Blinded with Science or Informed by Charts? A Replication Study

Pierre Dragicevic, Yvonne Jansen

▶ To cite this version:

Pierre Dragicevic, Yvonne Jansen. Blinded with Science or Informed by Charts? A Replication Study. IEEE Transactions on Visualization and Computer Graphics, 2018, 24 (1), pp.781-790. 10.1109/TVCG.2017.2744298 . hal-01580259

HAL Id: hal-01580259
https://hal.sorbonne-universite.fr/hal-01580259v1

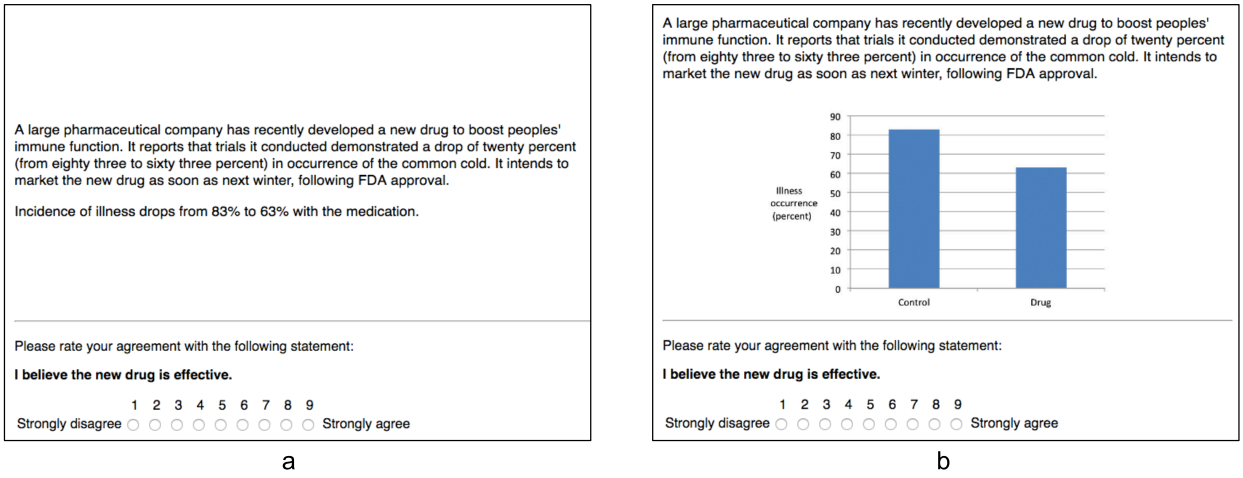
Submitted on 1 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Blinded with Science or Informed by Charts? A Replication Study

Pierre Dragicevic and Yvonne Jansen



A large pharmaceutical company has recently developed a new drug to boost peoples' immune function. It reports that trials it conducted demonstrated a drop of twenty percent (from eighty three to sixty three percent) in occurrence of the common cold. It intends to market the new drug as soon as next winter, following FDA approval.

Incidence of illness drops from 83% to 63% with the medication.

Please rate your agreement with the following statement:

I believe the new drug is effective.

Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 Strongly agree

A large pharmaceutical company has recently developed a new drug to boost peoples' immune function. It reports that trials it conducted demonstrated a drop of twenty percent (from eighty three to sixty three percent) in occurrence of the common cold. It intends to market the new drug as soon as next winter, following FDA approval.

Please rate your agreement with the following statement:

I believe the new drug is effective.

Strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 Strongly agree

Fig. 1. First page of our second experiment, replicating experiment 2 from Tal and Wansink [49]. *(a)* no-chart condition, with an extra sentence repeating the two quantities with numerals; *(b)* chart condition: the extra sentence is replaced with a bar chart.

Abstract—We provide a reappraisal of Tal and Wansink's study *"Blinded with Science"*, where seemingly trivial charts were shown to increase belief in drug efficacy, presumably because charts are associated with science. Through a series of four replications conducted on two crowdsourcing platforms, we investigate an alternative explanation, namely, that the charts allowed participants to better assess the drug's efficacy. Considered together, our experiments suggest that the chart seems to have indeed promoted understanding, although the effect is likely very small. Meanwhile, we were unable to replicate the original study's findings, as text with chart appeared to be no more persuasive – and sometimes less persuasive – than text alone. This suggests that the effect may not be as robust as claimed and may need specific conditions to be reproduced. Regardless, within our experimental settings and considering our study as a whole (N = 623), the chart's contribution to understanding was clearly larger than its contribution to persuasion.

Index Terms—Replication study, persuasion, charts, data comprehension, methodology.

1 Introduction

In 2014, Tal and Wansink [49] published a study entitled *"Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy"* in the journal *Public Understanding of Science*. The study shows that adding a chart to a statement about a new drug increases people's belief in the drug's efficacy. Fig. 1a shows the stimuli and question from the second experiment: when a chart was included, more people thought the drug was effective. The authors argued that the chart was redundant but persuasive because of its scientific aura. The article concludes by advising *"caution when encountering communications hinting at scientific credibility"*, and advising consumers to *"ignore spurious cues to a scientific basis"*. These findings were publicized by the media and are now cited in the scientific literature, including in information visualization [9, 40, 46].

The present article is a reappraisal of this study, which from now on we refer to as the *BwS study* (as in "Blinded with Science"). Our goal is to examine an alternative explanation for the results obtained in the BwS study, namely, that the charts allowed participants to better grasp the magnitude of the stated drug effect. Though the charts were informationally redundant, it is well known in information visualization that quantitative facts are often better understood if presented visually [4, 17, 37]. In the BwS study, the data only consisted of two quantities so the added value of the chart is unclear. Nevertheless, for the BwS study's conclusions to hold, it is important to ascertain that the chart did not give participants a better sense of the drug's efficacy.

The BwS study did attempt to establish that the chart had no positive effect on understanding, but as we will further discuss, the evidence is insufficient for several reasons: (1) the claim is based on accepting the null hypothesis with a single low-powered experiment, (2) the question did not test understanding but instead tested the retention of a specific number provided in the text (the percent reduction in illness), and (3) this number is not the most intuitive way of thinking of a drug's efficacy.

We conducted a series of four replications of the BwS study, which kept the stimuli and questions unchanged, and assessed people's understanding of the data provided about the drug. Considering our study as a whole, we found the chart's contribution to understanding to be small, but clearly larger than its contribution to persuasion. We conclude by discussing implications for persuasion evaluation methodology.

2 Background

Charts have been considered both as a means to influence people, and as a means to educate. We review previous work from both perspectives. We then specifically discuss what the literature has to say about the informativeness of minimalistic charts such as in Fig. 1.

- *Pierre Dragicevic is with Inria. E-mail: pierre.dragicevic@inria.fr.*
- *Yvonne Jansen is with Sorbonne Universités, UPMC Univ Paris 6, CNRS, ISIR. E-mail: jansen@isir.upmc.fr.*

Manuscript received 31 March 2017; accepted 19 July 2017. Date of Publication xx xxx. 201x; date of current version 27 June 2017.

This is an author version. The definitive version of record is published in IEEE Transactions on Visualization and Computer Graphics 24(1), 2018. DOI 10.1109/TVCG.2017.2744298

2.1 Using Charts to Influence

Persuasion can occur either through reason (i.e., by examining factual evidence and logical arguments) or under the influence of extraneous cues. These two distinct mechanisms are often referred to as the *"central route"* and the *"peripheral route"* to persuasion [39]. Many studies have confirmed that the peripheral route offers a variety of opportunities for deception, especially through the use of gratuitous scientific cues. Nonsense math can raise the perceived quality of research abstracts [16]. Irrelevant neuroscience information can make scientific explanations and articles appear more satisfying, stronger, and more convincing [41,55]. Gratuitous scientific jargon can increase the persuasiveness of messages promoting unproven medical remedies [24].

Although the BwS study takes its inspiration from this body of work, to our knowledge the association between charts and science has not been formally established, and evidence that charts can influence people through the peripheral route is scant. Admittedly, people use charts not only to inform, but also to create impressions [50]. Charts are often crafted to convey a chosen message, and are occasionally manipulated to conceal or distort data [26,50]. Though studies have confirmed the effectiveness of some of these manipulations [40], it is currently unclear whether plain, undistorted charts such as the bar chart in Fig. 1b can influence people's beliefs by their mere presence.

One study in information visualization has provided solid evidence for the persuasive power of charts in some contexts [39]. The authors however do not jump to conclusions as to the route through which persuasion occurred and call for more work in this area: *"we do not know if the more persuasive effect of charts over tables [...] is mostly due to having more information available or just because the medium itself (its visual appearance) is more persuasive."* [39]

Our paper's focus is not on testing the general hypothesis that charts can persuade through the peripheral route. Our goal is instead to determine whether the results from the BwS study admit an alternative explanation (i.e., the charts increased comprehension). If this were to be the case, it would not prove that persuasion through the peripheral route did not take place concurrently, nor would it prove that it cannot occur in other experimental setups. It would, however, have implications on how to better evaluate persuasion with charts in the future.

