Putting scientific results in perspective: Improving the communication of standardized effect sizes

Yea-Seul Kim
University of Wisconsin-Madison
Madison, Wisconsin, USA

Jake M. Hofman
Microsoft Research
New York, New York, USA

Daniel G. Goldstein
Microsoft Research
New York, New York, USA

ABSTRACT

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

KEYWORDS

statistics, effect size, visualization, perception

ACM Reference Format:
Yea-Seul Kim, Jake M. Hofman, and Daniel G. Goldstein. 2022. Putting scientific results in perspective: Improving the communication of standardized effect sizes . In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3502053

1 INTRODUCTION

As the world becomes more data-driven, people are increasingly exposed to statistical information about uncertain outcomes. In the field of HCI, for instance, researchers strive to quantify and communicate statistical uncertainty in their results [42, 43]. Likewise, other scientific domains face similar challenges in communicating results to audiences that may not be experts in their respective fields. 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* [33]).

While mean differences provide an indication of treatment effectiveness, they also rely on domain knowledge (e.g., familiarity with units of mmol/L in the green tea example and whether 0.62 mmol/L is large or small) and mask potentially important information about how outcomes vary around group averages. The latter is especially important for individual-level decision making, where one is concerned with what their own particular outcome is likely to be, as opposed to the average outcome for a large group of people.

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 cholesterol while those who take the second all have outcomes close to 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 [19]. 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: $d = \frac{\mu_1 - \mu_2}{\sigma}$.

Unfortunately—and despite calls from the HCI community [25, 38, 56] and many other scientific communities [2, 3, 18, 19, 59]—it remains rare that scientists report measures of standardized effect size in their published work. In fact, as we show below in a comprehensive review of every award-winning paper at CHI 2020, only a handful of these papers report standardized effect sizes. Furthermore, it is 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.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502053

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".[1] 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, or CLES [27, 51]), 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. 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 years old and 18 years old American women [21].

While there has been a great deal of discussion around alternatives for computing and reporting standardized effect sizes, there has not to our knowledge been any research to assess how people perceive effects when statistical results are presented in these different formats. In this work, we ask what can be done to accurately communicate the effectiveness of an uncertain treatment to laypeople. We contribute a sequence of four large-scale, pre-registered, randomized experiments involving close to 5,000 participants to investigate how to best communicate effect sizes, centered around two main research questions:

- **Research Question 1:** How effective do people think a treatment is when the treatment is summarized only in terms of its average effect?
- **Research Question 2:** How do these initial perceptions change after people are presented with information about how individual outcomes vary around the average effect?

All four of our experiments use a similar framework where participants read a scenario about a fictitious competition in which their performance can potentially be improved by paying for a treatment. We vary the way in which information about this treatment is presented to readers and measure how each format affects their willingness to pay for the treatment and their estimated probability of winning under it. We compare responses to reasonable norms to assess the biases introduced by each format.

In the first experiment, we assess the status quo by exploring ways of presenting the treatment that are commonly found in popular books and articles, ranging from simple directional statements to visualizations of statistical estimates. Regardless of the specific format, we find that summarizing the treatment in terms of only mean differences can lead to significant overestimation of treatment effectiveness, and, somewhat surprisingly, that including confidence intervals can, in some cases, exacerbate the problem.

In the subsequent three experiments, we attempt to remedy these issues by adding information about variability in individual outcomes in several different formats, including explicit statements about variance, probability of superiority for the treatment, and analogies that compare the treatment with more familiar effects, similar to the ones Cohen used in his textbook. We find that all of these formats substantially reduce potential misperceptions and that effect size analogies can be as helpful as more precise quantitative statements of standardized effect size.

[1]Cohen warned that standards for these categories would likely vary across the social sciences, which has since been confirmed [12, 53].

2 BACKGROUND & RELATED WORK

2.1 Communicating Effect Size

Null Hypothesis Significance Testing (NHST) is a standard practice in scientific reporting, but many have suggested that it be de-emphasized in favor of communicating effect sizes [9, 20, 46–48, 52, 58]. Broadly speaking, much of NHST focuses on whether differences between two or more groups systematically deviate from a fixed value (often taken to be zero), whereas effect sizes focus on how large of a difference exists between these groups [18]. Though there is no unified standard for how to report effect sizes, existing guidelines provide various options to calculate and communicate them [27, 30, 51, 57]. Some researchers advocate for presenting "simple" effect size measures, such as raw mean differences between groups [4, 5, 26, 54], whereas others exclusively consider "standardized" measures of effect size such as Cohen's d [37, 39]. However, Cummings et al. [23] show that, even after many calls to shift to reporting standardized effect sizes, fewer than half of the figures they surveyed show error bars of any type, encoding only mean differences; of those that do show error bars, those error bars most commonly represent one or two standard errors on a mean, the latter being one of the formats we test. In our work we compare mean differences (a simple effect size) to several standardized effect sizes that incorporate variation in individual outcomes.

While much has been written on developing and advocating for different measures of effect size and methods for estimating effect sizes (e.g., [44]), relatively little work has been done on how people *perceive* effect sizes that they are exposed to. One exception is work by Hofman et al. [35], which looks at how people perceive different visual representations of uncertainty commonly found in scientific publications. This work finds that visualizations depicting inferential uncertainty (e.g., plots containing standard errors or confidence intervals around parameter estimates) lead people to overestimate standardized effect sizes compared to visualizations that show outcome variability (e.g., plots showing standard deviations or prediction intervals). Here we extend this work to include a wider range of scenarios that are more commonly encountered by laypeople. Specifically, we broaden the focus from only visual representations of effect sizes to include text-based representations, and from scientific publications to more general reporting of treatment effectiveness that are more likely to be encountered in popular books and articles. We further suggest how to alleviate potential misperceptions of standardized effect sizes using simple text-based reporting and analogies for standardized effect sizes.

There are also studies of how people perceive effect sizes from the psychology literature. For instance, Funder and Ozer discussed how different ways of reporting effect sizes could be interpreted in the context of psychology studies [32]. Other work done by Brooks et al. [13] compares "traditional" measures of effect size to "non-traditional" measures like the probability of superiority. Though similar in spirit, there are a few key differences between their work and ours. First and foremost, Brooks et al. assume that the standard

for communicating effect sizes are measures such as Pearson's correlation coefficient (r) and the coefficient of determination (r^2), and compare alternative measures such as the probability of superiority to this baseline. We, however, use mean differences as a baseline, as these are much more commonly communicated to laypeople than measures like r and r^2. This allows us to assess biases introduced by the status quo in popular accounts of scientific studies. We also explore several ways to improve upon mean differences not explored by Brooks et al., including explicit statements about variance and analogies to more familiar terms, a technique that has been shown to help contextualize unfamiliar numbers in other settings [7, 36, 45, 55]. Another difference is that we collect a continuous measure of willingness to pay and compare this to a normative (risk-neutral) value, whereas Brooks et al. use an ordinal scale in a setting without any such normative value. Finally, the substantially larger sample size in our studies allows us to investigate effects that they are unable to estimate.

2.2 Relevance to the CHI Community

Despite many calls from the HCI community for improved statistical communication and reporting of effect sizes, our comprehensive review of every award-winning paper at CHI 2020 shows that these practices are still quite rare in the community. This is especially unfortunate given that HCI is an applied field that embraces a diverse set of methods, measurement techniques, and statistical approaches. As such, without effect size reporting HCI research does not always lend itself to easy comparisons across studies.

2.2.1 Transparent Statistical Communication in HCI. The HCI community has been mindful of developing strategies for communicating statistics in an accurate, transparent, and helpful manner [15, 17, 43, 60]. Among other concerns, conveying the actual magnitude of effects and the practical importance of findings have been emphasized by many HCI researchers [25, 42, 50, 56]. In particular, Dunlop and Baillie argue that reporting the results of statistical tests alone (e.g., p-values and test statistics) without presenting effect sizes can be particularly problematic in HCI, as compounding factors that introduce noise in measuring human behavior may distort the perceived value of effects [28]. Additionally, given the prevalence of small-sample studies and the relative lack of meta-analyses in the field, some researchers have advocated for Bayesian analyses in the HCI community [41, 43] to shift the focus from dichotomous significance testing to consider the magnitude and variability of estimated effects.

In addition to publishing papers calling for transparent statistical communication, HCI researchers have also developed systems that assist in designing experiments and analyzing results from them [29, 40, 49, 61]. For example, Touchstone2 provides an interactive environment for experimental design and facilitates power calculations based on targeted effect sizes [29], whereas Tea provides a language for automating statistical analyses given an experimental design and reports effect sizes as a result [40].

Our work contributes to this literature in two ways. First, it provides a quantitative assessment of potential misperceptions introduced by standard statistical reporting practices that many have criticized. Second, it demonstrates the benefits to be had by shifting focus to effect size reporting, as per the transparent statistics guidelines set forth by the HCI community [1].

2.2.2 Statistical reporting in the CHI community. Experiments are a prevalent method of evaluating hypotheses in the HCI community, ranging from systems research that aims to validate a system's efficacy to empirical findings that reveal insights about how people behave and interact. As such, the way in which results of HCI experiments are reported by authors and perceived by readers can have important implications for which methods and systems are adopted in the community.

To better understand statistical reporting practices in the CHI community, we collected all papers that a received a best paper or honorable mention award at CHI 2020 (151 papers total). Then we checked whether each paper contained an experiment by searching for the keywords 'experiment' or 'study', resulting in 109 papers. Then we filtered out papers that did not contain a quantitative experiment by reading each paper, leaving 49 papers. We coded the papers by noting whether they communicated 1) outcome variability (e.g., standard deviations or prediction intervals (PIs)), 2) inferential uncertainty (e.g., standard errors or confidence interval (CIs)), 3) simple effect sizes (e.g., mean or percentage differences between two or more groups), and 4) standardized effect sizes (e.g., Cohen's d).

Figure 1 shows the results of this analysis. First and foremost, we note that despite calls from many communities (HCI key among them) for increased effect size reporting, it is quite rare for even award-winning papers at CHI to report effect sizes (simple or standardized; 8 out of 49, 16%). Instead, it is most common that authors report outcome variability by writing out standard deviations in text (29 out of 49, 60.0%), as is suggested by APA style guidelines. While this is somewhat helpful, past work has shown that when both statistical visualizations and text-based statistics are reported, visualizations dominate text in terms of people's perceptions of statistical effects [35]. Looking at only visualizations contained in these papers, we see that they are nearly evenly split between displays of outcome variability (PIs; 11 out of 49, 22%) and inferential uncertainty (CIs; 14 out of 49, 28%), with slightly more of the latter than the former.

While these results are in line with a recent increase in reporting of CIs in HCI noted by Besançon & Dragicevic [10], they are also astonishingly similar to those found by Cummings et al. from over a decade ago [23]. As such, our analysis highlights the need for better communication of standardized effect sizes in the HCI community. Next, we turn to a series of four pre-registered experiments to assess how we might go about improving effect size reporting as a community, and the benefits this would confer.

3 OVERVIEW OF EXPERIMENTS

We conducted four interrelated experiments comprising responses from nearly 5,000 participants to investigate how to communicate effect sizes to laypeople, where the results of one study informed the design of the next. We pre-registered the entire sequence, summarized in Fig. 2, in advance[2].

[2]https://aspredicted.org/zd8w2.pdf

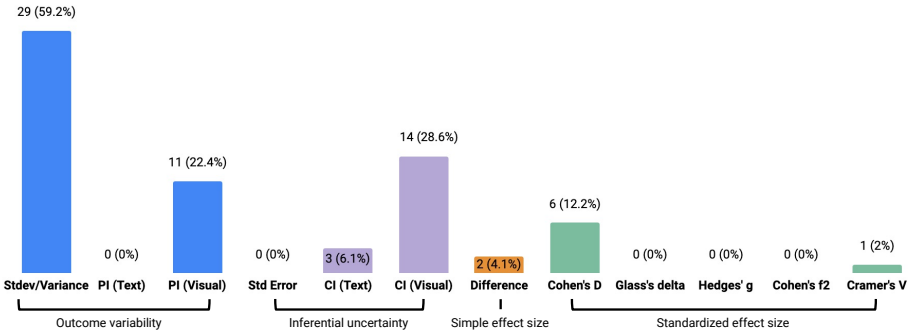


Figure 1: The frequency of various types of statistics reported in the 49 award-winning papers at CHI 2020 that contained quantitative experiments. Outcome variability is often reported in the text, whereas visualizations are nearly evenly split between displays of outcome variability and inferential uncertainty. In contrast, effect sizes (whether simple or standardized) are rarely reported.

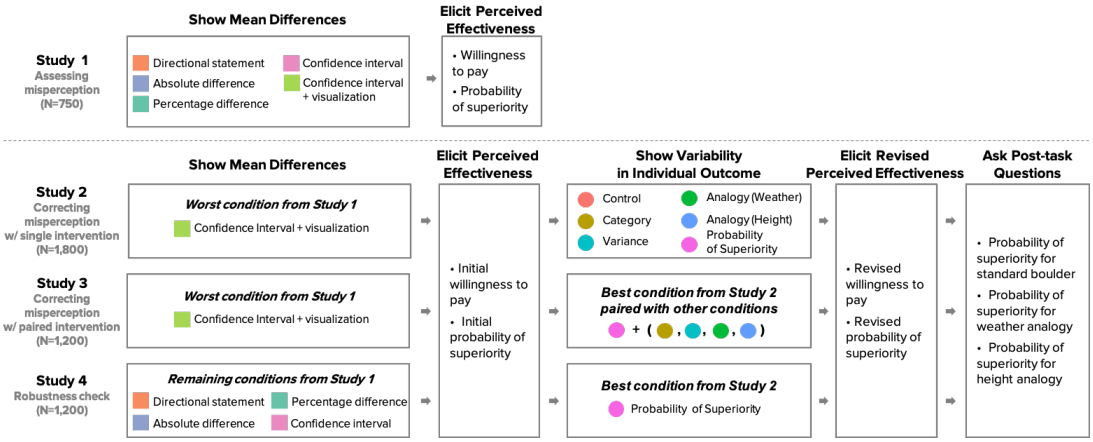


Figure 2: An overview of the sequence of four interrelated experiments we conducted. Each row represents one study.

All of our experiments presented participants with the same fictitious scenario used in prior work to evaluate people's effect size perception [35]. The scenario measures 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 told that they are athletes competing against an equally skilled opponent named Blorg. The goal is to slide their boulder farther than Blorg's, and there is an all-or-nothing 250 Ice Dollar prize for the winner. While Blorg is known to always use a standard boulder, participants have the option of renting a premium boulder (i.e., the treatment) known to slide further on average than Blorg's boulder. Participants were shown information about the effectiveness of the premium boulder, after which they were asked how much they were willing to pay for it and to estimate the probability of winning if they used it. We chose these outcomes because they reflect the types of individual-level decisions made by people on a daily basis (as opposed to, for instance, decisions made by policymakers that might place more emphasis on mean differences regardless of variation in individual-level outcomes). Willingness to pay is our primary dependent measure. In addition to being a standard measure of the value of treatments in health economics and consumer behavior [6], willingness to pay in our scenarios has the added advantage of having a normative value based on the probabilities and prize money presented. Finally, because people decide to invest in treatments they read about in the media (for example, buying running shoes that they claim to increase speed by 4%), willingness to pay is an ecologically valid measure.

We fixed the actual parameters of the standard and premium boulders across all four 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, which is the median effect size across a quasi-random sample of studies in psychology [22] and typical of effects studied in medicine, neuroscience, and the social sciences [8, 14, 16]. This is equivalent to an underlying probability of superiority of 57% for the premium boulder over the standard one. We achieved this by setting the mean of the standard and premium boulder sliding distances to 100 meters and 104 meters, respectively, each normally distributed with a standard deviation

of 15.3 meters so that 95% confidence intervals and 95% prediction intervals worked out to easily readable round numbers. This 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 ($250 \times 57\%$) and using the standard boulder ($250 \times 50\%$).[3] Our first experiment, summarized in the top row of Fig. 2, was the simplest of the four. Participants first saw information about the standard and premium boulder phrased in one of five mean difference formats and then stated their willingness to pay and perceived probability of superiority. This allowed us to determine which format caused people to overestimate treatment effectiveness the most, which turned out to be a visualization that depicted means and 95% confidence intervals for the standard and premium boulders.

We used this format as a starting point in each of our next two experiments to look at how well we could correct potential misperceptions of effect size. The idea was that if we could correct the biases introduced by showing 95% confidence intervals, we would be able to do the same for the other, less problematic formats.

Our second experiment started off identical to our first experiment, but all participants saw information about the premium boulder in the same mean difference format (a 95% confidence interval visualization), after which they were asked for willingness to pay and probability of superiority. At this point, we introduced additional information about variability in outcomes in one of five randomly selected formats, indicated in the second row of Fig. 2. After seeing this information (or nothing in a control condition), we asked participants if they would like to revise their previous answers and collected updated values for willingness to pay and probability of superiority. From this experiment, we learned that directly showing people the probability of superiority for the premium boulder was (directionally) the best format for reducing overestimation bias, with Cohen's height analogy and an explicit statement about individual outcome variance providing similar benefits.

In our third experiment, we asked whether we could improve upon the best single intervention (stating the probability of superiority for the premium boulder directly) by combining it with other formats. We repeated the previous experiment, but before asking for revised estimates, showed participants the probability of superiority for the premium boulder along with one of the other four formats for communicating outcome variability to provide additional context.

We used our fourth and final experiment as a robustness check for our previous findings. Specifically, we looked at the effectiveness of the best single format from Study 2 (probability of superiority) for correcting biases introduced by all mean difference formats from Study 1 other than the worst-performing format (the 95% confidence interval visualization), and found similar benefits.

We chose sample sizes for each experiment based on pilot data so that we would have 80% power in detecting effect sizes of a minimal interest (a 10% difference in relative error reduction) with a 5% false positive rate. Screenshots of all experiments and conditions are included as supplemental material[4], along with all data and secondary analyses from our pre-registration plan. In sum, these four experiments comprised of nearly 5,000 unique participants allowed us to address both of our main research questions in a reliable and robust manner. We provide further details of each experiment along with their results in the next four sections.

[3] A normative risk-*averse* willingness to pay would be even less. As will be seen, the choice between these common norms is not pivotal as the average willingness to pay is much greater than 17.50, even when participants revise initial answers.

[4] https://github.com/jhofman/effect-size-analogies-chi2022

4 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 books and articles. Participants were presented with information about a treatment in one of five formats with varying levels of detail. 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, it is perhaps the most common phrasing that one encounters in summaries of scientific findings (e.g., in scientific titles or abstracts, or in news stories covering these results) [24]. 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 conveying scientific results: showing 95% confidence intervals to convey uncertainty in estimating mean differences, both with and without a corresponding visualization.

4.1 Experimental Design

As mentioned above, participants were shown a fictitious scenario in which they are competing against an equally skilled opponent named Blorg 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:

- **Directional:** "The premium boulder slid further than the standard boulder, on average".
- **Absolute difference:** "The premium boulder slid 4 meters further than the standard boulder, on average".
- **Percentage difference:** "The premium boulder slid 4% further than the standard boulder, on average".
- **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".
- **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. 3.

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."

4.2 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 (i.e., Human Intelligence Task) 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).

4.3 Procedure

Participants were first presented with a brief introduction to the HIT and asked to sign a consent form indicating that they 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. Specifically, we asked them the following:

> Assume you and your friend are equally skilled at a game. If you were to play them at this game 100 times, how often do you think you would win (assuming this game does not have ties)?

If they answered "50", they were allowed to proceed. If not, they were shown a hint indicating that they should expect to win about half of the time and allowed to try again until they responded with "50".

On the next screen, we introduced the boulder sliding competition, as described above, and asked participants to check a box to confirm they understood the scenario before proceeding. At this point, they were shown a new screen with 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?

After submitting these two responses, participants were asked a final multiple choice question about their willingness to pay decision. This was an exploratory question to gain insight into if they made the decision based on the prize money, the feeling of winning, both, or neither. This concluded the experiment.

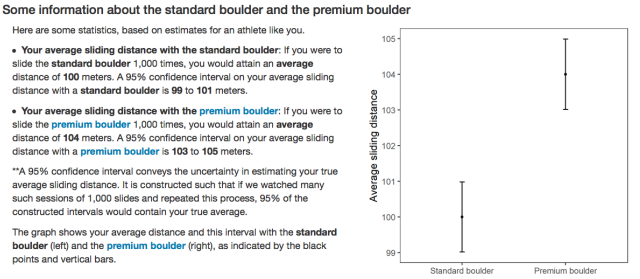


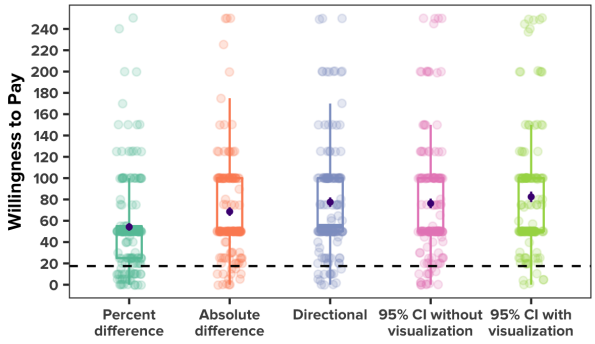
Figure 3: The 95% confidence interval visualization format used in our first three studies.

4.4 Results

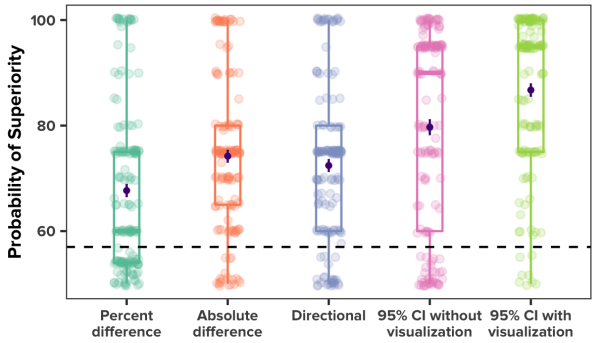
To measure how accurately participants perceived the effect of the premium boulder, we calculated the *error in willingness to pay* for the premium boulder by taking the absolute difference between each participant's stated willingness to pay for the treatment and the normative value (17.5 Ice Dollars, as calculated in the previous section by assuming a person is risk-neutral and maximizing their expected reward). We also computed participants' *error in the 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%). For the conditions that included confidence intervals, participants had enough information to compute both of these quantities exactly. The three other formats (directional, absolute difference, and percent difference) lacked complete information, but we still compare responses to normative values to measure the difference between how effective people perceive treatments to be compared to how effective they typically are. 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 identified the worst-performing format.[5]

Willingness to pay. As indicated in Fig. 4a, 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 to all others shows there is no statistically significant difference if the visualization is removed (t=-0.87, p=0.38) or between this condition and the directional statement (t=-0.71,

[5]In addition to analyzing error in these perceptions, we also analyzed the raw responses without comparing them to normative values, as declared in our pre-registration. The results, included in our supplemental material, show very similar patterns to what we present here. We also repeated our analyses using randomization inference (RI) to relax modeling assumptions (e.g., normalcy and sphericity required for ANOVAs), and confirmed that results are very similar to what we report under the pre-registered ANOVA analysis.



(a) The willingness to pay by condition. Jittered points show individual responses, with box plots overlayed to depict quantiles. Dark dots show the mean in each condition with error bars showing one standard error, and the dashed line shows the risk-neutral willingness to pay.



(b) The stated chance of winning by condition. Jittered points show individual responses, with box plots overlayed 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.

Figure 4: The result of Study 1.

p=0.48), whereas other conditions have comparatively lower error (percentage difference: t=-4.29, p<0.001, absolute difference: 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. 4b), with even more extreme results. Once again, participants who saw the 95% confidence interval visualization performed worst, followed by those who saw 95% confidence intervals without a visualization (t=-3.22,p<0.01). Relative to the 95% confidence interval visualization condition, participants that were exposed to percent differences (t=-10.49, p<0.001), absolute differences (t=-7.19, p<0.001), and the directional statement (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 95% confidence intervals.

We analyzed the final question in this experiment regarding participants' motivation of paying for the premium boulder to get a better sense of why responses deviated from the risk-neutral price. For instance, it could be the case that people have intrinsic value for the feeling of winning by itself, over and above the value of the payoff they would receive for doing so. Responses from this question, however, indicated that the majority of participants (68.3%) were solely concerned with the prize money alone, whereas a smaller fraction (21.1%) considered both the prize money and the feeling of winning. Relatively few people (6.9% of participants) considered only the feeling of winning.

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 [2] and providing readers with 95% confidence intervals—that is, strictly *more* information than simple mean differences—can in some cases exacerbate this problem. As per similar findings in [35], we suspect this is due to readers confusing inferential uncertainty with outcome variability (i.e., how precisely a mean is estimated with how much outcomes vary around the mean), which we investigate next.

5 STUDY 2: CORRECTING POTENTIAL 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) 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 [21] 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 [51]. Finally, inspired by Cohen's own suggestion from over 30 years ago, we test two other "analogy" conditions that compare the treatment to more familiar effects such as differences in height by age and weather over time. We hypothesize that the more direct the reported effect size measure is, the more accurately people will perceive that effect size. In particular, the condition in which participants are shown the true probability

of superiority represents an upper bound on how well we can expect people to perform, and thus it becomes a useful benchmark to compare other conditions, such as the effect size analogies, to.

5.1 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:

- **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".
- **Variance:** "Roughly speaking, 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."
- **Probability of superiority:** "Roughly speaking, 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."
- **Height analogy:** "Roughly speaking, 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."
- **Weather analogy:** "Roughly speaking, 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."
- **Control:** Participants in this condition are prompted to revise their willingness to pay and the probability of superiority without any additional information being given.

The height analogy was adapted directly from Cohen's textbook [19], where he contextualizes a d of 0.2 using this exact comparison. To make sure that this was still accurate, we calculated the actual probability of superiority for heights of 16 year old women compared to 15 year old women in the U.S. using data from the National Center for Health Statistics [31]. We found that Cohen's analogy matched the effect size of the premium boulder exactly and so used it directly in our studies.

