Effectively Communicating Effect Sizes

Yea-Seul Kim
University of Washington
Seattle, Washington
yeaseul1@uw.edu

Jake M. Hofman
Microsoft Research
New York, New York
jmh@microsoft.com

Daniel G. Goldstein
Microsoft Research
New York, New York
dgg@microsoft.com

ABSTRACT

How do people form impressions of effect size when reading the results of scientific experiments? We present a series of studies about how people perceive treatment effectiveness when scientific results are summarized in various ways. We first show that a prevalent form of summarizing scientific results—presenting mean differences between conditions—can lead to significant overestimation of treatment effectiveness, and that including confidence intervals can, in some cases, exacerbate the problem. We next attempt to remedy these misperceptions by displaying information about variability in individual outcomes in different formats: explicit statements about variance, a quantitative measure of standardized effect size, and analogies that compare the treatment with more familiar effects (e.g., differences in height by age). We find that all of these formats can substantially reduce initial misperceptions, and that effect size analogies can be as helpful as more precise quantitative statements of standardized effect size.

INTRODUCTION

As the world becomes more data-driven, people are increasingly exposed to statistical information about uncertain outcomes. For instance, newspaper articles often report the results of medical studies where some people are randomly assigned to receive an experimental treatment (e.g., green tea extract supplements) while others are not, after which the health of people in the two groups is compared (e.g., by measuring changes in cholesterol levels). In summarizing such studies, it is common for authors and journalists alike to present readers with information about the average outcome in each group, often emphasizing the difference in means between groups as evidence for treatment effectiveness (e.g., the group that was assigned to take the supplements lowered their cholesterol by 0.62 mmol/L more than the control group *on average* [7]).

While mean differences provide an indication of treatment effectiveness, they also rely on domain knowledge and mask potentially important information about how outcomes vary around group averages. For instance, consider two different supplements, each of which lowers cholesterol by the same amount on average, but those assigned to take the first supplement end up with highly variable blood pressures while those who take the second all have outcomes close the improved average for the group. Most people would value the second option higher than the first, as it represents a less uncertain choice in terms of their own individual health if they were to take the supplement.

The idea of conveying information about both average treatment effects and variation around these averages is not new. In fact, it has been around for decades and initially gained traction in scientific communities with the work of the statistician Jacob Cohen [4]. Cohen introduced measures of *standardized effect size* that incorporate information about both average outcomes *and* variation in outcomes, useful for comparing effects across different domains. One such measure of standardized effect size, known as Cohen's d, simply normalizes the mean difference between groups by the (pooled) standard deviation in individual outcomes: $d = \frac{\mu_1 - \mu_2}{\sigma}$.

Despite calls from scientific communities [1, 3, 4, 10, 5] it remains rare than scientists report measures of standardized effect size in their published work, and even more unlikely that such information is relayed in popular coverage of these studies. This may in part be due to the fact that people have limited experience and familiarity with standardized effect size measures. For instance, it is unlikely that a typical newspaper reader has an intuition for what a particular value of Cohen's d (e.g., $d = 0.42$ in the green tea example above) implies about treatment effectiveness.

Cohen recognized that this might be the case among scientists and laypeople alike, and so he proposed several ways to translate his d measure into terms that might be easier for people to understand. The first, simplest, and most widely adopted is a set of qualitative categories ("small", "medium", and "large"), under which the green tea effect mentioned above would be characterized as "medium-sized". Cohen also suggested re-expressing standardized effect sizes in terms of probabilities, such as the probability of superiority (also known as common language effect size), which captures how often a randomly selected member of the treatment group scores higher (or lower, in the case of cholesterol) than a randomly selected member of the control group [9, 6]. The probability of superiority for the green tea example is approximately 62%. Finally, Cohen even offered his readers analogies that compared values of d to more familiar effects, such as a difference in height by age. In this case, the difference in cholesterol between those who took green tea supplements and those who didn't is similar to the difference in height between 13 year old and 18 year old American women [4].

These alternative ways of communicating standardized effect sizes are potentially promising, but there has been relatively little work to assess how people respond to them. A rare exception is work by Brooks et al. [2] which compares the effectiveness of communicating effect sizes using traditional measures of effect size (e.g., r and r^2) to using nontraditional measures (e.g., probability of superiority). Our work is different in that we explore several other formats using a baseline (mean difference) that is more familiar to laypeople and commonly presented in popular accounts of scientific findings. We contribute a sequence of two large-scale, pre-registered[1], randomized experiments involving 2,500 participants to investigate how to best communicate effect sizes, centered around two main research questions: 1) How effective do people think a treatment is when the treatment is summarized only in terms of its average effect? and 2) How do these initial perceptions change after people are presented with information about how individual outcomes vary around the average effect?

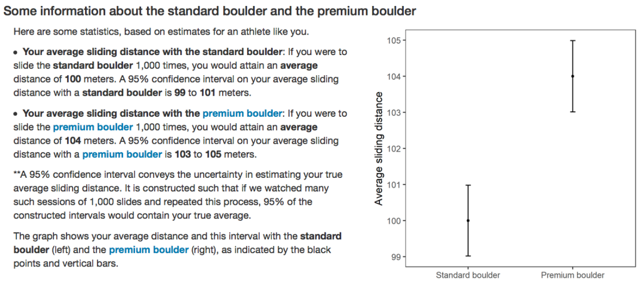
STUDY 1: ASSESSING (MIS)PERCEPTIONS

We designed our first study to evaluate how effective people perceive an uncertain treatment to be when it is phrased in terms of only mean differences between conditions, as is commonly the case in popular and scientific articles. Participants were presented with information about a treatment in one of five formats with varying levels of detail and asked how much they would be willing to pay for the treatment and to estimate its probability of superiority. The least informative format was a simple directional statement that merely indicated that the treatment led to better outcomes *on average*, without any precise statements about the size of the improvement. While this is missing important details needed to compute standardized effect sizes, it is perhaps the most common phrasing that one encounters in the news. Next were two formats that contained information about the magnitude of the improvement, showing the expected benefit from the treatment in absolute and percentage terms. This simulates scenarios where one may learn about the size of an improvement without necessarily having context for the scale on which outcomes are measured. Finally, we tested two other formats commonly used in scientific publications: showing 95% confidence intervals to convey uncertainty in estimating mean differences, both with and without a corresponding visualization.

Experimental Design

Both of our experiments presented participants with the same fictitious scenario that we designed and used in a previous study [8] to accurately measure perceptions of treatment effectiveness while remaining both easily understandable by laypeople and relatively free of biases or priors that might be attached to any particular real-world treatments. Specifically, participants were shown a fictitious scenario in which they are competing against an equally-skilled opponent named Blog in the up and coming sport of boulder sliding. The goal is to slide their boulder farther than Blorg's, and they alone have the option of renting a premium boulder (the treatment) that is expected (but not guaranteed) to slide farther than the standard boulder that Blorg will use. There is an all-or-nothing 250 Ice Dollar prize for the winner.

Participants were randomly assigned to see information about the standard and premium boulders in one of five formats: *i) directional:* "The premium boulder slid further than the standard boulder, on average"; *ii) absolute difference:* "The premium boulder slid 4 meters further than the standard boulder, on average"; *iii) percentage difference:* "The premium boulder slid 4% further than the standard boulder, on average"; *iv) confidence interval without visualization:* "The average sliding distance with the standard boulder is 100 meters and a 95 % confidence interval is 99 to 101 meters. The average sliding distance with the premium boulder is 104 meters, and a 95% confidence interval is 103 to 105 meters"; *v) confidence interval with visualization:* The same statement as in the previous condition, along with a visualization that displays the confidence interval, as shown in Fig. 1.



Some information about the standard boulder and the premium boulder

Here are some statistics, based on estimates for an athlete like you.

- **Your average sliding distance with the standard boulder**: If you were to slide the **standard boulder** 1,000 times, you would attain an **average** distance of **100** meters. A 95% confidence interval on your average sliding distance with a **standard boulder** is **99** to **101** meters.
- **Your average sliding distance with the premium boulder**: If you were to slide the **premium boulder** 1,000 times, you would attain an **average** distance of **104** meters. A 95% confidence interval on your average sliding distance with a **premium boulder** is **103** to **105** meters.

**A 95% confidence interval conveys the uncertainty in estimating your true average sliding distance. It is constructed such that if we watched many such sessions of 1,000 slides and repeated this process, 95% of the constructed intervals would contain your true average.