2.2 Using Charts to Inform

In educational psychology and cognitive science, it has long been established that pictures and diagrams can promote knowledge acquisition compared to text alone [5,23,43]. It has been further stressed that two *informationally equivalent* representations may not be *computationally equivalent*, i.e., information extraction can be more difficult with one than the other [32,43]. This distinction is crucial when studying persuasion with charts: in the BwS study, the no-chart and the chart conditions are *informationally equivalent*, but the BwS study has not formally established that they are also *computationally equivalent*.

Although the benefits of charts and visualizations are supported by a wealth of evidence accumulated in disciplines such as statistical graphics and information visualization, surprisingly little empirical data is available to determine whether the chart in Fig. 1b is useful.

The baseline condition in the BwS experiments has been referred to as *prose*[1], defined by MacDonald as *"ordinary language in written or printed form [which] may contain numerical data"* [33]. A few early studies have compared charts with prose in how well they convey quantitative facts. In 1927, Washburne [54] found prose to be inferior to charts for most purposes, concluding that *"it is a poor plan to present numerical data textually"*. In 1963, Feliciano [18] found that bar charts outperformed prose, and one year later Wilcox [56] found that prose supplemented with a bar chart was preferable to prose alone.

A much larger literature on tables vs. charts provides a more nuanced picture. Tables differ from (and are generally superior to [18,54,56]) prose due to their structured layout, but both have in common the use of numerals. For several decades, numerous studies comparing tables and charts have been conducted, with conflicting results [27,36,52]. Charts had many skeptics — for example, in a 1984 review, Desanctis [11] concludes that *"preliminary evidence suggests that a picture may not be worth a thousand words-or even a thousand numbers"*. Today, the consensus is that tables are best when exact individual quantities need to be extracted, while charts are generally superior for estimation and approximate comprehension, as well as comparisons and judgments of relationships within data [36,44,53]. If these results are translated to prose, it follows that conveying data through prose (as opposed to charts) must be a poor idea unless exact values are needed.

2.3 Can a Two-Bar Chart be Useful?

The previous findings cannot be easily applied to the chart in Fig. 1, because the studies involved datasets and charts that are substantially more complex. Authors sometimes expect their findings to generalize to smaller datasets, but are often evasive about where the limit is. While Washburne [54] warns against using prose *"if there are more than one or two items to be presented"*, Feliciano [18] strongly recommends against prose *"if more than a very few facts are to be presented"*.

Today, many visualization experts recommend against using charts for showing very few quantities. Tufte [51] deplores trivial charts and affirms that *"tables are preferable to graphics for many small data sets"*. For Duklan and Martin [15], *"it makes no sense to encode only a few numbers into an overblown graphic"*. For Kelly et al [29], *"graphs take up a lot of space if showing only a few data points. Hence they are best not used if there are only a few numbers to present."*. Gillan [20] similarly suggests that *"few data points might best be presented in the body of the text"* due to the cost of *"processing the data display and integrating the information from the display and text"*. Those are however only intuitions, without empirical evidence to support them.

Two studies come close to comparing prose with two-bar charts for estimating the difference between two quantities. Recently, Kim and Lombardino [30] presented participants with short statements involving three quantities (e.g., '*'the boy has three birds, five turtles, and one dog"*), or with equivalent three-bar charts. With questions involving ordinal comparison such as *"does the boy have more dogs than turtles?"*, participants were substantially faster with bar charts. Earlier on, Spence [45] compared how accurately and quickly people could estimate the relative difference between two quantities presented in various ways, including as a two-number table and as a two-bar chart. Both formats were about equally accurate, with possibly a slight advantage for the table, but the bar chart was clearly faster. Spence concludes that *"tables are preferable only if the audience is able to devote sufficient time and energy to their interpretation. With casual readers, who are less likely to linger, graphs may be superior to tables"*.

Participants to the BwS study could be considered "casual readers" because their task did not require them to carefully examine the data. However, both Spence's [45] and Kim and Lombardino's [30] studies involve repeated trials, where people may have been trained in rapidly extracting information from charts. In real settings and in the BwS study, even a simple graph such as in Fig. 1b needs to be parsed, which may incur extra costs [20,30], possibly cancelling the chart's benefits.

To summarize, despite the wealth of studies on charts, it is not at all clear whether the charts used in the BwS study are likely to have helped participants understand the information provided, or whether they were "trivial" as claimed by the authors.

3 Experiment 1 – First Replication and Rationale

The BwS study consists of three experiments. The first two test whether the addition of a simple bar chart can affect people's belief in medication efficacy. The third experiment tests whether the addition of a chemical formula ($C_{21}H_{29}FO_5$) produces a similar effect (with also positive results). Since our focus is on charts, we do not consider the third experiment in this article.

All the experiments we conducted are reported in full and summarized in Table 1. The experimental material (stimuli, data and R code) is available at http://www.aviz.fr/blinded.

Our first experiment is a partial [25] replication of the *first experiment* in the BwS study. The manipulations and the dependent variables were kept the same, and an additional dependent variable was collected that tested participants' understanding of the data.

[1] Also called *narrative* [11,56], *verbal format* [22], or simply *text* [18,56].

		Our experiment		Replicated BwS experiment
Stimulus data	#1	87% → 47% (40% drop)	#1	87% → 47% (40% drop)
Stimulus formats		Prose vs. Prose+Chart		Prose vs. Prose+Chart
Comprehens. test		Transfer task v.1 (Fig. 3)		None
Population		CrowdFlower		MTurk (US)
Stimulus data	#2	83% → 63% (20% drop)	#2	83% → 63% (20% drop)
Stimulus formats		Prose+Num. vs. Prose+Chart		Prose+Num. vs. Prose+Chart
Comprehens. test		Transfer task v.1 (Fig. 3)		Recall task
Population		CrowdFlower		Students (US)
Stimulus data	#3	83% → 63% (20% drop)	#2	See above
Stimulus formats		Prose+Num. vs. Prose+Chart		
Comprehens. test		Transfer task v.2 (Fig. 11)		
Population		CrowdFlower		
Stimulus data	#4	83% → 63% (20% drop)	#2	See above
Stimulus formats		Prose+Num. vs. Prose+Chart		
Comprehens. test		Transfer task v.2 (Fig. 11)		
Population		MTurk (US)		

Table 1. Summary table of the main characteristic of our four experiments and of the two original BwS experiments.

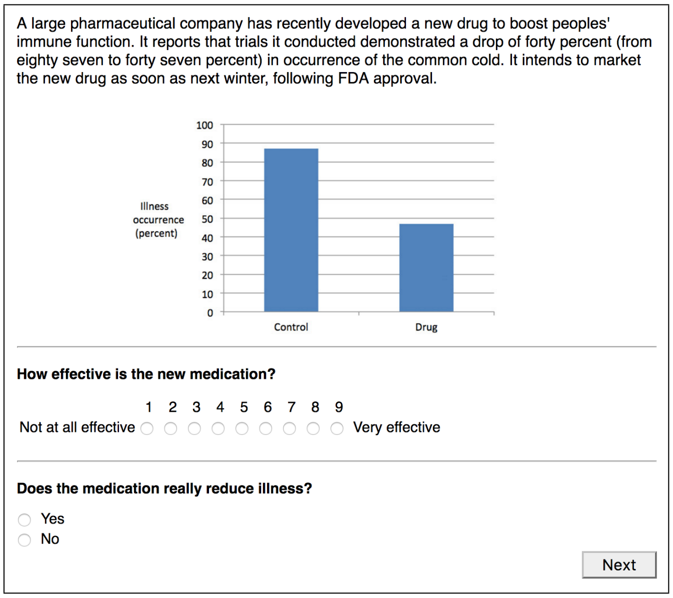
3.1 Replicated Stimuli and Questions

The replicated stimuli and questions are shown in Fig. 2. In contrast to the second BwS experiment mentioned in our introduction, the control condition in the first BwS experiment did not include an extra line repeating the two quantities with numerals (as in Fig. 1a). Thus, the two conditions only differed by having or not having a bar chart. This choice of control condition may have given an advantage to the bar chart, since all numbers in the text are fully spelled out and thus may require unnecessary effort to be parsed. We chose to replicate this first experiment nonetheless, because the authors present the two experiments as having comparable evidential strength [48, 49].

As can be seen in Fig. 2, two questions were displayed below the stimulus to assess participants' belief in the drug. The page for the control (no-chart) condition was identical in all respects except it was missing the bar chart.The paragraph of text was vertically centered as in Fig. 1a, so that both conditions occupy the same window size. The BwS paper does not report the layout of the stimuli used.