We independently designed a second analogy that compares the effect of the premium boulder to differences in weather over time. We chose weather because people have a relatively large and most representative sample of temperatures during different times of the year. Ideally, we would have personalized the weather analogy to each participant's location, but doing imposes a number of technical hurdles (e.g., collecting locations, downloading historical weather data for those locations, and constructing the corresponding personalized analogies) that we leave as future work. Instead, we used New York City as a benchmark because it is the most populated and frequently visited city in the country, making it a reasonable reference point for many people. We collected the daily maximum temperatures for the last 100 years in New York City using data from the National Oceanic and Atmospheric Administration provided through Google Big Query [11], and found a pair of days (January 15th and February 15th) that had a probability of superiority of 57%.

5.2 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 $1.00 for completing the task. We prevented workers from completing the HIT if they had completed Study 1 or previous pilots. The average time to complete the task was 6.3 minutes (SD=5.9 minutes), with no difference in the completion time between conditions ($F_{1,1798}$=1.82, p=0.177).

5.3 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 3.

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. Upon clicking a checkbox and continuing, 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. Their previous answers were shown alongside an empty text box that required them to enter their revised responses.

This was followed by three post-task questions. The first was a comprehension check that asked participants to estimate how often they would win if they and Blorg both used a standard boulder. Then we asked two questions to gauge how people perceived the effect size analogies we created. On one page, we asked participants how often they think the maximum temperature on February 15th was higher than the maximum temperature on January 15th, out of the last 100 years in New York City. On the following page, we asked how often they think that a randomly selected 16-year-old American woman would be taller than a randomly selected 15-year-old American woman, out of 100 such pairs. After each of these questions, we prompted participants to confirm or revise their responses. The final page was identical to the previous study.

5.4 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. Because probability of superiority is both a treatment condition (one of the effect size communication formats we present) and a dependent measure, willingness to pay is the primary dependent measure. 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 afterward. 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).

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

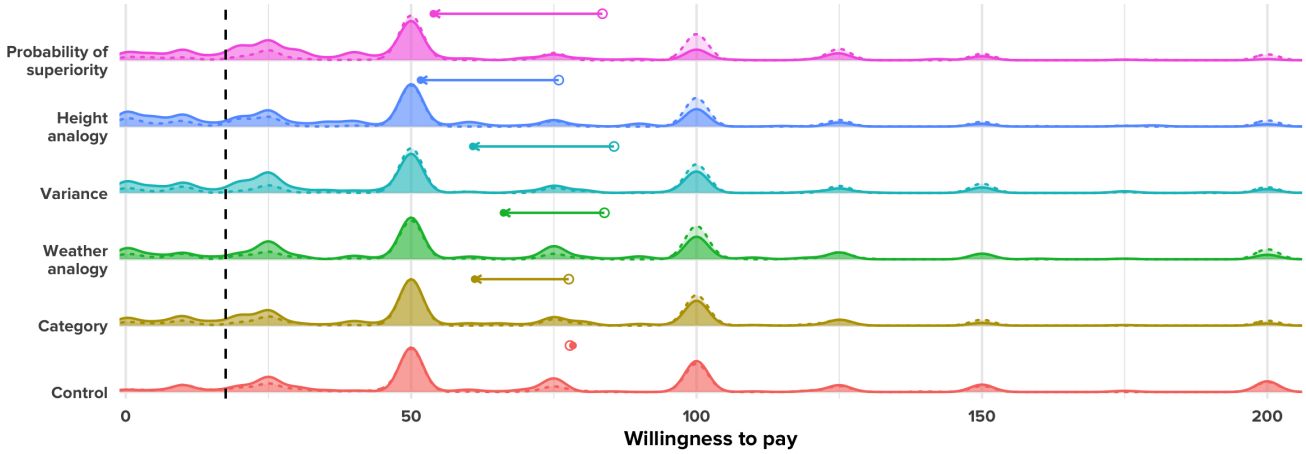


Figure 5: The distributions of initial willingness to pay (dashed lines) and the revised willingness to pay (solid lines) by condition. The empty circles indicate the mean of the initial responses in each condition, and the filled circles indicate the mean of the revised responses. The vertical dashed line shows the normative willingness to pay value. For readability, this plot excludes responses greater than 205 (3.9% of responses).

examined within-participant shifts in responses using 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:

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

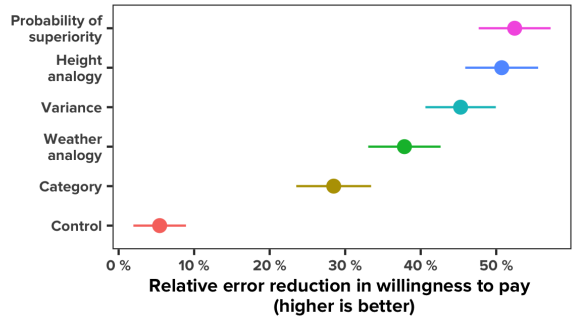
where i indexes each participant and c_i is the condition they were assigned to.

Willingness to pay. Figure 5 shows the distributions of willingness to pay for the premium boulder by condition both before (dashed lines) and after (solid lines) seeing outcome uncertainty information. The size and locations of the arrows show the shift in the average willingness to pay between initial and final responses in each condition. Three things are apparent from this plot. 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.

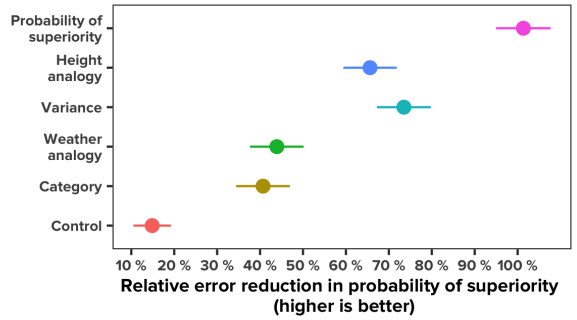
We used the linear model above to quantify these improvements at the individual participant level. Specifically, we computed the average within-participant reduction in error for each condition from the slopes of the fitted model, shown in Fig. 6a. Participants assigned to the probability of superiority condition had the largest error reduction (53% on average), however there was no statistically significant difference between this format and either the height analogy condition (t=0.37, p=0.71) or the variance condition (t=1.60, p=0.11). The weather analogy format and the category condition were significantly less effective for reducing errors in willingness to pay (t=3.16, p<0.01 and t=5.00, p<0.001) than the probability of superiority format.

Probability of Superiority. As shown in Fig 6b, 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, 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 (t=-9.37, p<0.001 for variance; t=-8.16, p<0.001 for height analogy; t=-4.65, p<0.001 for weather analogy; 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 potential 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. 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 this experiment, we think it might be more effective to phrase our results as follows: there is a 62% chance that error in willingness to pay for the premium boulder is higher when shown only mean differences compared to also seeing information about outcome variability. To put this in perspective, that is about equal to the probability that a randomly selected 18 year old American woman is taller than a randomly selected 13 year old American woman.