The graph shows your average distance and this interval with the **standard boulder** (left) and the **premium boulder** (right), as indicated by the black points and vertical bars.

Figure 1. The 95% confidence interval visualization format.

For the last two conditions we added the following text to help participants understand what a 95% confidence interval represents: "A 95% confidence interval conveys the uncertainty in estimating your true average sliding distance. It is constructed such that if we watched many such sessions of 1,000 slides and repeated this process, 95% of the constructed intervals would contain your true average."

We fixed the actual parameters of the standard and premium boulders in the two experiments, choosing values that were representative of treatment effects studied in practice. Specifically, the difference between the standard and premium boulders was set to correspond to a Cohen's d of 0.25. This is equivalent to an underlying probability of superiority of 57% for the premium boulder over the standard one, and corresponds to a normative risk-neutral willingness to pay of 17.5 Ice Dollars for the premium boulder, calculated as the difference in expected value between using the premium boulder (250 × 57%) and using the standard boulder (250 × 50%).

Participants

We recruited 750 participants from Amazon's Mechanical Turk and randomly assigned them to conditions (148 in directional statement, 145 in absolute difference, 162 in percent difference, 156 in 95% confidence interval without visualization, and 139 in 95% confidence interval with visualization). We made the HIT available to U.S. workers with an approval rating of 97% or higher and paid a flat fee of $0.50 for completing the task. We prevented workers from taking the HIT if they participated in any of our pilots. The average time to complete the task was 3.0 minutes (SD = 4.4 minutes), with no significant difference between conditions ($F_{(4,745)}$=1.69, p=0.149).

Procedure

Participants were first presented with a brief introduction to the HIT and asked to sign a consent form indicating that they

[1] http://aspredicted.org/blind.php?x=bk4dy5

agreed to partake in the study. Then they were told that they would be asked to make a decision about an uncertain event, and provided with a brief training on how to answer the types of questions they would be presented with later in the study.

Next they were introduced to the boulder sliding competition and shown information about the standard and premium boulders in one of the five formats listed above. We first asked them to estimate the probability of superiority for the premium boulder: *If you were to compete with Blorg 100 times where you had the premium boulder and Blorg had a standard boulder, what is your best estimate of the number of times you would win?* And next asked for their willingness to pay: *Given that you'll win 250 Ice Dollars if you beat Blorg, but nothing if you lose, what is the most you would be willing to pay to use the premium boulder?*

Results

To measure how accurately participants perceived the effect of the premium boulder, we calculated the *error in willingness to pay* for the premium boulder by taking the absolute difference between each participant's stated willingness to pay for the treatment and the normative value (17.5 Ice Dollars, as calculated in the previous section). We also computed participants' *error in probability of superiority* for the premium boulder by taking the absolute difference between each participant's stated probability of superiority and the true probability of superiority (57%). Following our pre-registration plan, we used a one-way ANOVA to evaluate whether the format in which mean differences are presented affects perceived effect size and to identify the worst-performing format.

Willingness to pay. Participants were willing to pay substantially more for the premium boulder than the risk-neutral price of 17.5 Ice Dollars across all conditions, with an average error of anywhere from 41 Ice Dollars in the percentage difference condition to more than 66 Ice Dollars when they were shown 95% confidence intervals. A one-way ANOVA confirms that these differences between conditions are statistically significant ($F_{(4,745)}$=5.92, p<0.001), with the 95% confidence interval visualization condition performing directionally worst. A linear regression comparing this condition (M=66.1, SD=56.7) to all others shows there is no statistically significant difference if the visualization is removed (M=61.0, SD=54.0, t=-0.87, p=0.38) or between this condition and the directional statement (M=61.9, SD=51.5, t=-0.71, p=0.48), whereas other conditions have comparatively lower error (percentage difference: M=41.2, SD=42.0, t=-4.29, p<0.001, absolute difference: M=53.1, SD=46.1, t=2.18, p<0.01).

