Communicating Uncertainties about the Effects of Medical Interventions Using Different Display Formats

Michelle McDowell.[1,2] and Astrid Kause[3]
[1]Harding Center for Risk Literacy, Faculty of Health Sciences, University of Potsdam
[2]Max Planck Institute for Human Development
[3]Leeds University Business School

Abstract

Communicating uncertainties in scientific evidence is important to accurately reflect scientific knowledge, increase public understanding of uncertainty, and to signal transparency and honesty in reporting. While techniques have been developed to facilitate the communication of uncertainty, many have not been empirically tested, compared for communicating different types of uncertainty, or their effects on different cognitive, trust, and behavioral outcomes have not been evaluated. The present study examined how a point estimate, imprecise estimate, conflicting estimates, or a statement about the lack of evidence about treatment effects, influenced participant's responses to communications about medical evidence. For each type of uncertainty, we adapted three display formats to communicate the information: tables, bar graphs, and icon arrays. We compared participant's best estimates of treatment effects, as well as effects on recall, subjective evaluations (understandability and usefuleness), certainty perceptions, perceptions of trustworthiness of the information, and behavioral intentions. We did not find any detrimental effects from communicating imprecision or conflicting estimates relative to a point estimate across any outcome. Further, there were more favourable responses to communicating imprecision or conflicting estimates relative to lack of evidence, where participants estimated the treatment would improve outcomes by 30-50% relative to a placebo. There were no differences across display formats, suggesting that, if well-designed, it may not matter which format is used. Future research on specific display formats or uncertainty types and with larger sample sizes would be needed to detect small effects. Implications for the communication of uncertainty are discussed.

Keywords: uncertainty, risk communication, visual displays

1 INTRODUCTION

Scientific evidence is often associated with uncertainties, such as imprecise, conflicting or even a lack of evidence about the effects of interventions (e.g., Budescu, Por, Broomell, & Smithson, 2014; Fischhoff & Davis, 2014; Glenton et al., 2010; Joslyn & LeClerc, 2012). Communicating these uncertainties to the public is paramount, to accurately reflect scientific knowledge, and to increase public understanding and tolerance towards uncertainty (Chalmers, 2004; Han, 2013; Johnson & Slovic, 1995; Spiegelhalter, 2017). Yet, it continues to present a challenge to experts and policy makers (Han, Klein, & Arora, 2011). A recent example can be found in communications about the emerging 2020 COVID-19 pandemic as experts have attempted to inform governments and the public on the basis of insufficient, imprecise or conflicting evidence (e.g., mortality rates, efficacy of masks or medications). One of the challenges to communicating uncertainty is that many of the techniques that have been developed have not been empirically tested or compared for communicating different types of uncertainty (Spiegelhalter, 2017; van der Bles et al., 2019). As such, evidence on the cognitive, psychological and behavioral responses to the communication of different types of uncertainty is lacking, as are clear evidence-based recommendations about how to integrate those uncertainties into existing risk communication formats (van der Bles et al., 2019).

The aim of the present study is to evaluate responses to different types of uncertainty information about the efficacy of medical interventions, and to develop and empirically test display formats for communicating those uncertainties. Specifically, we examine communications of a point estimate, an imprecise estimate, conflicting estimates or a lack of evidence. A *point estimate* represents the most common case in current risk communications, where the results of medical studies are aggregated to make a single numerical estimate of the number of people who are likely to experience an outcome (e.g., 4 in 1,000 patients are expected to experience a treatment benefit). However, estimates may be uncertain owing to measurement error (e.g., *imprecision* results in the estimation of a range of values for the expected benefit), opposing estimates from two different studies (e.g., *conflicting estimates*) or there may be an absence or insufficient data to make a clear numerical estimate (e.g., *lack of evidence*)[1]. The type of uncertainty may have different psychological effects or suggest different courses of action (Han et al., 2011). In order to facilitate the integration of these uncertainties into existing risk communications, we adapted evidence-based numerical and visual risk communication formats (Trevena et al., 2013), namely tables, bar graphs and icon arrays, to incorporate uncertainty information, drawing on insights from the uncertainty visualisation literature.

2 Communicating Uncertainty in Medical Evidence

To facilitate informed decision-making in health, information about risks, benefits, and uncertainties in medical evidence need to be communicated in ways that patients can understand (Trevena et al., 2013). There is an extensive literature on how to summarise and present point estimates in risk communications (Ancker, Senathirajah, Kukafka, & Starren,

[1]Han et al. (2011) classifies these types of uncertainty as ambiguity or the reliability, credibility or adequacy of the information, respectively.

2006; Garcia-Retamero & Cokely, 2017; Lipkus, 2007; Trevena et al., 2013). However, there are few empirically-based recommendations on how to communicate their associated *uncertainties*, and few health decision aids incorporate this type of information (Stacey et al., 2014). Further, as there are few studies that directly compare how people respond to different types of uncertainty, particularly in a medical context, there is a lack of evidence to inform risk commmunicators about the potential advantages or disadvantages of doing so. For instance, risk communicators lack evidence on how the communication of imprecision or conflicting estimates affects how people interpret the magnitude of benefits or harms of a treatment relative to a point estimate, or effects on other cognitive, trust or behavioral responses. Similarly, it is unclear when and to what effect communicating imprecision or conflicting estimates may be preferable to simply communicating there is a lack of evidence. We aim to provide guidance to risk communicators on these applied questions, and also explore whether responses are affected by the format in which the information is displayed.

2.1 Effects of communicating different types of uncertainty

Results from two recent reviews suggest that, on the whole, communicating uncertainty does not appear to have detrimental effects on a variety of attitudinal, behavioral and trust-related outcomes, although results are mixed for different uncertainty types, and tend to be more negative for conflicting estimates (Gustafson & Rice, 2020; van der Bles, van der Linden, Freeman, & Spiegelhalter, 2020). However, as the majority of studies have compared only one or two types of uncertainty, often imprecision against point estimates and/or numerical against verbal uncertainty (e.g., see Bansback, Harrison, & Marra, 2016; Han et al., 2011; van der Bles et al., 2020), it is unclear for which outcomes and for what types of uncertainty communications may have positive or negative effects. Studies that have compared three or more types of uncertainty, using either verbal statements or numerical estimates, report a mixture of results. For instance, whereas Kuhn (2000) found perceived riskiness of environmental hazards was similar across point, imprecise, conflicting estimates and a verbal description of uncertainty, Markon and Lemyre (2013) found conflicting evidence (a verbal statement describing conflicting evidence without numerical estimates) negatively affected risk acceptability, trust, and adherence to advisory warnings but only relative to lack of evidence. Similarly, Gustafson and Rice (2019) found negative effects of consensus uncertainty (verbal statement of conflicting evidence) on belief in scientific claims but no differences on perceived credibility or behavioral intentions relative to technical (imprecision), deficient (lack of evidence) or no uncertainty.

An important goal of risk communications is to help people to understand the absolute magnitude of the benefits and harms of medical treatments (Trevena et al., 2013). Yet few studies evaluate how people summarize or interpret treatment effects when provided with uncertainty information attributable to different types of uncertainty, or how much variability there is in their interpretations (for exceptions, see Benjamin & Budescu, 2018; Cabantous, Hilton, Kunreuther, & Michel-Kerjan, 2011). Studies that have used these metrics typically focus on how people interpret imprecision (e.g., Budescu et al., 2014; Dieckmann, Peters, & Gregory, 2015) and findings are mixed: people have been found to focus on either the upper or lower value in the range (Han et al., 2009; Highhouse, 1994) or the median (Benjamin & Budescu, 2018) and

around half of participants interpret numerical ranges as representing a uniform distribution with all outcomes equally likely (Dieckmann et al., 2015). In one of the few studies that have compared how people make estimates based on imprecision or conflicting estimates (in expert climate change forecasts), Benjamin and Budescu (2018) found that people made similar estimates across different types of uncertainty. However, when given expert evidence that varied in the breadth of or distance between two ranges (e.g., non-overlapping vs. overlapping ranges), participant's estimates deviated from expert forecasts and did so more in response to conflicting estimates relative to imprecision, suggesting differences in how people may summarise or interpret effects deriving from these types of uncertainty. It is not clear how uncertainty affects perceptions of the magnitude of benefits or harms in health information, or how much variability there is in interpretations across participants who view the same uncertainty information.

In the current study, we seek to evaluate how people interpret the effect of a medical intervention when given quantitative estimates either as a point estimate, imprecise or conflicting estimates or a verbal explanation of lack of evidence. Our interest in exploring responses to communications about a lack of evidence are twofold. First, evidence syntheses frequently conclude that there is a lack of sufficient evidence to provide clear estimates for some or even all study outcomes (e.g., Cochrane systematic reviews; e.g., Pollock, Gray, Culham, Durward, & Langhorne, 2014). Thus, understanding how people interpret such communications would be informative for risk communicators. Second, lack of evidence acts as an informative comparison to understand whether it is still better to communicate imprecision or conflicting estimates rather than to summarize that there is insufficient evidence to make a point estimate. There is some evidence that verbal statements of uncertainty can increase uncertainty perceptions or decrease trust relative to point estimates and/or imprecise ranges, yet at the same time may be more understandable to participants (Bansback et al., 2016; van der Bles et al., 2020).

2.2 Numerical and Visual Risk Communication Formats

As stated previously, while there are clear evidence-based recommendations for how to communicate point estimates, there are few empirically-based guidelines on how best to incorporate uncertainty information (Han, 2013; van der Bles et al., 2019). At the same time, a variety of visual design features have been proposed (e.g., see Spiegelhalter, 2017; Spiegelhalter, Pearson, & Short, 2011). We draw on these approaches, and evidence from the broader uncertainty visualisation literature, to design display formats to communicate different types of uncertainty. To facilitate integration into existing risk communications, we focused on adapting three common evidence-based formats for communicating numbers about medical evidence: tables, bar graphs, and icon arrays (Ancker et al., 2006; Lipkus, 2007; Trevena et al., 2013). We provide a brief overview of each of the formats along with studies that have explored integrating uncertainties into these formats. We describe the design of formats in more detail in the Method.

Tables are structured formats for communicating numerical probabilities (McDowell, Gigerenzer, Wegwarth, & Rebitschek, 2019; Schwartz, Woloshin, & Welch, 2009). Tables can be easily modified to communicate different types of uncertainty, by including multiple rows for conflicting study estimates or a verbal description for lack of evidence (left panel, Figure 1). Tables are currently used to communicate results from Cochrane systematic reviews of medical evi-

dence (Guyatt et al., 2011) and include a numerical representation of imprecision in the form of confidence intervals.

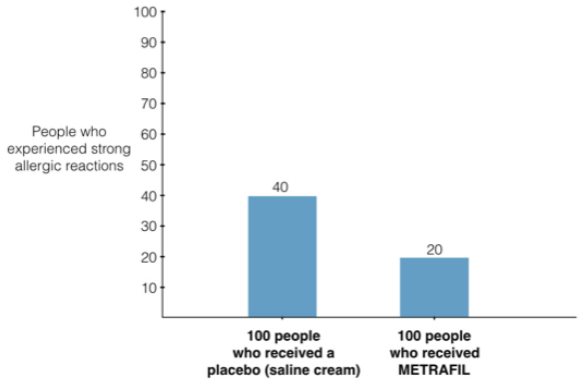
Bar graphs use the height or length of bars to display the magnitude of an estimate (middle panel, Figure 1). Bar graphs visualise part-to-whole relations when the risk is plotted proportional to the whole group (i.e., y-axis represents the reference class). As such, they build on visual processing capacities to compare quantities or magnitudes along a common scale (Ancker et al., 2006). Bar graphs have been used to communicate uncertainty related to imprecision, typically by using error bars to present confidence intervals (e.g., Correll & Gleicher, 2014; Han et al., 2011). However, error bars can result in 'within-the-bar bias': the tendency to consider values contained within the shaded portion of the bar to be more likely than those located outside the bar but within the confidence interval (Newman & Scholl, 2012). Adjusting the shading or transparency of error bars to indicate imprecision around a mean, for instance in a design similar to a box plot where only the range is shaded, reduces within-the-bar bias (Correll & Gleicher, 2014)[2].

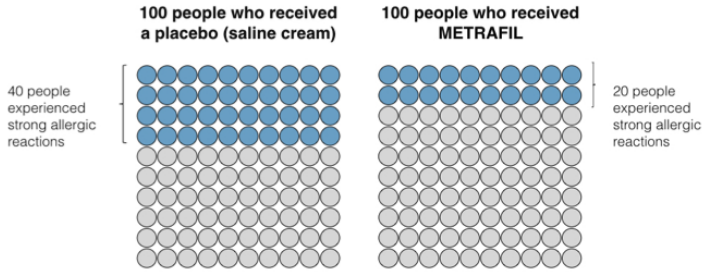
Icon arrays present frequencies out of 100 or 1,000 icons to allow for visual comparisons of quantities (Figure 1, right panel). Similar to bar graphs, icon arrays facilitate the comprehension of part-to-whole relations, and reduce potential biases in understanding, such as 'denominator neglect' (Ancker et al., 2006; Garcia-Retamero, Galesic, & Gigerenzer, 2010). Despite their efficacy for communicating point estimates, there are only a few studies of icon arrays adapted to communicate uncertainty, most often to communicate randomness (with scattered icons; Han et al., 2012; Kasper, Heesen, Köpke, Mühlhauser, & Lenz, 2011). Despite some practical examples suggesting how to communicate imprecision by incorporating shading around icons to indicate the confidence interval (e.g., Spiegelhalter et al., 2011), to our knowledge, only one study tested such a design and found that people understood uncertainty less than for verbal descriptions (Bansback et al., 2016).

[2]Box plots are currently not used in risk communications and may be unfamiliar to the lay public.

Point estimate

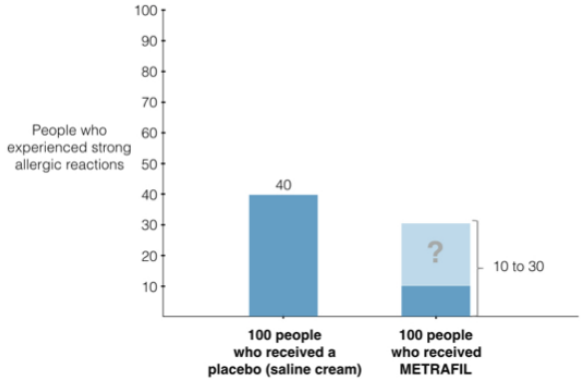
	100 people who received a placebo (saline cream)	100 people who received METRAFIL
Number of people who experienced strong allergic reactions	40	20

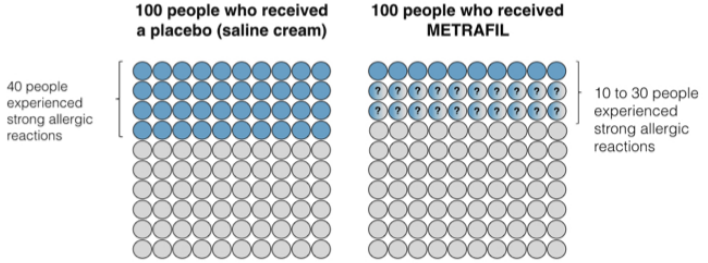




Imprecision

	100 people who received a placebo (saline cream)	100 people who received METRAFIL
Number of people who experienced strong allergic reactions	40	10 to 30





TABLES BAR GRAPHS ICON ARRAYS

Figure 1: Example *table*, *bar graph*, and *icon array* displays for each of the uncertain evidence types. For each type of uncertainty, we presented outcomes for people who do versus do not take a medication to facilitate comparisons of treatment effects relative to a placebo (Trevena et al., 2013). As shown in the examples, the expected risk for the *imprecision* and *conflicting estimates* conditions are consistent with the *point estimate* condition (the median of the two estimates is equivalent to the point estimate).

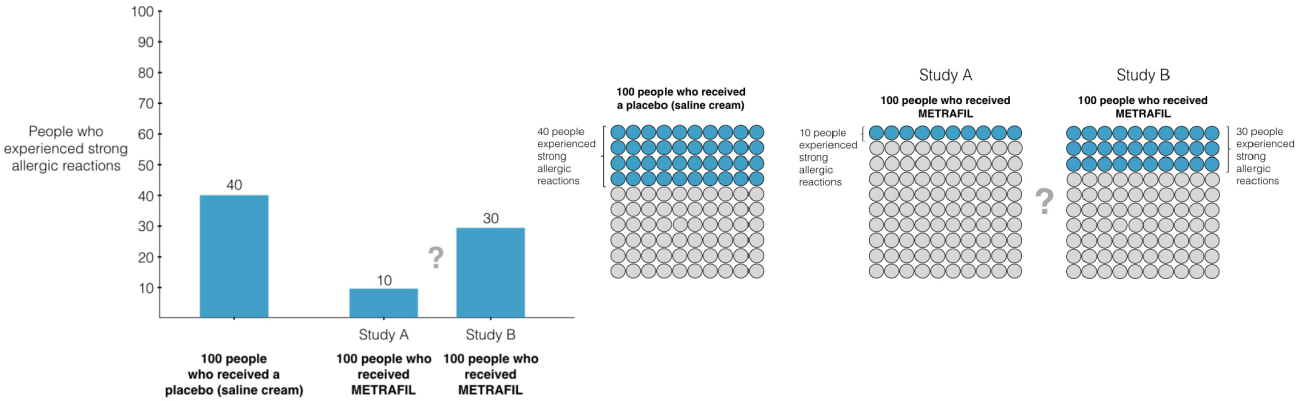
Conflicting evidence

Study A

	100 people who received a placebo (saline cream)	100 people who received METRAFIL
Number of people who experienced strong allergic reactions	40	10

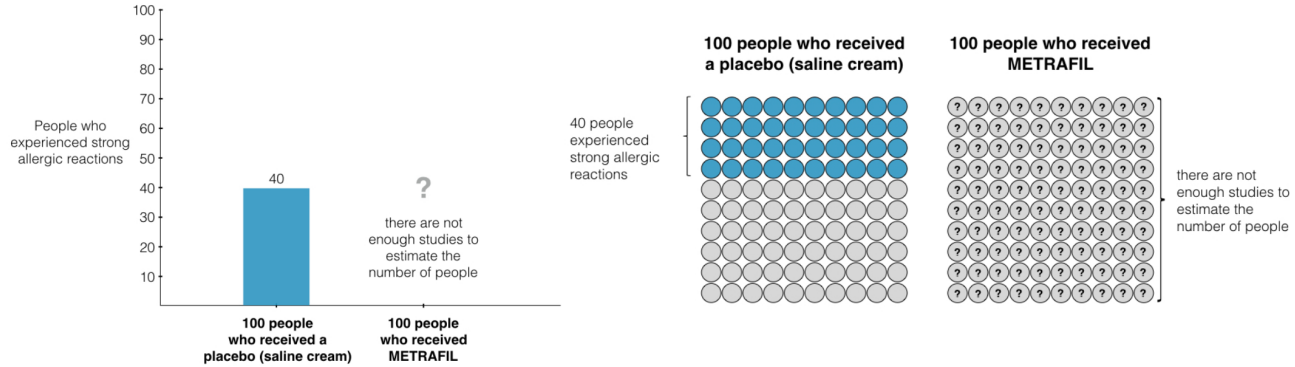
Study B

	100 people who received a placebo (saline cream)	100 people who received METRAFIL
Number of people who experienced strong allergic reactions	40	30



Lack of evidence

	100 people who received a placebo (saline cream)	100 people who received METRAFIL
Number of people who experienced strong allergic reactions	40	there are not enough studies to estimate the number of people



TABLES BAR GRAPHS ICON ARRAYS

Figure 1 (Cont.): Example *table*, *bar graph*, and *icon array* displays for each of the uncertain evidence types. For each type of uncertainty, we presented outcomes for people who do versus do not take a medication to facilitate comparisons of treatment effects relative to a placebo (Trevena et al., 2013). As shown in the examples, the expected risk for the *imprecision* and *conflicting estimates* conditions are consistent with the *point estimate* condition (the median of the two estimates is equivalent to the point estimate).

3 Overview of study and hypotheses

We examine the effect of communicating different types of uncertainty on people's interpretations of the effects of medical interventions. Specifically, we evaluate: (a) best estimates of treatment effects and the variation around participant's estimates, (b) recall, (c) subjective evaluations (usefulness and understanding), (d) perceptions of certainty, reliability, (e) trustworthiness, and (f) behavioral intentions. These outcomes map onto the cognition (a-d), trust (e), and behavioral responses (f) outlined in a recently published uncertainty communication framework (van der Bles et al., 2019). To facilitate practical applications, we highlight comparisons between imprecision or conflicting estimates relative to a point estimate and lack of evidence when reporting study results.

Our primary dependent variable is how people summarise estimates and how much estimates vary around the median of individual estimates (e.g., see Bruine de Bruin, van der Klaauw, van Rooij, Teppa, & de Vos, 2017). For instance, given one estimate or two estimates with the same median, upper and lower values but attributable to different types of uncertainty, how do participants summarise the evidence and is there consistency across uncertainty types and display formats? This metric reflects how much participant estimates differ from one another and has been used to evaluate the effect of survey modes and wording differences on estimates of inflation (Bruine de Bruin et al., 2017). It can therefore provide insights into how consistent participant's summaries are across uncertainty communications and whether certain display formats suggest more consistent interpretations than others.

We make the following predictions about how uncertainty type will affect the variation of estimates of treatment effects. Specifically, depending on whether imprecision and conflicting estimates are interpreted by participants as coming from the same or from two distinct distributions (e.g., that imprecise ranges represent a single but two conflicting estimates represent two distinct distributions), we propose two alternative hypotheses:

> H1: Relative to *imprecision* and *conflicting estimates*, variation will be (1) lower in the *point estimate* condition and (2) higher in the *lack of evidence* condition.
>
> > H1a: Assuming participants interpret *imprecision* as a single distribution and *conflicting estimates* as two distinct distributions, variation will be higher for *conflicting estimates* than for *imprecision*.
> > H1b: Assuming participants interpret both *imprecision* and *conflicting estimates* as coming from the same underlying distribution/s, variation will be similar for the two conditions.

For display format, recent studies have suggested people comprehend information about numerical risks similarly in tables and icon arrays (Hawley et al., 2008; McDowell et al., 2019) and some studies suggest similar low error rates for comprehension of icon arrays and vertical bar graphs (Feldman-Stewart, Brundage, & Zotov, 2007). However, given that studies of uncertainty communications based on bar graphs find that distributions are more poorly understood relative to other visual designs (e.g., Newman & Scholl, 2012; Okan, Garcia-Retamero, Cokely, & Maldonado, 2018), we hypothesise:

H2: Variation will be similar for *tables* and *icon arrays*, and both will be lower than variation in response to *bar graphs*.

To examine how well uncertainty communications achieve their intended goals, it is also important to understand how they affect different cognitive, trust, and behavioural responses (van der Bles etal., 2019). We pose the following general research question:

RQ: What is the effect of communicating uncertainty on recall, subjective evaluations, perceptions of uncertainty, trustworthiness, and behavioral intentions?

For these cognitive, trust, and behavioral outcomes, evidence is mixed or lacking so our analyses for these outcomes are more exploratory. In general, we expect results to follow the same pattern as H1a and H2 for type of uncertainty and display format. As there are no studies examining whether certain display formats are better for communicating different types of uncertainty information, we will conduct an interaction analysis. We control for numeracy and graph literacy as both factors are associated with how people understand numerical and visual information about point estimates (Hawley et al., 2008; Okan, Garcia-Retamero, Galesic, & Cokely, 2012). We also include an item assessing distributional perceptions for the *imprecision* and *conflicting estimates* conditions and report results descriptively.

4 METHOD

4.1 Participant recruitment

Participants were recruited via Amazon Mechanical Turk (MTurk; Paolacci, Chandler, & Ipeirotis, 2010). Participants were eligible for the study if they were above 18 years of age, had $\geq$98% approval rating and had completed $\geq$500 HITS (tasks), had not participated in prior studies conducted by the research team on visual formats, completed the study on a desktop or laptop computer, and completed a basic instruction task. The basic instruction task required participants to answer three multiple choice questions on the initial study description to ensure they understood the premise of the study. Participants were given two attempts to answer the items correctly. On average, the study took 18 minutes to complete and participants were paid US2.10 (average hourly rate of around US7.20)[3]. The study was approved by the Max Planck Institute for Human Development ethics committee. The study was pre-registered on the Open Science Framework (https://osf.io/gcf3r).

To determine the required sample size to detect small effects at .80 power and $\alpha = .02$ between different types of uncertainty and display formats for the main estimation outcome, a power analysis was conducted based on pilot data (https://osf.io/gcf3r). To reach an upper estimate of 140 participants per condition, and accounting for potential exclusions during preprocessing, data collection was terminated when $\sim$1700 surveys had been completed.

[3]The remuneration was increased from US1.80 for a 15 minute study based on average completion times for the first 30 participants who took longer than anticipated based on a pilot study. These participants received a bonus for the difference in payment. Changes to the remuneration were approved by the ethics committee.

4.2 Uncertainty communication stimuli

4.2.1 Uncertainty types

As an introduction to the study, participants received a short text about how the results of medical studies inform estimates of treatment benefits or harms. Following this introduction, participants received one of four explanations about how or why information about the effects of medical interventions may be uncertain[4]. These can be found in Table 1.

Table 1: Descriptions of different uncertainty types presented to participants in each condition.

In the best case, results from medical studies can be combined to make a precise estimate about the number of people who experienced a benefit (e.g., did not get sick) or a harm (e.g., experience a side-effect) in groups of people who do or do not take the medication. In some cases, there can be some uncertainty about this estimate.
Point estimate condition For example, there is uncertainty about **who (e.g., which individual)** will experience a benefit or harm as a result of the treatment. The results of medical studies can only provide a **precise estimate** of how many people experienced a benefit or harm as a result of the treatment. *Imprecision condition* For example, there may be uncertainty owing to **how well outcomes are measured** across medical studies. In such cases, the results of medical studies can only provide an **imprecise estimate** of how many people experienced a benefit or harm as a result of the treatment. *Conflicting estimates condition* For example, there may be uncertainty associated with **conflicting results** from medical studies. In such cases, medical studies may provide **different estimates** of how many people experienced a benefit or harm as a result of the treatment. *Lack of evidence condition* For example, there may be uncertainty owing to a **lack of medical studies** on a medication. In such cases, there are **not enough studies** to make an estimate of how many people would experience a benefit or harm as a result of the treatment.

For each type of uncertainty, we presented outcomes for people who do versus do not take a medication to facilitate comparisons relative to a placebo. The *point estimate* condition received a single numerical value, the *imprecision* and *conflicting estimates* conditions received two numerical values presented either as a range or two independent study values, respectively. The two values represented a 20-point difference and the expected risk was consistent with the *point estimate* condition (e.g., median of the two values equalled the point estimate value; see Table 2). The *lack of evidence* condition provided no numerical estimate with the statement: "there are not enough studies to estimate the number of people."

[4]To ensure the general concept of uncertainty was activated for all conditions, the aleatory uncertainty inherent in estimates (e.g., who/which individual would experience the outcome) was made explicit in the point estimate condition to avoid attributions of uncertainty to the numbers.

4.2.2 Display formats

We designed uncertainty communications by incorporating design features into *tables*, *bar graphs* and *icon arrays*. Across visual displays, we incorporated an extrinsic property[5], a question mark, to communicate uncertainty in *imprecision*, *conflicting estimates* and *lack of evidence*. A question mark indicates a query, doubt, or missing data and has been suggested as a glyph for communicating uncertainty (Gershon, 1998; Schünemann, Best, Vist, Oxman, & Group, 2003) but, to our knowledge, has not been empirically tested. Consistent with prior research, we also modified intrinsic properties[6] for *imprecision*. For the *bar graphs* and *icon arrays* we followed recommendations made by Correll and Gleicher (2014) to use visually symmetric and visually continuous representations and represent uncertainty across the range using shading (middle and right panel, Figure 1). For the *lack of evidence* condition, all display formats included a verbal statement alongside the numerical placebo comparison information.

4.3 Design

Participants were randomly assigned to one of 12 between-subjects conditions: 4 (type of uncertainty: *point estimate, imprecision, conflicting estimates, lack of evidence*) x 3 (display format: *table, bar graph, icon array*). Participants completed two tasks. Participants were presented with estimates about the effectiveness of two hypothetical medications, Parezon and Metrafil, to prevent strong allergic reactions in people with allergies. Both medications were presented in the same format. Across the two medications, the placebo value differed and the size of the treatment effect was manipulated within-subjects to be *small* or *moderate*, to assess whether responses were stable across varying effect sizes. Table 2 presents the numbers for each type of uncertainty and medication.

Table 2: Numbers for treatment and placebo effects for each type of uncertainty. Participants who received numbers for small effect size for Parezon received numbers for moderate effect size for Metrafil (and vice versa).

	Health condition and medication					
	Parezon (hay fever)			Metrafil (atopic eczema)		
		Treatment effect size			Treatment effect size	
Uncertainty type	Placebo	Small effect	Moderate effect	Placebo	Small effect	Moderate effect
Point estimate	50	40	25	40	30	20
Imprecision	50	30 to 50	15 to 35	40	20 to 40	10 to 30
Conflicting estimates	50	30 vs. 50	15 vs. 35	40	20 vs. 40	10 vs. 30
Lack of evidence	50	NA	NA	40	NA	NA

Notes: For *conflicting estimates*, numbers refer to effect for Study A vs. Study B. The order of presentation of the higher and lower study estimates was counterbalanced. Treatment effect size could not be manipulated for the *lack of evidence* condition as no numerical estimates were provided.

The two medications were presented in the same order (Parezon first, Metrafil second) but the treatment effect size was counterbalanced such that participants who received a *small effect*

[5] *Extrinsic* representations incorporate new objects or glyphs within the display, such as arrows, symbols or error bars (Bisantz et al., 2009; Gershon, 1998; Kinkeldey, MacEachren, & Schiewe, 2014).

[6] *Intrinsic* representations incorporate uncertainty within an existing display by altering visual variables, such as hue, brightness or transparency (e.g., increasing colour transparency with distance from the mean).

for Parezon received the *moderate effect* for Metrafil and vice versa. Treatment effect size could not be manipulated for the *lack of evidence* condition as no numerical estimates were provided.

4.4 Measures

Estimates of treatment effect. Participants were asked to make an estimate for the treatment group:

> Imagine a new group of 100 people who take [medication]. What would be your best estimate of the number of people who will experience strong allergic reactions:
>
> ___ out of 100 people who take [medication] will experience strong allergic reactions.
>
> Please provide a short description of how you reached your estimate. *Please be as specific as possible.*

The item was repeated for the placebo group. Estimation strategies were coded (see Table S1 in the Supplementary Material) and summarised descriptively. At the end of the study, participants in the *imprecision* and *conflicting estimates* conditions completed a multiple choice question about how they perceived the distribution of estimates. Results are reported in the Supplementary Material.

Recall. For the *point estimate* condition, participants were asked to recall the number of people who experienced allergic reactions when receiving the medication. Participants in the *imprecision* and *conflicting estimates* conditions were asked to recall the lower and higher estimates. Recall items were scored as correct if the participant provided the exact numbers as shown[7]. Participants in the *lack of evidence* condition were asked to provide a verbal description about the reason no numbers were provided about the treatment group. Responses were coded as correct if they mentioned "lack of evidence" or "not enough studies".

Subjective evaluations. Participants rated how understandable and useful the information was on 5-point Likert scales, ranging from 'not at all'-'very'. Higher scores indicated better evaluations. The two items were combined into a composite score (r=.46).

Certainty perceptions. Participants rated how reliable and how uncertain they perceived the information to be, on 5-point Likert scales ranging from 'not at all'-'very'. Prior to analysis, uncertainty was reversed-scored so that low values represented low perceived certainty. Items were combined into a composite 'certainty perception' score (r=.81).

Trustworthiness. Participants rated how trustworthy the information was on a 5-point Likert scale ranging from 'not at all'-'very'. Higher scores indicated greater trustworthiness[8].

Behavioral intentions. Participants were asked to imagine that they suffered from the health condition in each scenario and assume that the costs of the medication would be covered by

[7] e.g., in the *conflicting estimate* and *imprecision* conditions, participants were scored as correct if they identified both estimates correctly, irrespective of whether they correctly recalled the order.

[8] In the pre-registration, we stated that trustworthiness ratings would be analysed as part of a composite with subjective evaluations. We opted to adhere to the categorisation of trust as a distinct response for comparability with a recently published framework van der Bles et al. (2019). When combined with positive evaluations, results were largely consistent

their health insurance and there were no other medications available on the market. Participants rated how likely would it be that they would take the medication on an 11-point Likert scale ranging from 'I would definitely not take - definitely take [Parezon/ Metrafil]'.

Numeracy and graph literacy. Numeracy was assessed with the adaptive version of the Berlin Numeracy Test (BNT; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). Scores range from 1-4. Graph literacy was assessed with the graph literacy scale, short version (Okan, Janssen, Galesic, & Waters, 2019). Scores range from 0-4. For both scales, higher scores indicate greater numeracy/graph literacy.

4.5 Procedure

For each medication, participants made a best estimate of the treatment effect and completed a behavioral intention item on the same page as the display format, followed by subjective evaluations, trustworthiness, certainty perceptions, and recall items on separate pages. Distractor tasks were completed between the evaluation and recall items to increase the delay (details of the first distractor task can be found in the pre-registration; in between the evaluation and recall items for the second medication, participants completed the numeracy task). Graph literacy, distributional perceptions, and demographic items (age, gender, education, prior history of allergic rhinitis and/or atopic eczema) were completed at the end of the study.

4.6 Analysis Plan

The variation around estimates was calculated using the mean absolute deviation from the median of participant's estimates (see Bruine de Bruin et al., 2017). Specifically, within each of the 12 conditions (uncertainty type x display format), the median estimate was summarised for each medication and treatment effect size. We then calculated the absolute difference between a participant's estimate and the respective median. For example, if the median estimate for *imprecision* in the *table* display format group for Metrafil (small effect) was 30, the absolute deviation score for a participant who gave an estimate of 25 would be: $|25 - 30| = 5$. Recall of treatment and placebo numbers for both medications were coded such that each participant received a score of correct (vs. incorrect) for the treatment and placebo numbers they recalled.

Mixed effects regressions were used to evaluate the effect of uncertainty type and display format on outcomes. In each model, type of uncertainty and display format were included as fixed effects and participant as a random effect. That is, the model estimated the influence of uncertainty type and display format on outcomes, taking into account that the data are nested (or clustered) within participants. For the recall model, the recall value type (recall of treatment vs. recall of placebo numbers) was included as a fixed effect in models to examine whether there was any difference in recall of numbers for the treatment or placebo group. In each model, we controlled for numeracy and graph literacy. The uncertainty type *point estimate* and display format *table* served as the reference groups for analyses. We report 95% confidence intervals for fixed effects. Paired comparisons were made between uncertainty conditions and display formats using Tukey HSD adjustment for multiple comparisons (using the *emmeans* package in R). Exploratory analyses examined whether there were any interactions between uncertainty type and display format.

As secondary analyses, mixed effects regressions without the *lack of evidence* condition were also run to evaluate whether the manipulation of treatment effect size in the *point estimate*, *imprecision*, and *conflicting estimates* conditions had any effect on outcomes. Treatment effect size (small and moderate) was varied within-subjects for the two medications. Treatment effect size did not have an effect on outcome measures in a pilot study (except for a small effect on behavioral intention) and we made no hypotheses about its effect on responses. We report on secondary analyses in a subsection of the results and refer to detailed models in the Supplementary Material.

5 RESULTS

5.1 Sample

Of the 2432 participants who started the study, 2313 (95%) consented, 2221 attempted and 1925 (87%) successfully completed the basic instruction task. Two hundred and eight participants (11%) did not complete the study, leaving a final sample of 1717 (71%)[9]. All participants who completed the study were included in analyses[10]. On average, participants were 36 years old (SD=11.3, 18-79, median=34), 51.0% were male, and the majority had completed a university degree or higher (67.5%). One tenth indicated a prior history of eczema (10.6%) and one fifth a prior history of allergic rhinitis (22.5%). Compared to the US population, the sample was slightly younger (US population median=38.5) and much more highly educated (43.4% of US residents have completed a university degree or higher), with a similar gender distribution (49% male; Statista, 2020; US Census Bureau, 2018). The average numeracy score was 2.26 (SD=1.17, range 1-4) and 76.9% of participants answered two or more graph literacy items correctly (M=2.36, SD=1.18). Numeracy and graph literacy were moderately correlated (r=.38).

5.2 Overview of models

Across all analyses, there was an effect of uncertainty type but no effects associated with any of the display formats. Further, in no model did the inclusion of an interaction between uncertainty type and display format improve model fit. Accordingly, we report results without an interaction. To simplify the presentation of results, we focus on reporting differences in uncertainty type and refer to paired comparisons of estimated marginal means derived from the models (Table S2). We nevertheless report model parameters for display format in Tables and Figures, and provide paired comparisons tables in the Supplementary Material (Table S3). We briefly discuss potential reasons for the lack of effects for display format in the Discussion.

[9] All but four of these participants dropped out prior to completing the second medical topic. Owing to a technical issue, these four participants were not included as they were not assigned a treatment effect code.

[10] For quality assurance, we planned to remove respondents who took less than two standard deviations below the median completion time. No participants met this criteria. The median completion time was 15 minutes, with upper limits of one hour. Open-ended responses for fast completion times revealed meaningful responses.

5.3 Effect of uncertainty type on estimates of treatment effect

5.3.1 Best estimates and estimation strategies.

Table 3 presents a descriptive summary of participant's best estimates and most common estimation strategies for each uncertainty type and display format. Figure 2 visualises the estimates as a deviation from the *point estimate* or median of two given values (*imprecision* and *conflicting estimates* conditions), or from the placebo value for the *lack of evidence* condition. Two-thirds of participants (66.9%) used an identical strategy for the two treatments. Not all estimation strategies could be coded from the explanations provided.

For the *point estimate* condition, perhaps not surprisingly, participants made a best estimate that was in line with what they were shown. Over two thirds of participants provided an estimate that was identical to the point estimate number shown to them. The second most common strategy was to provide an estimate within +/- 5 points from the point estimate (Table 3).

The estimation strategies for the *imprecision* and *conflicting estimates* conditions were similar to one another, and the primary strategy used by over two thirds of participants – to estimate the median of the two treatment values – matched the expected risk for the *point estimate* condition. That is, the majority of participants perceived the mid-point of the two values to be the best estimate (see also Supplementary Material showing almost two thirds of participants in each condition perceived the underlying distribution for the imprecision and conflicting estimates values to be consistent with a normal distribution). The second most common strategy was to provide an estimate that was within +/- 5 points of the median value, and there was no clear preference to select the upper versus the lower of the two treatment values.

As seen in Figure 2, there appeared to be two strategies to make estimates on the basis of *lack of evidence*: to estimate the placebo value, or to estimate a treatment effect that suggested a benefit of at least 20 percentage points or more relative to the placebo (49.8% of participants). Indeed, one fifth of participants simply halved the placebo value, suggesting that they thought the best estimate of the treatment effect would be a 50% relative risk reduction. Around a fifth of participants considered that, in the absence of sufficient studies to make an estimate, the best estimate would be the baseline risk of the placebo group.

When asked to estimate a value for the placebo group, most participants provided an estimate that matched the original placebo value (70.7%) or +/- 5 from this value (5.4%). A small percentage gave estimates 10+ percentage points higher (4.1%) or lower (5.3%) than this value.

5.3.2 Variation in estimates across uncertainty types

Given the consistency in estimation strategies, variation – measured by the mean absolute deviations from each group's median – was small across type of uncertainty and display format conditions. The exception was for the lack of evidence, where there was greater variation around the median estimate. An analysis of the variation found an effect for type of uncertainty but not display format (Table 4, see also Table S3).

Paired comparisons of estimated marginal means found no differences between the *point estimate*, *imprecision*, and *conflicting estimates* conditions (Table S2). That is, despite participants in the *imprecision* and *conflicting estimates* conditions being given a 20-point range

for the estimate of a treatment effect, they did not provide more or less varied estimates relative to participants who received a single numerical point estimate. Within these conditions, participant's appeared to make estimates that were largely consistent to one another.

Given no numerical estimates in the *lack of evidence* condition, participants gave more varied estimates than in other uncertainty conditions. Specifically, there was less variation in the *imprecision* (difference: -11.45, 95%CI -12.70, -10.19, $p<.001$, Cohen's $d = -2.03$) and *conflicting estimates* conditions (difference: -11.57, 95%CI -12.83, -10.30, $p<.001$, $d = -2.06$) relative to the *lack of evidence* condition. There was also less variation in the *point estimate* condition (difference: -12.40, 95%CI -13.66, -11.14, $p<.001$, $d = -2.20$). Thus, even if study estimates of a treatment effect are imprecise or conflicting, providing those estimates to participants would result in more consistent interpretations of the effect than summarising the evidence as lacking in a verbal statement.

Table 3: Summary statistics and most common estimation strategies for best estimates of the treatment effect for each uncertainty type and display format

Uncertainty type	Display format			
Point Estimate	Table n=148	Bar graph n=140	Icon Array n=146	All formats
Summary statistics for best estimate				
Median deviation from point estimate[a]	0.0	0.0	0.0	0.0
Range	-25 – 40	-15 – 50	-13 – 60	-25 – 60
Mean variation (SD)[b]	2.77 (6.03)	2.40 (6.57)	3.19 (8.11)	2.79 (6.96)
Most common estimation strategies(%)[c]				
Point estimate number shown	65.9	71.4	69.2	68.8
Anchor +/- 5 on the point estimate	17.6	13.6	12.3	14.5
Not codeable	9.8	6.8	7.9	8.2
Imprecision	Table n=148	Bar graph n=150	Icon Array n=140	All formats
Summary statistics for best estimate				
Median deviation from median of two estimates[a]	0.0	0.0	0.0	0.0
Range	-18 – 80	-15 – 50	-10 – 45	-18 – 80
Mean variation (SD)[b]	4.09 (9.96)	3.91 (7.88)	3.80 (7.03)	3.93 (8.39)
Most common estimation strategies (%)[c]				
Median of two estimates	69.3	62.3	62.5	64.7
Anchor +/- 5 on median of estimates	9.5	13.3	10.7	11.2
Upper of two estimates	6.1	2.7	8.6	5.7
Lower of two estimates	2.7	8.0	6.4	5.7
Not codeable	6.8	7.3	7.5	7.2
Conflicting estimates	Table n=144	Bar graph n=141	Icon Array n=142	All formats
Summary statistics for best estimate				
Median deviation from median of two estimates[a]	0.0	0.0	0.0	0.0
Range	-30 – 55	-23 – 55	-20 – 70	-30 – 70
Mean variation (SD)[b]	3.46 (7.53)	4.10 (8.55)	3.91 (8.01)	3.82 (8.03)
Most common estimation strategies (%)[c]				
Median of two estimates	70.1	67.4	67.3	68.3
Anchor +/- 5 on median of estimates	8.3	8.2	7.4	8.0
Upper of two estimates	3.8	6.0	6.3	5.4
Lower of two estimates	4.5	3.2	3.5	3.7
Not codeable	7.6	7.1	8.8	7.8
Lack of evidence	Table n=142	Bar graph n=135	Icon Array n=141	All formats
Summary statistics for best estimate				
Median deviation from placebo value[a]	-20.0	-15.0	-20.0	-19.0
Range	-50 – 30	-50 – 35	-50 – 40	-50 – 40
Mean variation (SD)[b]	15.85 (11.09)	15.64 (10.05)	14.46 (10.69)	15.32 (10.63)
Most common estimation strategies (%)[c]				
10+ below placebo estimate	37.3	37.8	36.2	37.1
Half of placebo	19.0	15.2	21.3	18.5
Placebo as treatment	16.2	18.9	20.2	18.4
10+ above placebo estimate	9.9	9.3	5.7	8.3
Not codeable	11.6	8.9	11.0	10.5

Notes: [a]The median deviation from the point estimate value (*point estimate* condition) or median of the given values (*imprecision* and *conflicting estimates*), or the placebo value (*lack of evidence*). [b]Calculated as the mean absolute deviation from each group's median estimate. [c]Participant's estimation strategies were coded according to a pre-defined coding scheme. Full details of all the coding categories can be found in the Supplementary Material.

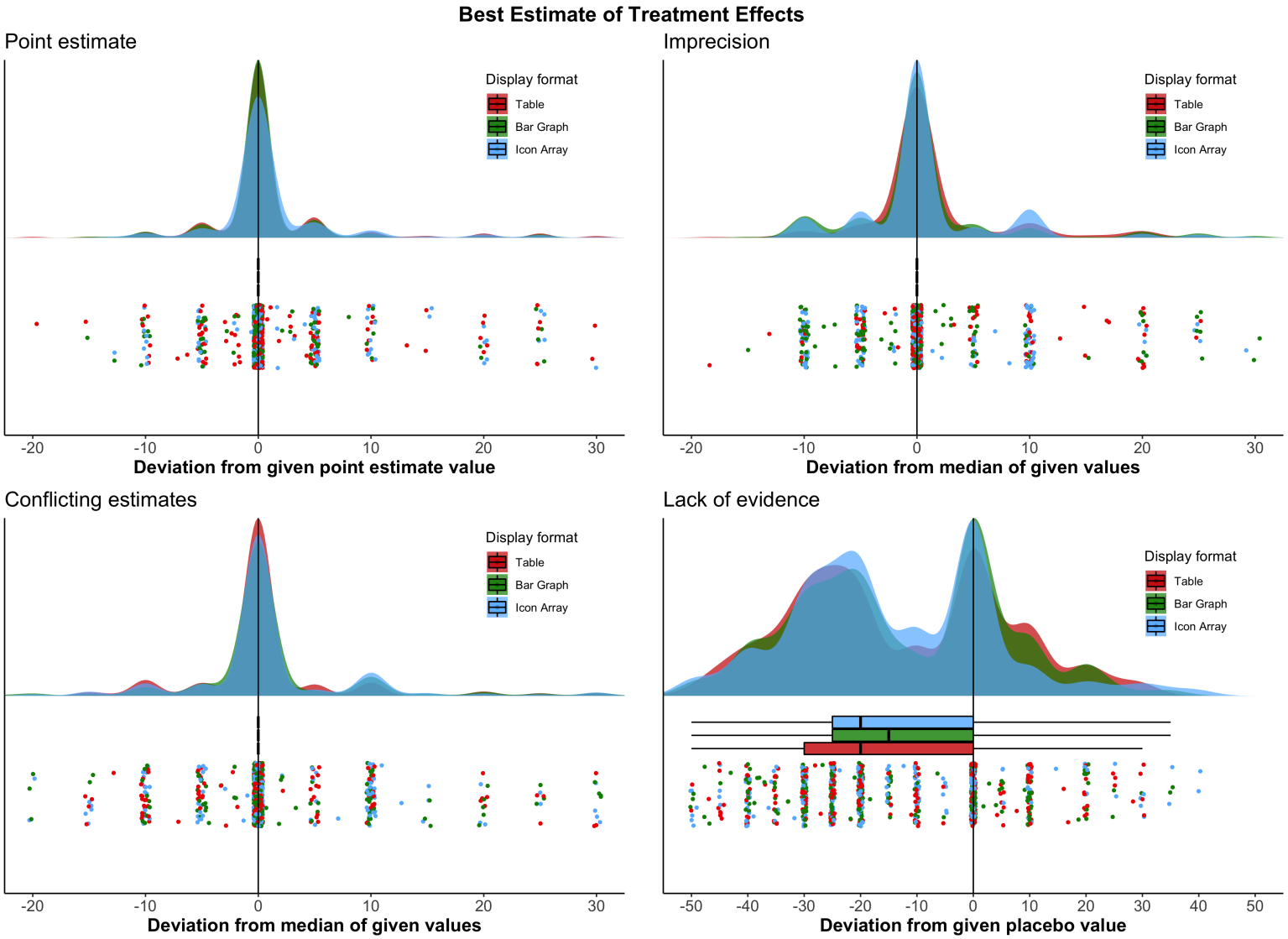


Figure 2: Raincloud plots showing estimates as deviation from the point estimate value (*point estimate* condition) or median of the given values (*imprecision* and *conflicting estimates*), or the placebo value (*lack of evidence*). Colours show values for display formats. For each display format, density plot, box plot, and jittered raw data are shown (note, as most responses were the point estimate/median value, boxplots are at the solid line). Extreme deviation scores were trimmed from the plot. Note: Participants who gave a prediction that matched the point estimate (72%), median of imprecise (66%) or conflicting values (69%) or placebo (*lack of evidence*; 23%) have a deviation score of zero.

Table 4: Results of linear mixed effects models on variation, subjective evaluations, certainty perceptions, trust, and behavioral intentions.

	Variation[a]	Subjective evaluations[b]	Certainty perceptions[b]	Trust[b]	Behavioral intentions[c]
	Estimate [95%CI]	Estimate [95%CI]	Estimate [95%CI]	Estimate [95%CI]	Estimate [95%CI]
Fixed effects					
Intercept	8.93 [7.75; 10.08]	4.34 [4.23; 4.46]	4.17 [4.03; 4.31]	4.09 [3.96; 4.21]	7.99 [7.57; 8.36]
Uncertainty type					
Point estimate	referent	referent	referent	referent	referent
Imprecision	0.95 [0.00; 1.86]	−0.00 [−0.10; 0.09]	0.02 [−0.09; 0.13]	0.01 [−0.11; 0.12]	−0.26 [−0.57; 0.06]
Conflicting estimates	0.83 [−0.12; 1.81]	−0.06 [−0.16; 0.03]	−0.10 [−0.22; 0.01]	−0.13 [−0.24; −0.02]	−0.34 [−0.66; −0.03]
Lack of evidence	12.40 [11.39; 13.33]	−0.54 [−0.63; −0.44]	−0.40 [−0.52; −0.28]	−0.29 [−0.41; −0.18]	0.12 [−0.20; 0.44]
Display format					
Table	referent	referent	referent	referent	referent
Bar graph	−0.08 [−0.92; 0.81]	−0.04 [−0.12; 0.05]	−0.07 [−0.18; 0.04]	−0.07 [−0.17; 0.02]	−0.22 [−0.49; 0.06]
Icon array	−0.12 [−0.94; 0.76]	0.01 [−0.07; 0.09]	−0.01 [−0.11; 0.09]	−0.02 [−0.11; 0.08]	−0.01 [−0.30; 0.29]
Covariates					
Graph literacy[d]	−1.89 [−2.20; −1.58]	−0.00 [−0.03; 0.03]	−0.14 [−0.18; −0.10]	−0.09 [−0.12; −0.05]	−0.18 [−0.27; −0.07]
Numeracy[e]	−0.66 [−0.98; −0.35]	−0.02 [−0.05; 0.01]	−0.04 [−0.08; −0.00]	0.02 [−0.02; 0.06]	−0.03 [−0.13; 0.07]
Random effects					
Intercept σ^2	35.39	0.42	0.68	0.58	3.73
Residual	31.64	0.15	0.15	0.20	3.84

Note: referent = reference group; Estimates are unstandardized coefficients and square brackets indicate bootstrapped confidence intervals. [a]Mean absolute deviations from each group's median estimate; [b]Measured on 5-point Likert scales ranging from 'not at all'–'very'. [c]11-point Likert scale 'I would definitely not – definitely take [medication]. [d]Scores range 0-4 with higher scores indicating greater graph literacy. [e]Scores range from 1-4 with higher scores indicating greater numeracy.

5.4 Recall

Figure 3 shows the percentage of participants who correctly recalled the presented values. Recall accuracy of the treatment values was high across all conditions, with recall slightly lower in the *lack of evidence* condition (74.8%) relative to the other uncertainty conditions (all >83%). More than 85% of participants across all conditions correctly recalled the placebo value. Analysis of total recall found an effect for type of uncertainty but not display format. Participants were more likely to recall correctly the placebo correctly than the treatment value (Table S5 and Figure 3).

Despite participants in the *imprecision* and *conflicting estimates* conditions having to recall two rather than one estimate relative to the *point estimate* condition, there were no meaningful differences in short-term recall between these conditions. The task of recalling a reason for the lack of numbers as opposed to actual numbers appeared to be a more difficult task for participants in the *lack of evidence* condition. Participants were more likely to recall values correctly in the *imprecision* (Odds Ratio=2.25, 95%CI 1.14, 4.46, p=.012) and there was a small advantage recalling *conflicting estimates* relative to the *lack of evidence* condition (OR=1.88, 95%CI 0.96, 3.68, p=.078; Table S2). Recall was also better in the *point estimate* condition (OR=3.18, 95%CI 1.58, 6.41, p<.001).

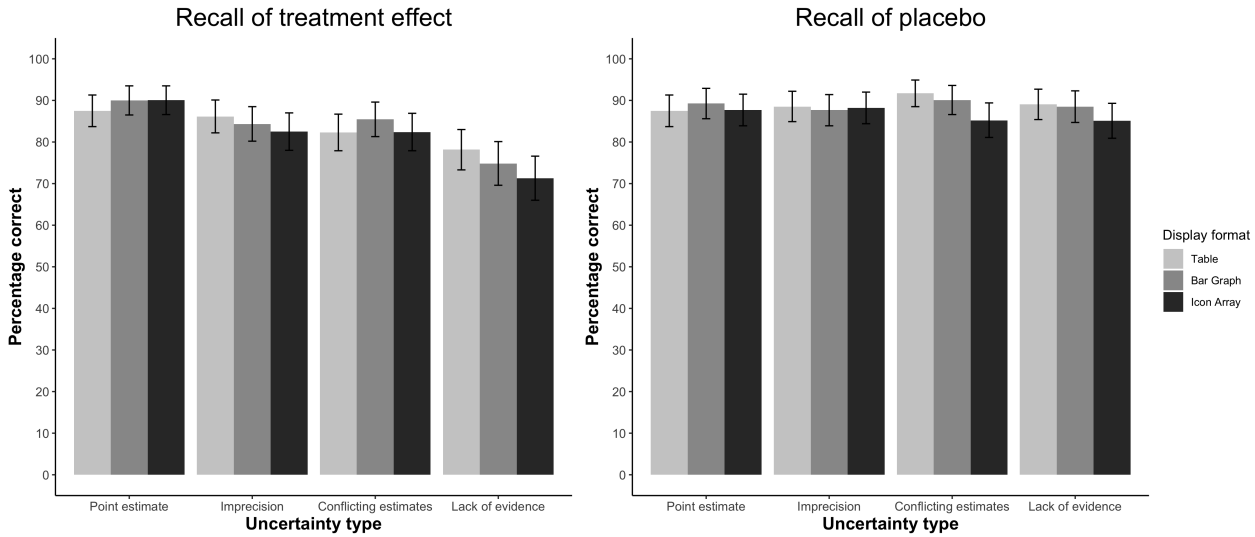


Figure 3: Recall by uncertainty type and display format. Error bars represent 95% confidence intervals. Note: participants in the *imprecision* and *conflicting estimates* conditions had to recall both study values correctly, and responses for participants in the *lack of evidence* condition were coded as correct if they mentioned "lack of evidence" or "not enough studies".

5.5 Subjective evaluations

The second column in Table 4 reports the results of analyses on the composite 'subjective evaluation' rating. Participants rated the information as similarly understandable and useful in the *point estimate*, *imprecision* and *conflicting estimates* conditions. Compared to the *lack of evidence*, participants rated *imprecision* (difference: 0.54, 95%CI 0.41, 0.66, p<.001, d = 1.37) and *conflicting estimates* (difference: 0.48, 95%CI 0.35, 0.60, p<.001, d = 1.22; Table S2) as

more understandable and useful. Participants also evaluated *point estimates* more positively than *lack of evidence* (difference: 0.54, 95%CI 0.41, 0.66, $p<.001$, $d = 1.38$).

5.6 Certainty perceptions

The results of analyses on the composite 'certainty' rating are reported in the third column of Table 4. Similar to subjective evaluations, there were no differences in how reliable or certain participants perceived the information in the *point estimate*, *imprecision* and *conflicting estimates* conditions. However, *imprecision* and *conflicting estimates* were perceived to be more certain than *lack of evidence* (difference: 0.41, 95%CI 0.26, 0.57, $p<.001$, $d = 1.05$ and 0.30, 95%CI 0.14, 0.45, $p<.001$, $d = 0.75$, respectively; Table S2). Not surprisingly, *point estimates* were rated as more certain than *lack of evidence* (difference: 0.40, 95%CI 0.24, 0.55, $p<.001$, $d = 1.01$).

5.7 Trustworthiness

Analyses of trustworthiness ratings are reported in the fourth column of Table 4. Ratings were high across all uncertainty types and display formats (all Means>3.6). There was a tendency to rate the trustworthiness of *point estimates* slightly higher than for *conflicting estimates* (difference: 0.13, 95%CI -0.01, 0.28, $p=.088$, $d = 0.29$) and to trust *imprecision* slightly more than *conflicting estimates* (difference: 0.14, 95%CI -0.01, 0.28, $p=.066$, $d = 0.31$), although differences were small. Trustworthiness ratings for *imprecision* and *conflicting estimates* were nevertheless higher than for *lack of evidence* (difference: 0.30, 95%CI 0.16, 0.45, $p<.001$, $d = 0.67$ and 0.16, 95%CI 0.02, 0.31, $p=.021$, $d = 0.36$, respectively; Table S2). *Point estimates* were also rated as more trustworthy than *lack of evidence* (difference: 0.29, 95%CI 0.15, 0.44, $p<.001$, $d = 0.65$).

5.8 Behavioral intentions

On average, intentions to take the medication were high across conditions (M=7.3, SD=2.8) and were influenced by the treatment effect size manipulation (see also section 5.10). Figure 4 shows that intentions were higher for a moderate (versus small) treatment effect size for the *point estimate*, *imprecision* and *conflicting estimates* conditions. Intentions to take the medication were lower in the *conflicting estimates* relative to *point estimates* in the main analysis, but the difference was not found when examining estimated marginal mean differences between these conditions in paired comparisons (Table S2). There were no other differences in intentions to take the medication between these uncertainty conditions (fifth column, Table 4).

Relative to the *lack of evidence* condition, participants reported lower behavioral intentions for *conflicting estimates* (difference: -0.46, 95%CI -0.88, -0.04, $p=.025$, $d = -0.23$) and a tendency for lower intentions for *imprecision* (difference: -0.38, 95%CI -0.80, 0.04, $p=.088$, $d = -0.19$; Table S2). However, as visible in Figure 4, behavioral intentions were higher than *lack of evidence* for the moderate treatment effect size manipulation in the *conflicting estimates* condition, suggesting that this effect may have been driven by lower intentions to take the medication when estimates conflicted and the treatment effect size was small.

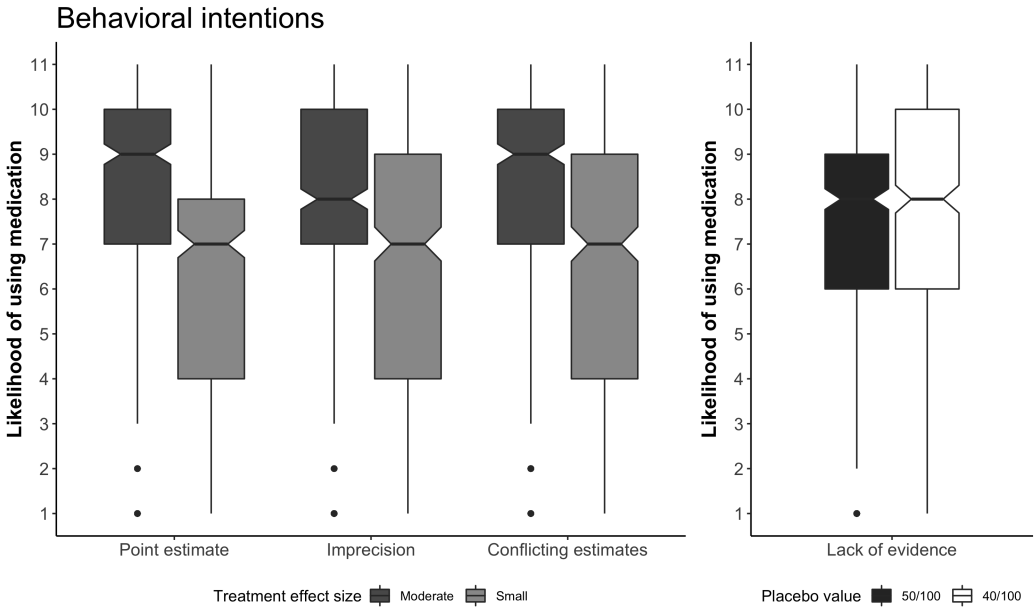


Figure 4: Left panel shows behavioral intention ratings when the treatment effect size was moderate or small for the *point estimate*, *imprecision* and *conflicting estimates* conditions. Right panel shows treatment intentions for the *lack of evidence* condition according to whether the placebo value was 50/100 (Parezon) or 40/100 (Metrafil).

5.9 Covariates: Graph literacy and numeracy

Graph literacy was associated with most outcome measures. In general, higher graph literacy scores were associated with less variation in estimates, better recall, lower certainty perception and trustworthiness ratings and lower behavioral intentions to take a medication (Tables 4 and S5). Higher numeracy scores were also associated with less variation in estimates, better recall, and lower certainty perception ratings.

5.10 Secondary analyses: Treatment effect size manipulation

Secondary analyses including the within-subjects treatment effect size (small versus moderate) manipulation in models are reported in Supplementary Table S4.[11] Across all models, results were consistent with the primary analyses reported above. However, the treatment effect size manipulation did have a main effect on outcomes. Specifically, compared to a moderate treatment effect size, a small treatment effect was associated with less variation in estimates, lower subjective evaluation ratings, lower certainty perceptions, lower trustworthiness ratings and lower behavioral intentions (see also Figure 4). The treatment effect size manipulation did not affect recall.

6 DISCUSSION

The aim of the present study was to evaluate how participants responded to different types of uncertainty in communications about medical evidence. We also examined whether responses

[11]These models excluded the *lack of evidence* condition as treatment effect could not be manipulated in this condition.

differed across three evidence-based display formats adapted to communicate these uncertainties. Our results have several implications for communicating uncertainty about medical evidence.

First, our results suggest no clear adverse affects of communicating about *imprecision* or *conflicting estimates* relative to a *point estimate* in risk communications. Our findings suggest that, when estimating the magnitude of benefit of a medication, participants interpreted *conflicting estimates* and *imprecision* in a similar way by focusing on values at or close to the median of conflicting or imprecise interval bounds (Benjamin & Budescu, 2018; Cabantous et al., 2011; Kuhn, 2000; Markon & Lemyre, 2013). Further, estimates in both of these conditions were similar and variation in estimation strategies was no greater than those those made by participants in the *point estimate* condition. These results partially support Hypothesis 1b in that there was no difference between *imprecision* and *conflicting estimates*. Indeed, there were almost no differences on any outcome measure between these three uncertainty conditions[12]. The only exception was in relation to trustworthiness, where conflicting estimates were associated with slightly lower ratings of trustworthiness in the information. Although we did not emphasise conflict or disagreement in our explanation for conflicting estimates, additional context for how and why estimates may conflict (e.g., differences in measurement, sample, or assessment time frame) may increase trustworthiness in the case of conflicting estimates. Taken together, the absence of clear differences between conditions suggests that it is not detrimental to communicate these scientific uncertainties to the general public.

Second, the results suggest that it would be preferable to communicate *imprecision* and *conflicting estimates* when available rather than to communicate that there is a *lack of evidence*, as people may perceive greater benefit in the absence of any estimate. Despite rating *lack of evidence* as less certain, less understandable or useful for decision-making, and less trustworthy relative to other types of uncertainty, participants in the *lack of evidence* condition reported high intentions to take the medication. At the same time, participants estimated the treatment benefit to be around 12-20 points greater than the placebo, or in other words, they estimated that the treatment would have a relative risk reduction of around 30-50%. The perception that a treatment being tested must therefore be beneficial is consistent with research on people's perceptions of the efficacy of new drugs. Thirty-nine percent of a representative sample of the US general public believed that the Federal Drug Administration (FDA) only approved drugs that were extremely effective and over two-thirds of participants would select a newer over an older drug when both were described as equally effective (Schwartz & Woloshin, 2011). Our results suggest caution when communicating a lack of evidence as people may presume that any treatment being tested is expected to be effective. Additional disclaimers or information about the conduct of clinical trials may help to reduce these perceptions (Schwartz & Woloshin, 2011).

Third, estimates of treatment effects did not vary between display formats and there were no interactions between display format and uncertainty type. These results are in contrast to our Hypothesis 2, and may be a result of using display formats that were designed and adapted in

[12]The general concept of uncertainty was activated for the point estimate condition by making explicit the aleatory uncertainty inherent in estimates (e.g., which individual would experience the outcome). Prior studies on point estimates do not typically communicate this information explicitly. Future studies could assess whether this uncertainty communication could also affect people's interpretation of the precision of the point estimate.

accordance with recommendations for communicating point estimates (McDowell et al., 2019; Trevena et al., 2013). Future research could evaluate whether display format effects emerge when manipulating the degree of uncertainty (e.g., larger ranges) or when multiple pieces of information require integration (e.g., comparing multiple benefits and harms). Nevertheless, our results suggest that the three evidence-based display formats adapted to communicate uncertainty in the present study did not affect how people responded to the information. Accordingly, it may not matter which of the three tested formats is used, although additional research on specific formats or with larger sample sizes is needed in order to detect small effects between formats.

Individuals with higher graph literacy varied less in their estimates of treatment effect, were more likely to recall the provided values, and reported lower certainty perceptions, trustworthiness ratings, and behavioral intentions relative to those with lower graph literacy. Similarly, more numerate participants varied less in their estimates, were more likely to recall the provided values and had lower certainty perceptions compared to less numerate participants. Consistent with prior studies on risk communications, these results suggest that basic graphical and numerical competencies affect how people understand uncertainty information (Hawley et al., 2008; Okan et al., 2012). Future studies could systematically explore how these individual differences moderate responses to uncertainty communications.

Our study included a within-subjects manipulation of treatment effect size (small vs. moderate) for the *point estimate*, *imprecision* and *conflicting estimates* conditions to assess whether any observed differences were stable across varying treatment effect sizes. Although treatment effect size had an influence on multiple outcomes, including treatment effect size in analyses did not change any of the observed effects between uncertainty type and display format. In our study, we opted to hold the size of uncertainty around estimates constant across medications, and to vary the size of the treatment effect. