HIV and AIDS—a barrier for control programmes. *AIDS Care*, *5*(2), 231–241. https://doi.org/10.1080/09540129308258604

Norman, S. M., Avolio, B. J., & Luthans, F. (2010). The impact of positivity and transparency on trust in leaders and their perceived effectiveness. *The Leadership Quarterly*, *21*(3), 350–364. https://doi.org/10.1016/j.leaqua.2010.03.002

Nurse, P. (2006). US biomedical research under siege. *Cell*, *124*(1), 9–12. https://doi.org/10.1016/j.cell.2005.12.029

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., & Thompson, W. H. (2017). Research: The readability of scientific texts is decreasing over time. *ELife*, *6*, e27725. https://doi.org/10.7554/eLife.27725

Posavac, E. J., & Sinacore, J. M. (1984). Improving the Understanding of Statistical Significance: Reporting Effect Size. *Knowledge*, *5*(4), 503–508. https://doi.org/10.1177/107554708400500404

Research Councils UK. (n. d.). Public engagement with research strategy. Retrieved from http://www.rcuk.ac.uk/documents/publications/rcukperstrategy-pdf/

Richard, F. D., Bond Jr., 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

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. https://doi.org/10.1007/BF00055564

Schindler, P. L., & Thomas, C. C. (1993). The structure of interpersonal trust in the workplace. *Psychological Reports*, *73*(2), 563–573. https://doi.org/10.2466/pr0.1993.73.2.563

Science Media Center. (2012). 10 best practice guidelines for reporting science & health stories. Retrieved from http://www.sciencemediacentre.org/wp-content/uploads/2012/09/10-best-practice-guidelines-for-science-and-health-reporting.pdf

Sedgwick, P. (2014). Clinical significance versus statistical significance. *BMJ*, *348*, g2130. https://doi.org/10.1136/bmj.g2130

Spence, A., Poortinga, W., Butler, C., & Pidgeon, N. F. (2011). Perceptions of climate change and willingness to save energy related to flood experience. *Nature Climate Change*, *1*(1), 46–49. https://doi.org/10.1038/nclimate1059

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., … Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, *349*(dec09 7), g7015–g7015. https://doi.org/10.1136/bmj.g7015

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32.

Thomson, M. E., Önkal, D., Avcioğlu, A., & Goodwin, P. (2004). Aviation risk perception: A comparison between experts and novices. *Risk Analysis*, *24*(6), 1585–1595. https://doi.org/10.1111/j.0272-4332.2004.00552.x

Tromovitch, P. (2015). The lay public's misinterpretation of the meaning of 'significant': A call for simple yet significant changes in scientific reporting. *Journal of Research Practice*, *11*(1), Article P1.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–27.

Appendix

Below is the R code for the three effect sizes, which were judged to be most informative, including Cohen's d.

```
# Cohen's d
install.packages("effsize")
library(effsize)
cohen.d(x, y, na.rm = T) # Computes Cohen's d. x and y are vectors, containing
the data. Add after "= T" ", paired = T" if appropriate.
d <- cohen.d(x, y, na.rm = T)[[3]] # extracts only Cohen's d

# Cohen's U3
u3 <- function(d){pnorm(d)} # After copy + pasting this to R, you can compute
the U3 for a given d. For example, you can get Cohen's U3 for d = 0.50 with
"U3(0.50)". Do this analogue for the other effect sizes listed below

# Common language effect size (CLES; McGraw & Wong, 1992)
cl <- function(d){pnorm(d/sqrt(2))}

# Probability of superiority A (Ruscio, 2008), which is the non-parametric
version of the CLES
ps <-
function(x,y){suppressWarnings(as.numeric(wilcox.test(x,y)[1])/(length(x)*leng
th(y)))0} # Enter "ps(x, y)" in R, with x and y being vectors

# Overlapping coefficient (OVL):
OVL <- function(d){2*pnorm((-abs(d))/2)}
```