Probability of Superiority. We found a similar pattern for participants' perceptions of the probability of superiority for the premium boulder (Fig. 2), with even more extreme results. Once again, participants who saw the 95% confidence interval visualization performed worst (M=30.4, SD=13.6), followed by those who saw 95% confidence intervals without a visualization (M=25.3, SD=15.2, t=-3.22, p<0.01). Relative to the 95% confidence interval visualization condition, participants that were exposed to percent differences (M=13.7, SD=13.2, t=-10.49, p<0.001), absolute differences (M=18.7, SD=12.8, t=-7.19, p<0.001), and the directional statement (M=17.5, SD=12.5, t=-7.96, p<0.001) perceived the effectiveness of premium boulders more accurately, but participants overestimated the effectiveness of the premium boulder by more than 15 percentage points across all conditions. To our surprise, a treatment with a 57% probability of superiority was perceived as having around 90% probability of superiority when results were presented with a graph of means and 95% confidence intervals.

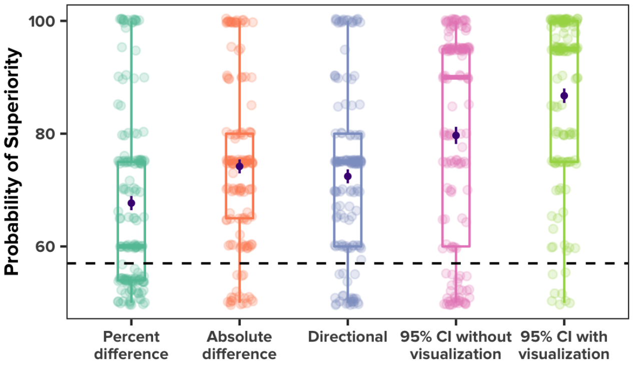


Figure 2. The stated chance of winning by condition. Jittered points show individual responses, with box plots to depict quantiles. The dark dots show the mean in each condition with error bars showing one standard error, and the dashed line shows the true probability of superiority.

The results of our first experiment demonstrate that phrasing treatments in terms of mean differences alone can lead people to overestimate their effectiveness. Interestingly, we see that following conventional guidelines [1] and providing readers with 95% confidence intervals—that is, strictly *more* information than simple mean differences—can in some cases exacerbate this problem. We suspect this is due to readers confusing inferential uncertainty with outcome uncertainty (i.e., how precisely a mean is estimated with how much outcomes vary around the mean) [8], which we investigate next.

STUDY 2: CORRECTING MISPERCEPTIONS

Our previous study showed that common ways of communicating treatments—specifically in terms of mean differences—can cause readers to overestimate treatment effectiveness. In this experiment, we explore ways to correct this. We first present readers with the most biasing condition from our previous study (the 95% confidence interval visualization, Fig 1) and elicit willingness to pay and perceived probability of superiority. Then we present additional information about variability in individual outcomes and give participants the opportunity to revise their responses to the previous questions.

We explore five formats to convey outcome uncertainty, the simplest being Cohen's categorical labels [4] that classify an effect as "small", "medium", or "large" according to Cohen's *d*. We compare this to a variance condition where we directly give participants information about how much outcomes vary around their average values. This contains all of the information necessary to compute a standardized effect size, but does not present the reader with effect size information directly. We also look at direct measures of standardized effect size that simultaneously incorporate information about both mean differences and variation in individual outcomes. Specifically, in one condition we show readers the probability of superiority for the treatment, which is thought to be easily understood by laypeople [9]. Finally, inspired by Cohen's own suggestion from over 30 years ago, we test two other conditions that compare the treatment to more familiar effects such as differences in height by age and weather over time.

Experimental Design

Participants were randomly assigned to see information about outcome uncertainty in one of five formats or no such information in a control condition: *i) category:* "The difference in the average sliding distance between the standard boulder and the premium boulder is *small* relative to how much individual slides vary around their long-run average"; *ii) variance:* "95% of your next 1,000 slides with the standard boulder would be between 70 and 130 meters and 74 and 134 meters with the premium boulder"; *iii) probability of superiority:* "If you were to play 100 times where you had the premium boulder and Blorg had a standard boulder, you would expect to win 57 times"; *iv) height analogy:* "The premium boulder will beat the standard boulder about as often as a randomly selected 16 year old is taller than a randomly selected 15 year old, among American women"; *v) weather analogy:* "The premium boulder will beat the standard boulder about as often as the maximum temperature on February 15th is higher than the maximum temperature on January 15th in New York City"; *v) control:* Participants in this condition are prompted to revise their willingness to pay and the probability of superiority without any additional information being given.

Participants

We recruited 1,800 participants from Amazon's Mechanical Turk and randomly assigned them to conditions (298 in control, 304 in category, 309 in variance, 302 in probability of superiority, 289 in height analogy, and 298 in weather analogy). We made our HIT available to U.S. workers with 97% or more approval rate and paid $1.00 for the task. We prevented workers from completing the HIT if they had completed Study 1 or previous pilots. The average time to complete the task was 6.3 minutes (SD=5.9 minutes), with no difference in the completion time between conditions ($F_{1,1798}$=1.82, p=0.177).

Procedure

The first part of this experiment was identical to the previous study, with the exception that all participants initially saw information about the premium boulder in the same format, the 95% confidence interval visualization shown in Figure 1.

After participants submitted their willingness to pay and probability of superiority for the premium boulder, they were told that they would have a chance to revise their estimates. They were shown additional information in one of the five formats mentioned above (or no extra information in a control condition) and asked to update their willingness to pay and probability of superiority.

Results

Similar to the previous experiment, we analyzed participants' willingness to pay for the premium boulder and their estimated probability of winning if they used it. In contrast to the previous experiment, however, we had two measurements for each of these quantities: an initial measurement before they saw information about individual outcome uncertainty and a revised measurement afterwards. We computed the absolute error in all four quantities by comparing each to its normative value (17.5 Ice Dollars for willingness to pay and 57% for probability of superiority).

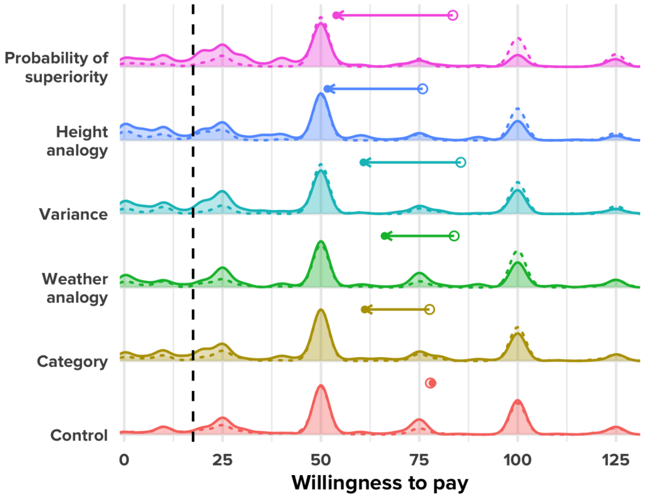


Figure 3. The distributions of initial willingness to pay (dashed lines) and the revised willingness to pay (solid lines) by condition. The empty circles indicate the mean of the initial responses in each condition, and the filled circles indicate the mean of the revised responses. The vertical dashed line shows the normative willingness to pay value. For readability this plot excludes responses greater than 130 (13.7% of responses).

We looked at shifts in each dependent variable in two ways. First, we compared the full distributions of responses before and after showing outcome variability information to each other. Then we examined within-participant shifts in responses using mixed effects linear models (one for willingness to pay and another for estimated probability of superiority). The models estimate the absolute error in a participant's revised response for each measure based on the absolute error in their initial response, with a variable slope and intercept for each condition k:

$$y_i^{revised} = \alpha_0 + \beta_0 \, y_i^{initial} + \sum_k \mathbf{1}_{c_i=k} \left(\alpha_k + \beta_k \, y_i^{initial} \right),$$

where i indexes each participant and c_i is the condition they were assigned to.

Willingness to pay. Figure 3 shows the distributions of willingness to pay for the premium boulder by condition before (dashed lines) and after (solid lines) seeing outcome uncertainty information. First, there is a strong round number effect in responses across all conditions, with many people submitting initial values of 50 or 100. Second, showing outcome uncertainty of any kind substantially improved the accuracy of responses compared to the control condition, where responses mostly remained unchanged. Much of this improvement comes from moving people away from round number responses (e.g., from 100 to lower values). And third, a larger fraction of participants revised their estimates downwards in the probability of superiority condition than in other conditions, with the height analogy and variance formats showing similar improvements.

We used the linear model above to quantify these improvements at the individual participant level. Specifically, we computed the average within-participant reduction in error for each condition from the slopes of the fitted model. Participants assigned to the probability of superiority condition (M=39.9, SD=46.9) had the largest error reduction (53% on average), however there was no statistically significant difference between this format and either the height analogy condition (M=39.5, SD=41.3, t=0.37, p=0.71) or the variance condition (M=47.5, SD=50.9, t=1.60, p=0.11). The weather analogy format and the category condition were significantly less efficient at reducing errors in willingness to pay (M=52.2, SD=48.2, t=3.16, p<0.01 and M=46.7, SD=45.6, t=5.00, p<0.001) than the probability of superiority format.

Probability of Superiority. We see a similar ranking of formats for error reduction in estimating the probability of superiority of the premium boulder as we saw with willingness to pay. Unsurprisingly, participants who were shown the actual probability of superiority did best (92.8%, on average), as all they had to do was recall a value they had previously seen. The variance and height analogy formats were next, with the weather analogy and category conditions reducing errors the least. Regardless, all formats for conveying outcome uncertainty showed statistically significant improvements over the control condition (M=16.5, SD=14.0, t=-9.37, p<0.001 for variance; M=16.0, SD=13.2, t=-8.16, p<0.001 for height analogy; M=22.9, SD=15.0, t=-4.65, p<0.001 for weather analogy; M=19.4, SD=14.9, t=-4.12, p<0.001 for category).

The results of our second experiment demonstrate that while showing only mean differences can cause people to overestimate treatment effectiveness, adding information about variability in individual outcomes can substantially reduce these misperceptions. Stating outcome variability in terms of probability of superiority was (directionally) best, although a non-quantitative analogy in terms of differences in height by age performed similarly, as did showing variance explicitly.

DISCUSSION AND CONCLUSION

How effective do people think treatments are when they are summarized in terms of only their average effects? Four common ways of summarizing results led participants to overestimate treatment effectiveness, as proxied through two variables: willingness to pay for a treatment (relative to a reasonable norm) and perceived probability of superiority. A surprising result of this study was that the inclusion of 95% confidence intervals increased both error and variance in perceptions of probability of superiority. A treatment with a 57% probability of superiority was perceived as having around 90% probability of superiority when results were presented with a graph of means and 95% confidence intervals. We do not suggest omitting confidence intervals in descriptions of scientific results. On the contrary, we endorse their use. However it is important to know they can—at least in circumstances like those tested here—have a biasing effect and that these biases can be countered with information about variability in outcomes. How do these initial perceptions change after people are presented with information about how individual outcomes vary around the average effect? We investigated how five textual information formats that convey this information cause people to update their willingness to pay for a treatment. Of the formats tested, probability of superiority was most effective and not substantially different than showing the variance in outcomes or simply using an analogy comparing people's heights at different ages. The latter format is notable in that it does not require much in the way of statistical literacy to comprehend.

We could summarize these results by saying that formats such as probability of superiority cut errors by more than half, on average. But, in the spirit of our findings, we think it might be more effective to phrase our results as follows: there was a 62% chance that people who received outcome variability information made better inferences than those who did not. To put this in perspective, that is about equal to the probability, among American women, that a randomly selected 18 year old is taller than a randomly selected 13 year old.

REFERENCES

[1] American Psychological Association and others. 1996. Task force on statistical inference initial report. *American Psychological Association PsycNET* (1996).

[2] Margaret E Brooks, Dev K Dalal, and Kevin P Nolan. 2014. Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology* 99, 2 (2014), 332.

[3] Robert Coe. 2002. It's the effect size, stupid: What effect size is and why it is important. (2002).

[4] Jacob Cohen. 1988. Statistical power analysis for the social sciences. (1988).

[5] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern Statistical Methods for HCI*. Springer, 291–330.

[6] William P Dunlap. 1994. Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin* 116, 3 (1994), 509.

[7] Louise Hartley, Nadine Flowers, Jennifer Holmes, Aileen Clarke, Saverio Stranges, Lee Hooper, and Karen Rees. 2013. Green and black tea for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews* 6 (2013).

[8] Jake M Hofman, Daniel G Goldstein, and Jessica Hullman. 2020. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.

[9] Kenneth O McGraw and SP Wong. 1992. A common language effect size statistic. *Psychological bulletin* 111, 2 (1992), 361.

[10] Gail M Sullivan and Richard Feinn. 2012. Using effect size or why the P value is not enough. *Journal of graduate medical education* 4, 3 (2012), 279–282.