(a) The relative error reduction in willingness to pay, estimated by regressing each participant's final error against their initial error.



(b) The relative error reduction in stated probability of superiority, estimated by regressing each participant's final error against their initial error.

Figure 6: The relative error reduction in willingness to pay and the stated probability of superiority.

6 STUDY 3: PAIRED INTERVENTIONS

In our previous experiment we saw that several relatively different formats for communicating variability in individual outcomes were equally helpful for reducing potential misperceptions about treatment effectiveness. In this study, we investigate whether there are any complementarities between these formats. Specifically, we pair the best format from Study 2 (probability of superiority) with each of the four remaining outcome variability conditions and test for reductions of error. As in the example in the previous paragraph, we showed participants the probability of superiority for the standard boulder first, followed by a sentence that said, "To put this in perspective, ... " and showed either the category, variance, height analogy, or weather analogy formats.

6.1 Procedure & Participants

The procedure for this experiment was identical to Study 2 except that participants saw the probability of superiority format combined with one of the four other outcome variability formats (category, variance, height analogy, or weather analogy). There was no control condition in this experiment because that from Study 2 suffices.

We recruited 1,200 participants from Amazon's Mechanical Turk and randomly assigned them to conditions (301 in probability of superiority with category, 303 in probability of superiority with height analogy, 304 in probability of superiority with variance, 292 in probability of superiority with weather analogy). We again recruited U.S. workers with 97% approval rating or higher and paid $1.00, excluding workers had participated any of our previous pilots or studies. The average time to complete the task was 6.4 minutes (SD=4.4 minutes), with no difference in the completion time between conditions ($F_{1,1198}$=0.6, p=0.431).

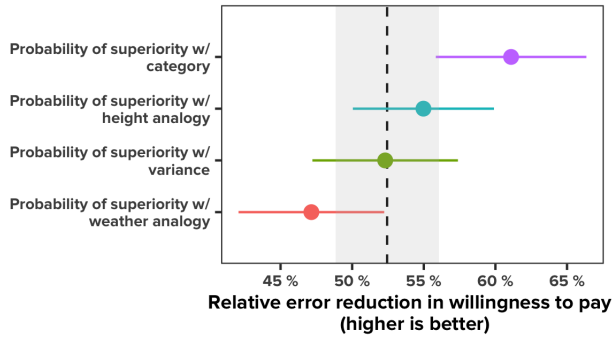


Figure 7: The relative error reduction in willingness to pay after seeing the combined interventions. The dashed line shows the mean error reduction from the probability of superiority condition alone from Study 2 (the shaded area shows one standard error).

6.2 Results

We analyzed the data using the same linear model as in Study 2. Only willingness to pay and revised willingness to pay were analyzed because the true probability of superiority was shown to all participants in all conditions. Figure 7 depicts the relative error reduction in willingness to pay after seeing the combined interventions. No combination of probability of superiority with another format was significantly better than probability of superiority alone (with category: t=-1.65, p=0.099, with height analogy: t=-0.52, p=0.607, with variance: t= 0.03, p=0.977, with weather analogy: t= 1.04, p=0.297).

The results of this experiment show that it is difficult to improve upon probability of superiority for reducing errors in perceived effect sizes. At the same time, we do not see any detrimental effects to showing additional information to help readers contextualize treatment effectiveness.

7 STUDY 4: ROBUSTNESS CHECK

Studies 2 and 3 demonstrated that explicitly showing information about outcome variability corrected potential misperceptions introduced by showing mean differences alone. However in both of those studies participants initially saw information about the premium boulder in just one of the mean difference formats that people frequently encounter: the 95% confidence interval visualization, which was the *most misleading* format we tested. In this study we check the robustness of our findings by first showing people information about the premium boulder in the *other* mean difference formats

from Study 1 and seeing if exposure to the probability of superiority format has the same normalizing effect.

7.1 Participants & Procedure

We recruited 1,200 participants from AMT who were randomly assigned to conditions (317 in percent difference, 271 in absolute difference, 311 in directional, 301 in 95% confidence interval without a visualization). We again recruited U.S. workers with 97% approval rating or higher and paid $1.00, excluding workers had participated any of our previous pilots or studies. The average time to complete the task was 5.5 minutes (SD=5.6 minutes), with no difference in the completion time between the conditions ($F_{1,1198}$=1.15, p=0.284).

Study 4 was similar to Studies 2 and 3, except that what varied between conditions was the mean difference format that participants saw before submitting their initial willingness to pay and probability of superiority. Participants were randomly assigned to one of four mean difference formats: a directional statement, percentage difference, absolute difference, or 95% confidence interval without a visualization. After submitting their initial responses, all participants saw outcome variability information in the probability of superiority format and were asked to revise their estimates, as in previous studies. Other details were identical to the previous two studies.

7.2 Results

We used the same linear model as in the previous two studies to analyze participants' willingness to pay. As shown in Fig. 8, we found large reductions in error for all conditions. Comparing these to the control condition from Study 2, we find that all gains are substantial and statistically significant (26.1% for percent difference, t=-5.91 p<0.001; 39.0% for absolute difference, 39.0%, t=-6.35 p<0.001; 39.8% for directional, t=-9.19 p<0.001; 40.8% for 95% confidence intervals without visualization, t=-9.57 p<0.001).

We confirmed that, regardless of which initial mean difference format people are shown, exposure to outcome variability in the form of probability of superiority statements reduces potential misperceptions of treatment effectiveness.

8 PERCEPTIONS OF EFFECT SIZE ANALOGIES

In three of our four studies we asked participants how often they thought the maximum temperature in New York City was higher on February 15th compared to January 15th and how often a randomly selected 16 year old American woman would be taller than a randomly selected 15 year old American woman.

We aggregated participants' responses from Studies 2 and 4 and compared this to the ground truth (57%), as shown in Fig. 9.[6] Participants had accurate perceptions for both of these analogies, on were only off by a few percentage points on average. Bias and variance are lower for the height analogy compared to the weather analogy, in line with our results that the height analogy was more effective than the weather analogy in debiasing. Cohen's height analogy proved to be surprisingly accurately perceived.

[6]We excluded Study 3 from this analysis because some participants saw the ground truth alongside the analogies during the study.

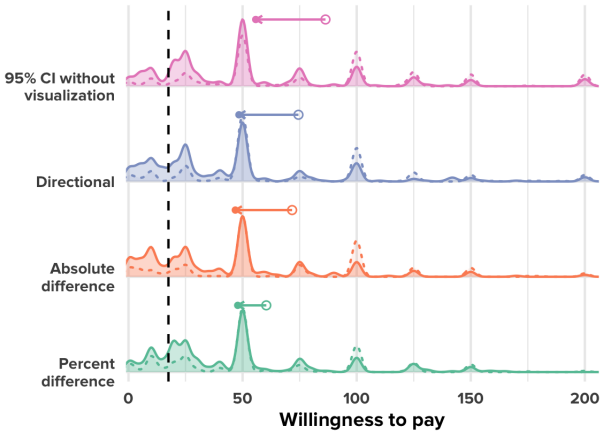


Figure 8: The distribution of initial willingness to pay (dotted lines) and final willingness to pay after seeing probability of superiority (solid lines) by condition for Study 4. The empty circles indicate the mean of the initial responses in each condition, and the filled circles indicate the mean of the revised responses. The vertical dashed line shows the normative willingness to pay value. For readability this plot excludes responses greater than 205 (3.25% of responses).

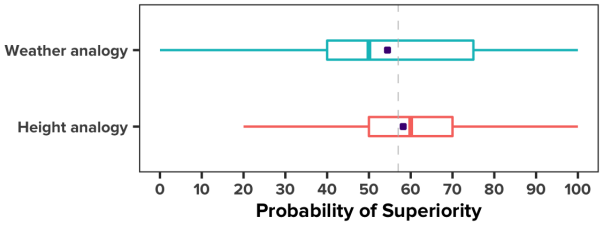


Figure 9: Boxplots showing distributions for the stated probability of superiority for the height and weather analogies. Points show the mean for each analogy. Error bars showing one standard error are present but exceedingly small. The horizontal gray line shows the true probability of superiority for the analogies (57%).

9 DISCUSSION

When results are summarized only in terms of average effects, as is common in news reporting, how do people perceive treatment effectiveness? Our studies found that four common ways of summarizing results led to potential misperceptions of treatment effectiveness, as proxied through two variables: willingness to pay for a treatment and perceived probability of superiority. A surprising result 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 confidence intervals. While we do not suggest omitting confidence intervals in descriptions of scientific results, we feel it is worth noting that they have biasing effects that

can be countered by providing simple information about outcome variability.

How do these initial perceptions change after people are presented with outcome variability information? We investigated how five textual information formats conveying outcome variability cause people to update their willingness to pay for a treatment. Of the formats tested, stating the probability of superiority was most effective, reducing error in willingness to pay by about 50%. Specifically, we observed a 62% chance that error in willingness to pay for a treatment was higher when participants only saw mean differences compared to seeing information about the probability of superiority. To put this in perspective, this shift is about equal to the probability that a randomly selected 18 year old American woman is taller than a randomly selected 13 year old American woman. More support for the use of probability of superiority came in the last study, in which it was shown that it was robust: it had a similar debiasing effect when applied to four different ways of presenting scientific results, from those found in journal articles to the merely directional claims that are common in everyday media.

For reducing error, showing the variance in outcomes or simply using an analogy comparing people's heights at different ages was not substantially different than showing the probability of superiority on average. These findings have a practical impact for scientists and journalists because the proposed statements can be formulated with little overhead; one only needs simple summary statistics to create them. In testing whether the best single format (probability of superiority) could be made more effective by combining it with other formats, we found that it could not, implying that authors are not making compromises when sticking with easy-to-read statements.

We were pleasantly surprised that the height analogy is, on average, about as effective as other precise quantitative measures such as the probability of superiority and variance. We speculate that beyond the benefits demonstrated in the paper, analogies may have other potential advantages, such as better user engagement and appeal to populations with low numerical literacy. In the optional feedback textbox we provided at the end of the study, we found some anecdotal evidence for how participants used the analogies to make their judgements. For instance, one participant mentioned why they thought the effect was small: "I think that most girls have stopped growing at around age 15, so that is why I think 15 and 16-year-olds would be about the same height." Some expressed their curiosity about the provided analogies as well (e.g., "I would like to know the actual stats about the weather!").

Given the results of our studies, we encourage authors to do the following when communicating their findings. First, provide the probability of superiority (or common language effect size) where applicable, which can be easily derived from Cohen's d: $P_{\text{superiority}} = \Phi(d/\sqrt{2})$, where Φ is the cumulative distribution function for the standard normal distribution, or as the area under the ROC curve (AUC) between conditions. This information may be especially helpful when authors provide visual confidence intervals to mitigate potential misconceptions. Using discrete outcomes to express the probability of superiority (e.g., "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") is preferred compared to using the percentage format (e.g., if you were to play where you had the premium boulder and Blorg had a standard boulder, you would expect to win with a 57% probability) as shown in prior work [34]. In addition, authors should consider reporting variance in outcomes. Reporting variance was one of the most prevalent strategies from the award-winning collection of CHI papers, appearing in around 60% of them. While our investigation found that reporting variance was less effective in decreasing misconceptions, it has the advantages of being familiar and is already provided by statistical software. Lastly, authors might consider providing analogies to make their results more engaging. We provide a script in the supplemental material that will calculate the height analogy for authors who would like to use it in their papers.

As for limitations of our work, here we have studied just one (hypothetical) setting, involving one (representative) effect size [22], with a particular population (laypeople) and a specific (winner-takes-all) payoff function. Future work could explore how results vary for different scenarios (e.g., real-world decisions with high stakes) or for different underlying effect sizes. It is also possible that the relative benefits of communicating outcome variability and standardized effect sizes are smaller for expert populations compared to laypeople, although past work suggests there may still be benefits to explicit presentations of outcome variability [35]. Furthermore, there are of course scenarios for which raw (non-standardized) effect sizes are appropriate, and outcome variability is less important. Finally, we have not looked at the issue of communicating inferential uncertainty about standardized effect sizes themselves—for instance, it would be interesting to think about how to convey that there is uncertainty in an analogy used to communicate probability of superiority itself. We see all of these directions as possibilities for future work.

In sum, in this paper, we present a series of experiments that investigate how to effectively communicate the effectiveness of a treatment. We find that for typical effect sizes in behavioral research, simply showing averages can lead people to overestimate treatment effectiveness, and this can be exacerbated by presenting information pertaining to statistical significance. Then we investigate how these initial perceptions change after people are shown information about variability in individual outcomes. We find that such information substantially reduces potential misperceptions. One applied recommendation stemming from this work would be to employ simple statements that describe how often one group outperforms another, or analogies grounded in other familiar phenomena, to help readers contextualize the effectiveness of treatments.

ACKNOWLEDGMENTS

We thank Hongtao Hao for helping us collect and annotate the CHI award-winning paper collection.

REFERENCES

[1] [n.d.]. Transparent Statistics in HCI. https://transparentstatistics.org. Accessed: 2020-09-14.

[2] American Psychological Association et al. 1996. Task force on statistical inference initial report. *Washington, DC: American Psychological Association PsycNET* (1996).

[3] American Psychological Association et al. 2010. Publication manual of the American psychological association Washington. *DC: American Psychological Association* (2010).

[4] Thom Baguley. 2009. Standardized or simple effect size: What should be reported? *British journal of psychology* 100, 3 (2009), 603–617.

[5] Thomas Baguley. 2012. *Serious stats: A guide to advanced statistics for the behavioral sciences*. Macmillan International Higher Education.

[6] Mohan V Bala, Josephine A Mauskopf, and Lisa L Wood. 1999. Willingness to pay as a measure of health benefits. *Pharmacoeconomics* 15, 1 (1999), 9–18.

[7] Pablo J Barrio, Daniel G Goldstein, and Jake M Hofman. 2016. Improving comprehension of numbers in the news. In *Proceedings of the 2016 chi conference on human factors in computing systems*. ACM, 2729–2739.

[8] Melanie L Bell, Mallorie H Fiero, Haryana M Dhillon, Victoria J Bray, and Janette L Vardy. 2017. Statistical controversies in cancer research: using standardized effect size graphs to enhance interpretability of cancer-related clinical trials with patient-reported outcomes. *Annals of Oncology* 28, 8 (2017), 1730–1733.

[9] Joseph Berkson. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Assoc.* 33, 203 (1938), 526–536.

[10] Lonni Besançon and Pierre Dragicevic. 2019. The continued prevalence of dichotomous inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[11] Google BigQuery. 2019. Global Surface Summary of the Day Weather Data. https://console.cloud.google.com/bigquery?project=api-project-821773148614&folder&organizationId&p=bigquery-public-data&d=noaa_gsod&page=dataset

[12] Frank A Bosco, Herman Aguinis, Kulraj Singh, James G Field, and Charles A Pierce. 2015. Correlational effect size benchmarks. *Journal of Applied Psychology* 100, 2 (2015), 431.

[13] 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.

[14] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 5 (2013), 365.

[15] Paul Cairns. 2007. HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*. British Computer Society, 195–201.

[16] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2, 9 (2018), 637.

[17] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. Hark no more: on the preregistration of chi experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 141.

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

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

[20] Jacob Cohen. 1992. A power primer. *Psychological bulletin* 112, 1 (1992), 155.

[21] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

[22] Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[23] Geoff Cumming, Fiona Fidler, Martine Leonard, Pavel Kalinowski, Ashton Christiansen, Anita Kleinig, Jessica Lo, Natalie McMenamin, and Sarah Wilson. 2007. Statistical reform in psychology: Is anything changing? *Psychological science* 18, 3 (2007), 230–232.

[24] Jasmine M DeJesus, Maureen A Callanan, Graciela Solis, and Susan A Gelman. 2019. Generic language in scientific communication. *Proceedings of the National Academy of Sciences* 116, 37 (2019), 18370–18377.

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

[26] Pierre Dragicevic. 2018. Can we call mean differences "effect sizes"? https://transparentstatistics.org/2018/07/05/meanings-effect-size/

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

[28] Mark D Dunlop and Mark Baillie. 2009. Paper rejected (p> 0.05): an introduction to the debate on appropriateness of null-hypothesis testing. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 1, 3 (2009), 86–93.

[29] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E Mackay. 2019. Touchstone2: An Interactive Environment for Exploring Trade-offs in HCI Experiment Design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 217.

[30] Paul D Ellis. 2010. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

[31] National Center for Health Statistics. 2016. Anthropometric Reference Data for Children and Adults: United States, 2011 to 2014. https://www.cdc.gov/nchs/data/series/sr_03/sr03_039.pdf

[32] David C Funder and Daniel J Ozer. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science* 2, 2 (2019), 156–168.

[33] 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).

[34] Ulrich Hoffrage and Gerd Gigerenzer. 1998. Using natural frequencies to improve diagnostic inferences. *Academic medicine* 73, 5 (1998), 538–540.

[35] 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*. 1–12.

[36] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. 2018. Improving Comprehension of Measurements Using Concrete Re-Expression Strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 34.

[37] John E Hunter and Frank L Schmidt. 2004. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

[38] Transparent Statistics in Human-Computer Interaction Working Group. 2019. Transparent Statistics Guidelines. https://doi.org/10.5281/zenodo.2226616

[39] Bob Ives. 2003. Effect size use in studies of learning disabilities. *Journal of Learning Disabilities* 36, 6 (2003), 490–504.

[40] Eunice Jun, Maureen Daum, Jared Roesch, Sarah E Chasins, Emery D Berger, Rene Just, and Katharina Reinecke. 2019. Tea: A High-level Language and Runtime System for Automating Statistical Analysis. *arXiv preprint arXiv:1904.05387* (2019).

[41] Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1105–1114.

[42] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanotham. 2017. Moving transparent statistics forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 534–541.

[43] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4521–4532.

[44] Ken Kelley and Kristopher J Preacher. 2012. On effect size. *Psychological methods* 17, 2 (2012), 137.

[45] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating personalized spatial analogies for distances and areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 38–48.

[46] Roger E Kirk. 1996. Practical significance: A concept whose time has come. *Educational and psychological measurement* 56, 5 (1996), 746–759.

[47] Geoffrey R Loftus. 1996. Psychology will be a much better science when we change the way we analyze data. *Current directions in psychological science* 5, 6 (1996), 161–171.

[48] David T Lykken. 1968. Statistical significance in psychological research. *Psychological bulletin* 70, 3p1 (1968), 151.

[49] Wendy E Mackay, Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Yangzhou Du, Jean-Daniel Fekete, and Yves Guiard. 2007. Touchstone: exploratory design of experiments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1425–1434.

[50] Jean-Bernard Martens. 2019. Insights in Experimental Data through Intuitive and Interactive Statistics. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, C09.

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

[52] Paul E Meehl. 1992. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. (1992).

[53] Ted A Paterson, PD Harms, Piers Steel, and Marcus Credé. 2016. An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership & Organizational Studies* 23, 1 (2016), 66–81.

[54] Jolynn Pek and David B Flora. 2018. Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological methods* 23, 2 (2018), 208.

[55] Christopher Riederer, Jake M Hofman, and Daniel G Goldstein. 2018. To put that in perspective: Generating analogies that make numbers easier to understand. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 548.

[56] Judy Robertson and Maurits Kaptein. 2016. *Modern statistical methods for HCI*. Springer.

[57] Robert Rosenthal and Donald B Rubin. 1982. A simple, general purpose display of magnitude of experimental effect. *Journal of educational psychology* 74, 2 (1982), 166.

[58] Patricia Snyder and Stephen Lawson. 1993. Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education* 61, 4 (1993), 334–349.

[59] 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.

[60] Radu-Daniel Vatavu and Jacob O Wobbrock. 2015. Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1325–1334.

[61] Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding novices in statistical analysis. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2693–2702.