3.2 Task Framing

The BwS paper does not provide information on task framing and instructions. Their first experiment was conducted as *"part of a longer session containing multiple unrelated studies"* [49], but no information was provided about the other studies, and on the stated purpose of the entire session. Thus it seems difficult to reproduce the original context of the experiment. Nevertheless, the authors present their findings as general findings that can be applied to other situations such as court trials [48]. Thus, we used a task framing that is as minimal and as general as possible. We simply presented the experiment as a study on judgment, and told participants that they *"will be asked a few questions about [their] perception of medication effectiveness"*. As in previous studies on judgment and decision making [12, 35], the stimulus and the judgment questions were presented on the same page (Fig. 2). Thus the questions acted as hints on how to interpret and process the stimulus.


A large pharmaceutical company has recently developed a new drug to boost peoples' immune function. It reports that trials it conducted demonstrated a drop of forty percent (from eighty seven to forty seven percent) in occurrence of the common cold. It intends to market the new drug as soon as next winter, following FDA approval.

How effective is the new medication?

Not at all effective 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ Very effective

Does the medication really reduce illness?

○ Yes
○ No

Fig. 2. First page of of experiment 1, chart condition.

3.3 Comprehension Test

The first experiment of the BwS study did not test comprehension. In the second BwS experiment, however, an extra test was administered to participants to rule out the possibility that charts helped them process the information. We first discuss this test and its limitations.

3.3.1 Test Used in the BwS Study

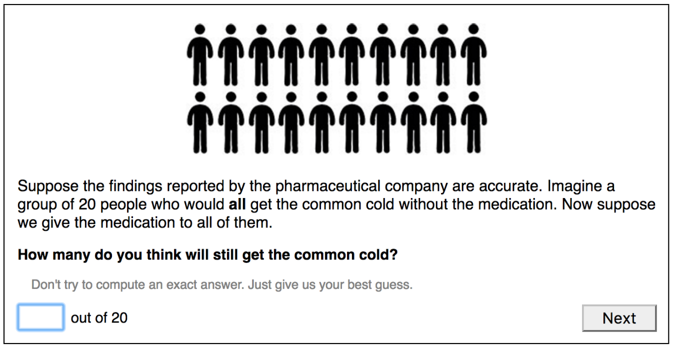
In the second BwS experiment, about 30 min after participants gave their answer, they were asked to report *"the percent by which the medication reduced illness"* (the experiment was also part of a longer session and the question was asked at the very end). The correct answer was 20% (see Fig. 1). No clear difference was found between the two conditions, leading the authors to conclude that *"the effects of graphs [...] is not moderated by increased understanding or retention of information"* [49]. This claim however lacks support for five reasons:

1. The BwS paper is ambiguous as to what the test is measuring. The experiment section accurately presents it as testing *retention* of information, but both retention *and understanding* are mentioned in the initial motivations and final interpretation of the results.
2. Retention in itself does not seem directly relevant to the question of what caused the chart to persuade participants: testing participants with a 30-min delay may not reflect what they understood *at the time* they indicated their degree of belief in the drug.
3. The question only requires participants to repeat a number. The text paragraph only includes three numbers, one of which is the right answer. The right answer is also the only round number (twenty), and likely the easiest to recall. Being able to recall this number from memory does not necessarily indicate understanding.
4. The number to be recalled is the reduction of illnesses in *percentage points*, a unit with which few people are familiar [19]. This unit can easily confuse, e.g., a reduction from 100% to 80% is the same in percentage points as a reduction from 20% to 0%. To add to the confusion, the wording used in the paragraph is often employed to indicate a *percent change*, which is a different unit[2] [13].
5. Independently from the above issues, the conclusion that the chart had no effect was based on accepting the null in a statistical significance test, without considering the uncertainty in the data. The 95% confidence interval we calculated for the difference between the proportions of correct answers is [-17%, 30%], which is wide.

3.3.2 Comprehension Test Used In This Experiment

Our goal is to test to what extent participants intuitively understand the magnitude of the drug's effect reported in the fictional study. Gigerenzer wrote extensively on how to express drug efficacy to a general audience in a way that promotes good decision making [19]. He recommends among other things to convey *absolute risks*. Both the text and the chart in Fig. 2 are doing this: the absolute risk of getting the cold is 87%, and becomes 47% with the drug. However, the text does not provide any information about the reference class, such as the type of population involved in the trials or the duration across which the absolute risks were measured. This makes it difficult to ask comprehension questions that require participants to apply knowledge rather than simply repeat numbers. Thus we chose to focus our test on the *relative risk reduction* (40%), which does not depend on the base incidence of the disease (87%). Relative risks are less preferred because they can exaggerate the benefits of an intervention when the base risk is low (e.g., halving the

[2] Expressed as a percent change, the 20% drop is a $\frac{63-83}{83} = 24.1\%$ drop.



Suppose the findings reported by the pharmaceutical company are accurate. Imagine a group of 20 people who would **all** get the common cold without the medication. Now suppose we give the medication to all of them.

How many do you think will still get the common cold?

Don't try to compute an exact answer. Just give us your best guess.

Fig. 3. Second page testing participants' understanding of the data.

risk could just mean going from 2 cases out of 10,000 to 1 case) [19]. However, in the BwS scenario the base incidence is high.

The test question used in the BwS study was also a question about relative risk reduction but as pointed out by Gigerenzer [19], percentage points are not the most intuitive way of thinking of a reduction. Percent changes are almost equally confusing [13]. *Ratios* such as in "halving the risk" are easier to grasp. In Fig. 2, the ratio of people who get sick with vs. without the drug is $47/87 \approx 0.54$, which translates into a meaningful probability. If we make the fair assumption that the drug never causes anyone who would *not* have gotten sick to get sick, then the probability of anyone getting sick with the drug, conditional on being in the group of people who would have gotten sick otherwise, is $p \approx 0.54$. However, Gigerenzer has shown that people often do not understand probabilities, and recommends using frequencies instead [19]. For example, if we consider a group of 1000 people whom we know will get sick and we give them the drug, we should expect about 540 of them to get sick. We chose 20 as the denominator, because it is easier to picture a small group and 20 yields a sufficient precision.

Our comprehension test is shown in Fig. 3 and was presented just after the first page (Fig. 2). The text asks participants to assume that the previously reported findings are *accurate*. This being a data comprehension question, if a participant understood the data but has extraneous reasons to doubt its reliability, it is important that their answer is not impacted. We additionally asked participants to give an estimate rather than trying to compute an exact answer. Our education system trains people to expect questions to admit a single correct answer, which can sometimes cause them to underperform in estimation tasks due to miscalculation [35]. Besides, we wanted to make it clear that this question was not an attention check that would have caused the rejection of the participant's contributions if answered incorrectly. To further prime participants into performing an estimation, we inserted a simple image to help them picture a group of 20 people in their mind's eye.

3.3.3 Error Metric

Dichotomizing answers into correct and incorrect wastes information and yields low statistical power [35,47]. Thus we assess the correctness of participants' answers using a continuous error metric.

We interpret participant answers as probabilities. For example, giving "5 out of 20" as the answer is the same as stating that the probability of any particular person getting sick is $p = 0.25$. We look at how far this probability is from the true probability $p = \frac{47}{87} \approx 0.54$. Absolute difference is not a good distance metric for probabilities since, for example, it considers that the separation between $p = 0.4$ and $p = 0.5$ is similar to that between $p = 0.1$ and $p = 0.0001$ [35]. Therefore we convert probabilities into log-odds before computing errors. The log-odds (or logit) function $p \mapsto \log(\frac{p}{1-p})$ projects probabilities onto the real line in a way that magnifies distances between probabilities near 0 and 1. In short, we use as our error metric the absolute difference between the implicitly stated log-odds of getting sick and the true log-odds. Since the log-odds is undefined for 0 and 1, we substitute the answers 0 and 20 with the values 0.1 and 19.9. Fig. 4 shows how answers map to errors. The best answer, 11, yields an error of ≈ 0.039.

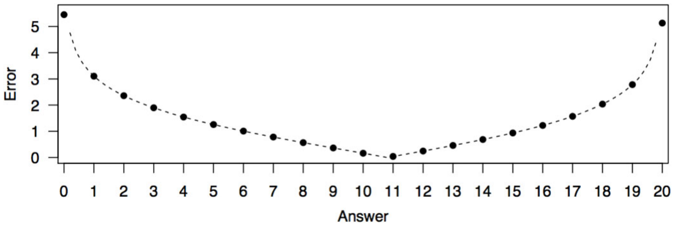


Fig. 4. Error for each possible response to the question in Fig. 3.

3.4 Additional Data Collected

We asked additional questions besides the three questions mentioned before. The first was shown only to participants who responded "No" to the second question in Fig. 2, and invited them to briefly justify their answer. The reason is that a "No" seems at odds with the information given, and yet in the BwS experiment the rate of "No" answers was quite high in the no-chart condition (10 out of 31 vs. 1 out of 29 with the chart). If we were to replicate this result, we wanted to know more about what motivated these answers. The second extra question asked participants whether they have heard or participated in a very similar study before.

3.5 Procedure

We ran the experiment through the crowdsourcing platform CrowdFlower (the original experiment used Amazon Mechanical Turk). The job was titled "Research study on judgment" and as discussed in Sect. 3.2, the job description stated that contributors will be asked a few questions about their perception of medication effectiveness.

Upon accepting the job, a new window opened showing an externally-hosted five-page web form. The form prevented contributors from reviewing previous pages. Page 1 consisted of the stimuli and questions from the BwS experiment (Fig. 2). Page 2 was the comprehension test (Fig. 3). Page 3 contained the additional questions mentioned in Sect. 3.4. Page 4 contained an attention check (explained next) and an optional text field for leaving comments. Page 5 displayed a thank you message with a job completion code when the job was accepted, or an explanation of why the job was rejected.

Contributors were offered a reward of 12¢ for an estimated completion time of one minute. The actual median completion time was 2.4 minutes. Job batches were posted on the CrowdFlower platform until the number of valid jobs approximately reached the target sample size.

3.5.1 Crowdsourcing Quality Control

Jobs were open to contributors with a performance level of 3 (the highest on CrowdFlower). After the tasks were completed, an attention check question asked which disease was mentioned in the study, with six possible answers including the common cold. The job description previously informed contributors that the job may include one or more attention tests. A job was rejected and not analyzed if:

- The job completion code was incorrect or already used;
- The answer to the attention check question was incorrect (5% of all completed jobs);
- Total job completion time was abnormally low or high, i.e., less than 30 seconds or more than 15 minutes (3% of jobs).

As crowdsourcing subject pools are becoming increasingly familiar with scientific studies [6], a job was accepted but discarded from our analyses if the contributor reported having heard of or participated in a very similar study in the past (12% of jobs).

The above job rejection rules were decided prior to running the experiment. All jobs that were not rejected according to these rules were considered valid and were analyzed.

3.6 Design and Research Questions

The between-subjects independent variable was *condition* $\in$ {no-chart, chart}. The three dependent variables were:

- *perceived effectiveness* $\in$ [1..9], which is the answer to the first question in Fig. 2,

- *belief in efficacy* $\in$ {yes, no}, which is the answer to the second question on that same page, and
- *comprehension error* $\in$ [0.039, 5.45], i.e., the error (Fig. 4) of the answer to the question on the second page (Fig. 3).

Our goal was to study comprehension, rather than to test whether the previous results from the BwS experiment replicate. Still, since we closely replicated the manipulations and dependent variables from the BwS study but could not ensure that we replicated the original experiment in all of its details, we had the opportunity to verify whether the results hold in a partial replication. In summary, our questions were:

Q1. Will the results from the BwS study replicate?
Q2. Will the chart yield improved comprehension?

These questions were formulated prior to conducting the experiment. Our expectations were that the results would replicate (i.e., perceived effectiveness and belief in efficacy both higher in the chart condition), but that the chart would also yield lower comprehension error.

3.7 Participants

Our planned sample size was $N = 120$ (the BwS experiment had $N = 61$). We received a total of $N = 123$ valid jobs, 62 for the no-chart condition and 61 for the chart condition.

Participants were 35% female, with a mean age of 35. They were from 32 different countries covering Europe (55%), Americas (28%) and Asia (17%). When asked about their education, 48% reported a 4-year college education or more, 28% reported some college education, 22% reported high school, and 2% reported none of the above.

3.8 Planned Analysis

We report all our results using interval estimates conveyed graphically [14]. All the analyses reported here, including the definition of error from Sect. 3.3.3, were planned before the data was collected.

3.8.1 Perceived Effectiveness

Perceived effectiveness is the response to the first question "How effective is the new medication?" (from 1 to 9). Like the original BwS study, we use as point estimate the sample mean. Since the measure is bounded and therefore not normally distributed, we use as interval estimates 95% BCa bootstrap confidence intervals for individual means and for the difference between two independent means [31].

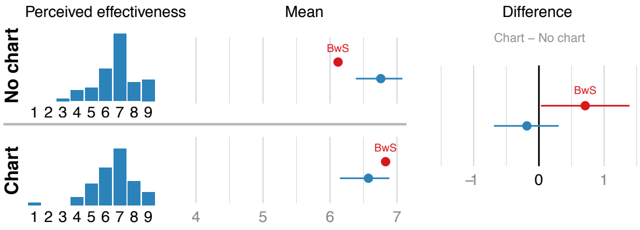


Fig. 5. Exp 1 – *Left:* responses to "how effective is the new medication?". *Right:* means and difference in means. Error bars are 95% CIs.

Raw responses are shown as histograms on the left side of Fig. 5. Although responses vary, most participants thought the drug was relatively effective, with 7 being the most common answer in both conditions. The mean response in each condition is reported in the middle of the figure: the two blue dots are the point estimates, and the two error bars are 95% confidence intervals. Thus there seems to be no evidence for a positive effect of the chart, as confirmed by the difference in means and its 95% CI, shown on the right side of Fig. 5. Point estimates from the original BwS experiment are shown in red for reference, with the 95% CI for the difference derived from the reported p-value [1].

3.8.2 Belief in Efficacy

Belief in efficacy is the response to the second question "Does the medication really reduce illness?". Like the original study, we use proportions of 'Yes' answers as point estimates. We use as interval estimates Wilson's confidence interval for a single proportion, and the score interval for difference of proportions and independent samples [57].

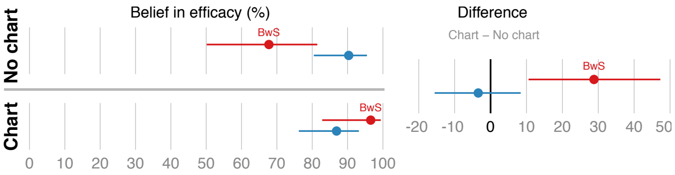


Fig. 6. Exp 1 – Percentages of "Yes" responses to the question "does the medication really reduce illness?". Error bars are 95% CIs.

As can be seen in Fig. 6, the vast majority of participants replied 'Yes' in both conditions (90% for no-chart and 87% for chart). Again, our data provides no evidence for a persuasive effect of the chart.

3.8.3 Comprehension Error

Comprehension error is the error of the response to the question "How many do you think will still get the common cold?". The response itself is between 0 and 20, the error is between 0.039 and 5.45. We use as point estimate the geometric mean instead of the arithmetic mean in order to reduce the influence of highly erroneous answers (which are more likely to be anomalous observations) and increase the weight of near-correct answers[3] [28]. As a result, differences between conditions will be expressed as error ratios, with 1 meaning no difference [42].

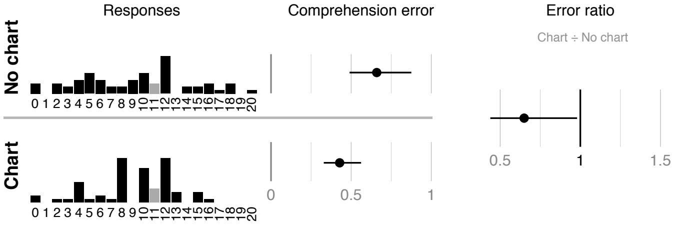


Fig. 7. Exp 1 – *Left:* responses to the comprehension question; *Right:* mean response errors and their ratio. Error bars are 95% CIs.

Raw responses (from 0 to 20) are shown as histograms on Fig. 7-left. Few participants gave the best possible answer (11) and the responses are widely spread. More responses however seem to cluster near 11 in the chart condition. This is supported by the interval estimates of the mean errors and of their ratio: the average error with chart was 0.65 times the average error without the chart, 95% CI [0.44, 0.98].

3.9 Additional Analyses

Though participants seemed to have a *lower understanding* of the drug's efficacy without the chart, we did not find evidence that they were *less convinced* of its efficacy. This may be due to the fact that misjudgments were both underestimations *and* overestimations of the drug's true efficacy, as suggested by the histograms in Fig. 7. To assess whether there was a general difference, we compared the mean of all responses (0–20) in the two conditions. The mean was 9.2, 95% CI [7.9, 10] in *no-chart*, and 9.1, 95% CI [8.1, 10] in *chart*. The difference was -0.06, 95% CI [-1.7, 1.4]. Thus there is no evidence for a substantial bias.

Concerning the justification question (see Sect. 3.4), we did not replicate the remarkably large proportion of 'No' answers in the no-chart condition, and thus we did not examine the data further.

3.10 Discussion

Contrary to our expectations, we were unable to replicate the results from the first experiment of the BwS study, which found a substantial effect of charts on the two metrics for belief in the drug. Although the effect may even appear reversed at first sight, the interval estimates are way too wide to support such a conclusion. The possible explanations

[3] Taking the geometric mean is the same as log-transforming all observations, taking the arithmetic mean, and then anti-logging it [42]. Logging the error function in Fig. 4 substantially reduces (but does not eliminate) the upward curvatures near 0 and 20, while introducing a sharp downward curvature near 11. Simulations conducted prior to the experiment suggested that this approach yields a slightly higher statistical power in the presence of uniform noise.

for this failure to replicate will be discussed later on. For now, our major focus is on whether the chart can promote comprehension.

Our comprehension test yielded many inaccurate answers in both conditions, but we did find evidence for a positive effect of charts overall. It is conceivable that the numbers provided in the text were hard to grasp, especially since they were fully spelled out. The two key quantities were made clearly visible by the bar chart, possibly helping (at least some) participants better appreciate the stated drug efficacy. At the same time, the improved understanding did not seem to cause participants to judge the drug as more effective on average, which may explain why it did not translate into a higher persuasion.

Despite the chart's seemingly positive effect on data comprehension, the choice to fully spell out all numbers in the text seems rather odd, and it is natural to ask whether a facilitating effect of the chart would have been observed had the numbers been provided as numerals instead. This question has been addressed to some extent in the second experiment of the BwS study, which we replicate next.

4 Experiment 2 – Second Replication

Our second experiment is a replication of the second BwS experiment, whose stimuli are shown in Fig. 1. In order to control for a possible repetition effect of the chart, the authors modified the no-chart condition to include an extra sentence that repeats the two quantities. Although the first paragraph of text shared by the two conditions still has its numbers spelled out, the no-chart condition now shows the numbers also as numerals, which puts it on a more equal footing with the chart.

4.1 Replicated Stimuli and Questions

We replicated the stimuli from the second BwS experiment (Fig. 1), which in addition to revising the no-chart condition used different numbers, yielding a smaller drug effect (the 40% drop becomes 20%).

The BwS paper is ambiguous as to whether the extra sentence in no-chart was inserted within the paragraph or added afterward. Since we could not find a spot in the paragraph where it would logically fit, we placed it underneath. This way, the experimental manipulation is also clear: the sentence is replaced by an informationally equivalent chart, aside from some differences in terminology.

The chart is not illustrated in the original BwS paper [49] but is shown in a subsequent publication [48]. Note in Fig. 1b that the y-axis stops at 90, which is Microsoft Excel's default. Although 100 may have been a better choice, we kept the original design to remain as close as possible to the original experiment. In the figure from [48], the y-axis is also missing a title, but it unclear whether it was also the case for the stimulus or if the figure was cropped. Since the title seems important to interpret the chart, we chose to include it, as in the first experiment.

The two questions from the first experiment were replaced with a single question. Experiment 1's first question was asking participants to estimate a quantity (the drug's effectiveness) rather than reporting a degree of belief. Thus this new question is closer to Experiment 1's second question, with the difference that it admits answers on a 1–9 scale instead of simply Yes/No answers.

4.2 Elements not Replicated

In the BwS study, the second experiment introduced three additional modifications: *(i)* it measured participants' degree of belief in science, with the finding that belief in science moderates the effect of the chart on persuasion. We chose not to include this extra dependent variable in order to keep our experiment design simple; *(ii)* it presented a chart with a non-zero origin to a subset of the chart group, but the difference was statistically non-significant. We did not include this variation in order to keep the experiment simple, and because the effectiveness of this manipulation has already been established [40]; *(iii)* it measured participant's retention of information, as already discussed in Sect. 3.3. We substitute this test with our own comprehension test.

4.3 Comprehension Test

The comprehension test was the same as in the previous experiment. The error function was updated to account for the new numbers (see Fig. 8). Now the best possible answer is 15, with an error of ≈ 0.049.

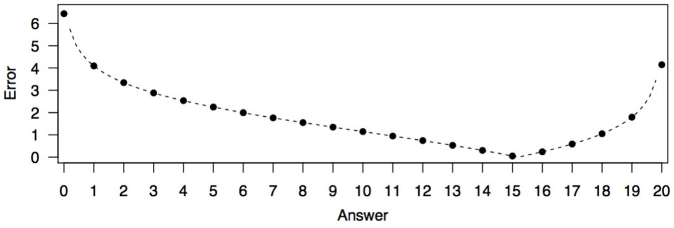


Fig. 8. Error for each possible response to the question in Fig. 3.

4.4 Additional Data Collected

We timed job completion time as before and kept the same extra questions, except for the question asking participants to justify their answer.

4.5 Procedure

We ran the experiment again on CrowdFlower (the second BwS experiment was a lab study with students) using the same procedure as before, with new contributors. Contributors were offered a reward of 20¢, and their median job completion time was 2.0 minutes. We used the same quality control procedure as before. Among all contributors who completed the job, 3% failed the attention check, 2% had an abnormal job completion time, and 10% reported having heard of or done a very similar study before. These results were not analyzed.

4.6 Design and Research Questions

As before, the between-subjects independent variable was *condition* $\in$ {no-chart, chart}. The two dependent variables were:

- *belief in efficacy* $\in$ [1..9], which is the answer to the question on the first page (Fig. 1),
- *comprehension error* $\in$ [0.049, 6.44], i.e., the error (Fig. 8) of the answer to the question on the second page (Fig. 3).

The research questions were the same as before.

4.7 Participants

Our planned sample size was $N = 160$ (the second BwS experiment had $N = 56$). We received a total of $N = 164$ valid jobs, 79 for the no-chart condition and 85 for the chart condition. Participant demographics was similar to the first experiment (see experimental material for details).

4.8 Planned Analysis

Belief in efficacy if the degree of agreement to the first question "I believe the new drug is effective" (from 1 to 9). We again report sample means and their 95% BCa bootstrap confidence intervals.

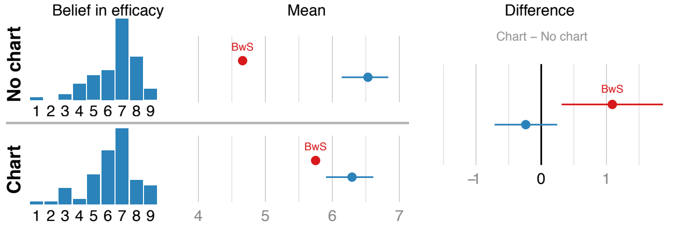


Fig. 9. Exp 2 – *Left:* degree of agreement to "I believe the new drug is effective". *Right:* means and difference in means. Error bars = 95% CIs.

Raw responses are shown as histograms on the left side of Fig. 9. The distributions are similar to the distributions of responses to the effectiveness estimation question from experiment 1 (Fig. 5). Again, there is no evidence of a positive effect of the chart. Results from the second BwS experiment are shown in red for reference. There is a particularly dramatic difference in the point estimates in the no-chart condition (4.7 in BwS and 6.5 for us on the 9-point scale).

Raw responses to the comprehension test are shown on Fig. 10. The accuracy of participants was rather poor: while the best answer was 15, responses were widely distributed between 0 and 20. This time, there is no evidence of a difference in mean errors: the mean error with chart was 0.91 times that without the chart, 95% CI [0.63, 1.35].

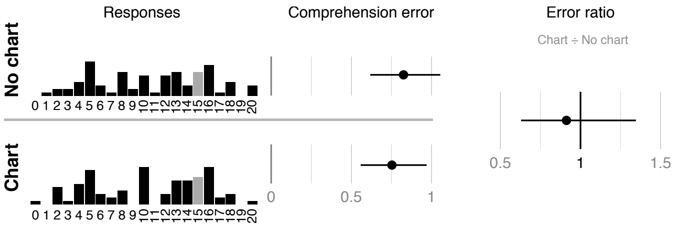


Fig. 10. EXP 2 – *Left:* responses to the comprehension question; *Right:* mean response errors and their ratio. Error bars are 95% CIs.

4.9 Discussion

We were still not able to replicate the original BwS findings in this second experiment. This time, though, we did not find evidence that the chart promoted comprehension either. It is possible that putting the no-chart condition on a more equal footing with the chart makes the effect disappear or become negligible. Alternatively, it may be that two-bar charts are less effective at showing the relative difference between two numbers when the difference is 20% as opposed to 40%. This seems consistent with Spence's study suggesting that people are more accurate at visually judging relative differences near 50%, 0% and 100% [45]. There is however a fair degree of overlap between the CI of the error ratio here and in Fig. 7, so the apparent difference between the two experiments may also be statistical noise.

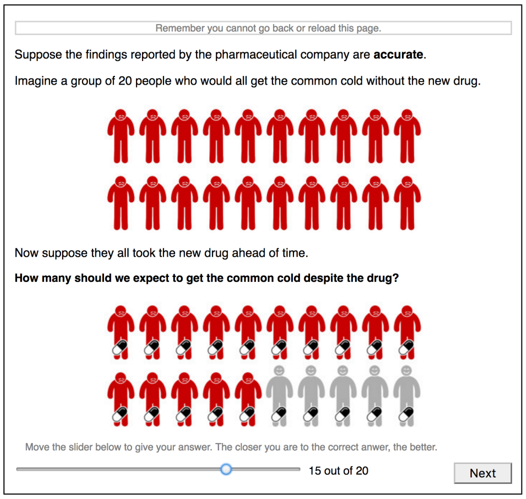
We were perplexed by the large proportion of incorrect answers and noticed a symmetry in their distribution, especially in the chart condition (Fig. 10): the distribution from 0 to 10 seems to mirror the distribution from 10 to 20. We reasoned that many participants may have inverted their answer. These possible inversions may be diagnostic of a data comprehension error, but could also result from a response error. In particular, it is possible that these participants entered the number of people who will *not* get sick instead of entering the number of people who will get sick (Fig. 3). In order to eliminate this possibility, we redesigned the comprehension test and ran another experiment.

5 EXPERIMENT 3 – REVISITED COMPREHENSION TEST

This third experiment is the same as the previous one (i.e., it uses the same replicated stimuli and questions from the second BwS experiment), but the comprehension test has been redesigned.

5.1 Modifications to Experiment 2

The redesigned comprehension test is shown on Fig. 11. The major difference is that the response is given by setting the number of sick people in an *icon array*, using a slider. Icon arrays are commonly used for communicating health risks to the public [2, 35]. The default value of the slider was set to 20.



Remember you cannot go back or reload this page.

Suppose the findings reported by the pharmaceutical company are **accurate**.

Imagine a group of 20 people who would all get the common cold without the new drug.

Now suppose they all took the new drug ahead of time.

How many should we expect to get the common cold despite the drug?

Move the slider below to give your answer. The closer you are to the correct anwer, the better.

15 out of 20

Fig. 11. Our redesigned comprehension test on page 2. Initially, all people at the bottom are selected and shown in red (20 out of 20).

A second attention check question was used, asking *"In the scenario described on the second page, how many people took the drug?"* (11% failed to answer one of the two questions). In addition, jobs were discarded if less than 8 seconds were spent on the first or the second page (1% of all submitted jobs). 4% reported doing a very similar study before. Demographic questions were removed. Contributors were rewarded 15¢ and the median job completion time was 2.4 minutes.

5.2 Participants

Our planned sample size was $N = 160$. We received a total of $N = 176$ valid jobs, 88 for the no-chart condition and 88 for the chart condition.

5.3 Planned Analysis

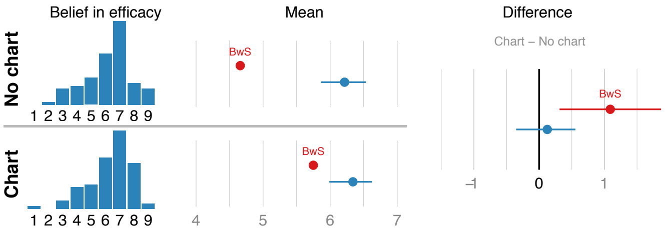


Fig. 12. EXP 3 – *Left:* belief in efficacy responses. *Right:* means and difference in means. Error bars are 95% CIs.

Responses to the statement "I believe the new drug is effective" (Fig. 12) are virtually indistinguishable from the previous experiment. There is again no evidence for a positive effect of chart.

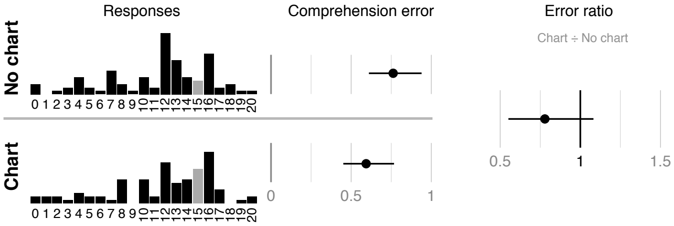


Fig. 13. EXP 3 – *Left:* responses to the comprehension question; *Right:* mean response errors and their ratio. Error bars are 95% CIs.

Raw responses to the comprehension test are shown in Fig. 13. The symmetry observed in Fig. 10 has almost disappeared and responses tend to cluster towards 12–16. There may be a positive effect of chart overall, but the evidence is relatively weak. The mean error with chart was 0.78 times the mean error without the chart, 95% CI [0.55, 1.08].

5.4 Additional Analyses

The mean response to the comprehension question was 11.3, 95% CI [10.2, 12.2] in *no-chart* and 11.8, CI [10.7, 12.7] in *chart*, with a difference of 0.48, CI [-0.91, 1.85]. Thus there is still no evidence for a substantial bias, even after addressing the issue of inverted responses.

5.5 Discussion

Our new comprehension test yielded distributions of responses closer to what we should expect, suggesting that the symmetry in experiment 2 was due to response errors rather than comprehension errors. We found weak evidence for a positive effect of the chart overall, but cannot at this point draw definitive conclusions concerning experiment 2 stimuli.

After three replications, we still could not find anything suggestive of a persuasive effect of charts. One clear difference between our experiments and the original experiments is in the population studied: the BwS study involved an American population (recruited through Amazon MTurk in the first experiment, college students in the second), while our population is multinational. Although our Crowdflower contributors were English speakers, many are likely not native speakers. There can also be cultural differences affecting how the stimuli are processed. For example, non-US residents may not necessarily understand what the FDA is. Therefore, we ran the experiment again, this time with US residents recruited through MTurk.

6 Experiment 4 – US Population

Our fourth and last experiment replicates experiment 3 on the Amazon MTurk platform. We also introduce two covariates in order to better understand what drives persuasion or the lack thereof.

6.1 Additional Data Collected

We re-introduced demographic questions and added two questions at the end of the experiment. The first one was *“Do you generally believe in science?”*, on a 9-point scale (1 = not at all, 9 = very much). The second BwS experiment included a similar question (see Sect. 4.2) and found that chart persuasiveness is stronger for people with a high belief in science, presumably because the chart signals science. Thus it is possible that responses from science skeptics diminished our effects.

The second question read: *“The drug we mentioned was fictional. Nevertheless, do you think that a large pharmaceutical company can design an effective drug for preventing the common cold?”*, with 1 = extremely unlikely and 9 = extremely likely. The question measured to what extent people’s core beliefs make them inclined to believe in the drug irrespective of the data, similarly to what Pandey et al [39] call *polarization*. It captures three causes of negative polarization we saw in the answers to the justification question from Sect. 3.4: general disbelief in medicine, mistrust of the pharmaceutical industry, and informed skepticism about the feasibility of such a drug. Pandey et al found charts to be more persuasive for neutral and positively polarized people, and tables to be more persuasive for negatively polarized people.

6.2 Procedure

Only MTurk contributors who reside in the US and have a job approval rate of 97% could participate. Our experiment, hosted on an external page, was the same as experiment 3 except for the changes mentioned above. Of all contributors, 3% failed the attention checks and none reported having completed a very similar study before. Contributors were rewarded 20¢ and the median completion time was 1.7 minutes.

6.3 Design and Research Questions

The independent and dependent variables were the same as in the previous experiment, but the design also included two covariates: *belief in science* $\in$ [1..9], and *polarization* $\in$ [1..9]. Besides the two research questions from the previous experiments, we wanted to determine whether belief in science and polarization predict belief in efficacy.

6.4 Participants

We received $N = 160$ valid jobs, 80 per condition, for the same planned sample size. Participants were 44% female, with a mean age of 37. 48% reported a 4-year college education or higher, 38% reported some college education, 14% reported high school. All were from the US.

6.5 Planned Analysis

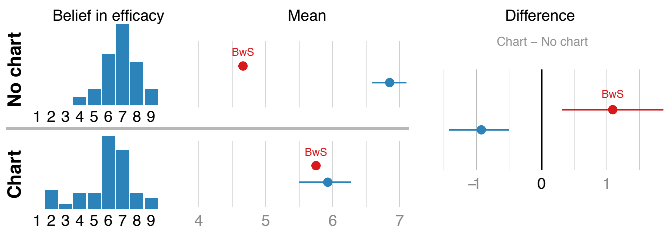


Fig. 14. Exp 4 – *Left:* belief in efficacy responses. *Right:* means and difference in means. Error bars are 95% CIs.

Results for belief in effectiveness are shown in Fig. 14. This time, there is strong evidence that the chart had a *negative* effect on persuasion. Responses to the comprehension question are shown in Fig. 15. Participants are more accurate than in previous experiments, but there is no evidence for a positive effect of chart overall. The mean error with chart was 0.97 times the mean error without, 95% CI [0.69, 1.42].

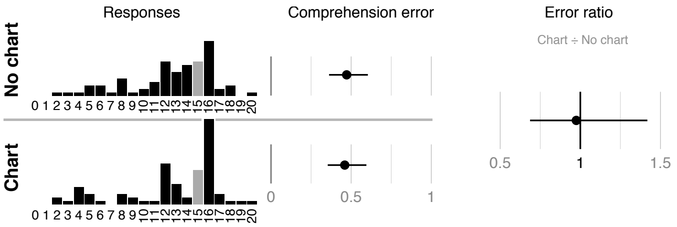


Fig. 15. Exp 4 – *Left:* responses to the comprehension question; *Right:* mean response errors and their ratio. Error bars are 95% CIs.

Concerning the effect of covariates on chart persuasiveness, although prior studies dichotomized [49] or trichotomized [39] the covariate of interest, we use linear regression for more statistical power. The 1st and 2nd plots in Fig. 16 suggest that while belief in science is overall high among our participants, belief in science predicts belief in efficacy in both conditions. However, the regression slopes are similar, which is inconsistent with the hypothesis that the difference between chart and no-chart increases with belief in science. The difference between the regression slopes is -0.05, 95% CI [-0.61, 0.37] (BCa bootstrap CI).

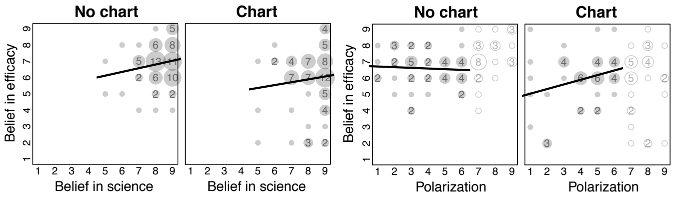


Fig. 16. Exp 4 – Categorical scatterplots with regression line for *belief in science* vs. *belief in efficacy* (left), and for *polarization* vs. *belief in efficacy* (right). Hollow dots are excluded from the regressions.

The 3rd and 4th plots in Fig. 16 show the results for the polarization covariate. Since Pandey et al [39] mostly emphasized the difference between negative and neutral polarization, values > 6 are excluded from the regression. The relationship between polarization and belief in efficacy seems to differ between the two conditions, with a difference in regression slope of 0.33, 95% CI [0.31, 0.75]. Thus, there is evidence for an interaction effect: consistent with Pandey et al [39], the relative persuasive power of the chart (compared to no-chart) is lower for negatively polarized participants than for neutral participants.

6.6 Discussion

In this experiment with US participants, we did not find a facilitating effect of charts. Even more surprising is the *negative* effect of charts on persuasion. The cause is likely not a low belief in science. Compared to our previous CrowdFlower population, participants differed not only in nationality, but also in their overall accuracy. An Mturk approval rate of 97% may be harder to attain than a CrowdFlower level of 3.

Although the effect of the polarization covariate is consistent with previous findings [39], the near-horizontal regression slope in the no-chart condition (3rd plot in Fig. 16) is suspicious: participants’ belief in the drug’s efficacy does not seem affected by their core beliefs overall. Why participants were more skeptical with the chart also calls for explanation. Thus we conducted a short follow-up survey.

6.7 Follow-up Survey

We contacted the 19 contributors for whom belief in efficacy was *i)* at least 5 points higher than polarization in the no-chart condition or *ii)* at least 3 points lower than polarization in the chart condition. Respondents were offered 10¢ plus a bonus of 80¢ to 150¢. We presented the stimulus again, reminded them of the discrepancy between their response to the polarization question and the first (belief in efficacy) question, and asked them to explain their response to the first question.

In the no-chart condition, 8 out of 9 contributors responded. Three of them stated being skeptical of the feasibility of such a drug, while 1 was skeptical of a drug that would completely prevent the common cold. Yet 6 respondents stated that the numbers were suggestive of an effective drug, including 4 who explicitly mentioned the 20% drop.

Two stated that the trial was “fictional” or “theoretical”. Thus it seems that at least some participants gave the drug a high rating because they tended to focus on the numbers irrespective of their prior beliefs, sometimes considering the question as purely hypothetical. Yet it does not explain why the chart condition exhibited different trends.

In the chart condition, 3 out of 10 contributors (who all gave a low belief score but thought such a drug was possible) responded. One respondent stated that technology can now cure anything, but large pharmaceutical companies have a vested interest in keeping people sick. The two other respondents stated that the reduction in illness was not sufficient for calling the drug effective, and that the difference could be due to small sample sizes or uncontrolled variables. Although no one explicitly mentioned the chart, perhaps the chart (with its bar labelled ‘Control’) reminded participants that the numbers came from a clinical trial, and should be therefore treated cautiously.

7 Meta-Analysis

Before concluding, we report a meta-analysis to better quantify the strength of statistical evidence in our four experiments. Meta-analyses make it possible to aggregate the results from multiple heterogeneous studies asking the same research question, and can be used to combine the results from multiple experiments within the same study [10].

Since there is lots of variability in the data, we report standardized effect sizes (Cohen’s d, i.e., the difference in means divided by the standard deviation [8]) in order to assess how large the effects are relative to individual differences. This measure being unitless, it also allows us to compare dependent variables expressed in different units.

We report effect sizes with their 95% BCa bootstrap CI for *belief in efficacy* and *comprehension error*. For experiment 1 where belief in efficacy is a binary measure, we use perceived effectiveness as a substitute. Consistently with our previous analyses (Sect. 3.8.3), comprehension errors are log-transformed. Aggregate effect sizes are obtained by performing a contrast weighted by sample size [31].

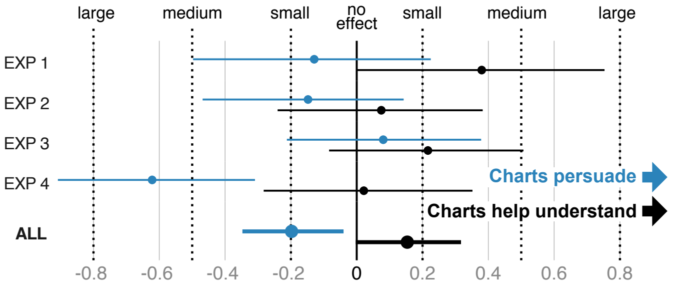


Fig. 17. Cohen’s d for the chart’s tendency to *increase* belief in efficacy (blue) and to *reduce* comprehension error (black). Error bars = 95% CIs.

Fig. 17 shows the effect sizes for each of our experiments, and for all our experiments considered together (at the bottom). For blue estimates, the higher the value the larger the chart’s contribution to *persuasion*. For black estimates, the larger the value the larger the chart’s contribution to *understanding*. The results of the meta-analysis suggest that overall, the chart’s contribution to understanding was most likely positive and its contribution to persuasion was very likely negative.

8 Conclusions

Our goal was to examine whether the results from the BwS study can be explained by a facilitating effect of charts on data comprehension. As we were unable to replicate these results, we cannot provide a definitive answer. A meta-analysis of our four replications does point towards a likely – although small – facilitating effect of the chart overall. This suggests that a similar effect could have occurred in the BwS study. It remains unclear whether the aggregate effect we observed is only due to the first experiment, or whether the effect generalizes when numerals are provided and drug efficacy is lower (20% instead of 40%).

We have no explanation for our inability to replicate the effects on persuasiveness. The no-chart condition seems to exhibit the largest discrepancies in the results, as it systematically elicits more persuasion in our study than in the original (see Figs 5,6,9,12,14). Our experiments were designed to be as close as possible to the original experiments for the replicated dependent variables (belief in drug efficacy). Since these variables were collected first, all modifications we did to the experiment (e.g., the comprehension question, or the extra questions in experiment 4) cannot have affected our results for the replicated variables. It also seems unlikely that the differences are due to the populations involved. Our experiment 4 involved US MTurk contributors like the first BwS experiment, but found an opposite effect. Our participants overall had a strong belief in science and levels of education similar to the BwS study. There may be other differences on aspects that were not fully described in the BwS paper. Regardless, our study suggests that the effect may not be as general as the original study claims and might require very specific conditions to be observed.

The main lesson from our study is that with charts, the peripheral route of persuasion cannot be studied independently from the central route: in order to establish that a chart biases judgment, it is necessary to also rigorously establish that it does not aid comprehension. Our study illustrates one way this can be done. Although it is impossible to statistically establish that a manipulation has no effect, we suggest a possible workaround — if a chart’s contribution to understanding is substantially lower than its contribution to persuasion in terms of standardized effect sizes, it seems reasonable to assume that some persuasion has taken place through the peripheral route. One difficulty is the method’s reliance on a particular comprehension test, which may be addressed by using a battery of comprehension tests.

Our comprehension test is not without limitations. Perhaps as with any comprehension test, there is no way to make sure that no additional comprehension took place during the test, *after* participants expressed their judgment about the drug’s efficacy. The test could have prompted participants to recall the chart from their visual memory, even if they did not pay attention to it previously. Similarly, participants could have recalled numbers and performed calculations. Why participants were so inaccurate also remains to be understood. Judgment of relative magnitudes can be inaccurate with charts [45], but it does not explain why some participants gave highly erroneous answers. Previous work suggests that removing all numbers may improve accuracy [35].

Since our focus was on the BwS study, our study was not designed to investigate how charts in general affect data comprehension and persuasion. There are many ways the experiments could be improved, e.g., by clarifying both the text and the chart, by using a consistent terminology between the two conditions, and by performing single manipulations to isolate the effects of different factors. In addition, the task we used may have been too abstract and too artificial to capture persuasion in the real world. There seems to be more promise in testing realistic tasks, e.g., asking people to assess real facts on meaningful social issues [39], or exposing them to ads in their personal environment [34]. Asking the right questions is also important, especially since stated beliefs do not necessarily reflect real intentions or behavior [38]. In any case, task context and instructions are likely to affect both how people report their beliefs and how they process information, thus they require particular attention and should be easily replicable.

In the future, it could be interesting to study if individual differences such as education level, visualization literacy [3] or cognitive style [7] can affect the results. The authors of the BwS study later reported they asked participants to identify themselves as “visual thinkers” or “verbal thinkers”, but could not find an effect [48]. Detecting such effects will likely require designing more specific, higher-powered experiments.

Finally, our replication opens many relevant questions for infovis. Are charts really associated with science? More generally, what associations do charts or visualizations trigger depending on their visual design? When exactly is a chart trivial? Two arguments against minimalistic charts is that they take up space and they break the flow of the text. How do word-scale visualizations [21] change these trade-offs?

Acknowledgments

We would like to thank Frédéric Vernier, Gilles Bailly, and Geoff Cumming for their feedback on early revisions of this paper.

References

[1] D. G. Altman and J. M. Bland. How to obtain the confidence interval from a p value. *BMJ*, 343:d2090, 2011.

[2] J. S. Ancker, Y. Senathirajah, R. Kukafka, and J. B. Starren. Design features of graphs in health risk communication: a systematic review. *JAMA-J AM MED ASSOC*, 13(6):608–618, 2006.

[3] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A principled way of assessing visualization literacy. *IEEE TVCG*, 20(12):1963–1972, 2014.

[4] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[5] R. N. Carney and J. R. Levin. Pictorial illustrations still improve students' learning from text. *Educational psychology review*, 14(1):5–26, 2002.

[6] J. Chandler, P. Mueller, and G. Paolacci. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46(1):112–130, 2014.

[7] T. L. Childers, M. J. Houston, and S. E. Heckler. Measurement of individual differences in visual versus verbal information processing. *Journal of Consumer Research*, 12(2):125–134, 1985.

[8] J. Cohen. Statistical power analysis for the behavioral sciences lawrence earlbaum associates. *Hillsdale, NJ*, pp. 20–26, 1988.

[9] M. Correll and M. Gleicher. Bad for data, good for the brain: Knowledge-first axioms for visualization design. In *IEEE VIS 2014 (workshop DECISIVE)*, 2014.

[10] G. Cumming. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2013.

[11] G. DeSanctis. Computer graphics as decision aids: Directions for research. *Decision Sciences*, 15(4):463–487, 1984.

[12] E. Dimara, A. Bezerianos, and P. Dragicevic. The attraction effect in information visualization. *IEEE TVCG*, 23(1):471–480, 2017.

[13] P. Dragicevic. My technique is 20% faster: Problems with reports of speed improvements in HCI. Research report, Oct. 2012.

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

[15] K. Duklan and M. A. Martin. *Communicating effectively with words, numbers and pictures: Drawing on experience*. School of Finance and Applied Statistics, Australian National University, 2002.

[16] K. Eriksson. The nonsense math effect. *Judgment and decision making*, 7(6):746, 2012.

[17] J.-D. Fekete, J. Van Wijk, J. Stasko, and C. North. The value of information visualization. *Information visualization*, pp. 1–18, 2008.

[18] G. D. Feliciano, R. D. Powers, and B. E. Kearl. The presentation of statistical information. *Educational Technology Research and Development*, 11(3):32–39, 1963.

[19] G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 8(2):53–96, 2007.

[20] D. J. Gillan, C. D. Wickens, J. G. Hollands, and C. M. Carswell. Guidelines for presenting quantitative data in hfes publications. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(1):28–41, 1998.

[21] P. Goffin, W. Willett, J.-D. Fekete, and P. Isenberg. Exploring the placement and design of word-scale visualizations. *IEEE TVCG*, 20(12):2291–2300, 2014.

[22] G. L. Grace. Application of empirical methods to computer-based system design. *Journal of Applied Psychology*, 50(6):442, 1966.

[23] S. Guri-Rozenblit. Impact of diagrams on recalling sequential elements in expository texts. *Reading Psychology: An International Quarterly*, 9(2):121–139, 1988.

[24] J. Haard, M. D. Slater, and M. Long. Scientese and ambiguous citations in the selling of unproven medical treatments. *Health communication*, 16(4):411–426, 2004.

[25] K. Hornbæk, S. S. Sander, J. A. Bargas-Avila, and J. Grue Simonsen. Is once enough?: on the extent and content of replications in human-computer interaction. In *Proceedings of CHI*, pp. 3523–3532. ACM, 2014.

[26] D. Huff. *How to lie with statistics*. WW Norton & Company, 2010.

[27] S.-L. Jarvenpaa and G. W. Dickson. Graphics and managerial decision making: Research-based guidelines. *Communications of the ACM*, 31(6):764–774, 1988.

[28] O. N. Keene. The log transformation is special. *Statistics in medicine*, 14(8):811–819, 1995.

[29] D. Kelly, J. Jasperse, and I. Westbrooke. Designing science graphs for data analysis and presentation. *Department of Conservation Technical Series*, 32, 2005.

[30] S. Kim and L. J. Lombardino. Comparing graphs and text: Effects of complexity and task. *Journal of Eye Movement Research*, 8(3), 2015.

[31] K. N. Kirby and D. Gerlanc. BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior research methods*, 45(4):905–927, 2013.

[32] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100, 1987.

[33] M. Macdonald-Ross. How numbers are shown. *AV Communication Review*, 25(4):359–409, 1977.

[34] A. Mehta. Advertising attitudes and advertising effectiveness. *Journal of advertising research*, 40(3):67–72, 2000.

[35] L. Micallef, P. Dragicevic, and J.-D. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE TVCG*, 18(12):2536–2545, 2012.

[36] J. Morgan and G. Michaelson. A comparative evaluation of tabular and graphical presentation styles for information retrieval search results, 2012.

[37] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[38] B. Nyhan and J. Reifler. Does correcting myths about the flu vaccine work? an experimental evaluation of the effects of corrective information. *Vaccine*, 33(3):459–464, 2015.

[39] A. V. Pandey, A. Manivannan, O. Nov, M. Satterthwaite, and E. Bertini. The persuasive power of data visualization. *IEEE TVCG*, 20(12):2211–2220, 2014.

[40] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini. How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In *Proceedings of CHI*, pp. 1469–1478. ACM, 2015.

[41] R. E. Rhodes, F. Rodriguez, and P. Shah. Explaining the alluring influence of neuroscience information on scientific reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5):1432, 2014.

[42] J. Sauro and J. R. Lewis. Average task times in usability tests: what to report? In *Proceedings of CHI*, pp. 2347–2350. ACM, 2010.

[43] W. Schnotz and M. Bannert. Construction and interference in learning from multiple representation. *Learning and instruction*, 13(2):141–156, 2003.

[44] M. Schonlau and E. Peters. Comprehension of graphs and tables depend on the task: empirical evidence from two web-based studies. *Statistics, Politics, and Policy*, 3(2), 2012.

[45] I. Spence. Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):683, 1990.

[46] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE TVCG*, 22(1):629–638, 2016.

[47] D. L. Streiner. Breaking up is hard to do: the heartbreak of dichotomizing continuous data. *The Canadian Journal of Psychiatry*, 47(3):262–266, 2002.

[48] A. Tal. Looks like science, must be true! graphs and the halo of scientific truth. *The Jury Expert*, 27(2):1–8, 2015.

[49] A. Tal and B. Wansink. Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science*, 25(1):117–125, 2016.

[50] N. Tractinsky and J. Meyer. Chartjunk or goldgraph? effects of presentation objectives and content desirability on information presentation. *MIS Quarterly*, pp. 397–420, 1999.

[51] E. R. Tufte. The visual display of quantitative information. *Journal for Healthcare Quality*, 7(3):15, 1985.

[52] I. Vessey. Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2):219–240, 1991.

[53] I. Vessey. The effect of information presentation on decision making: A cost-benefit analysis. *Information & Management*, 27(2):103–119, 1994.

[54] J. N. Washburne. An experimental study of various graphic, tabular, and textual methods of presenting quantitative material. *Journal of Educational Psychology*, 18(7):465, 1927.

[55] D. S. Weisberg, F. C. Keil, J. Goodstein, E. Rawson, and J. R. Gray. The seductive allure of neuroscience explanations. *Journal of cognitive neuroscience*, 20(3):470–477, 2008.

[56] W. Wilcox. Numbers and the news: Graph, table or text? *Journalism Quarterly*, 41(1):38–44, 1964.

[57] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.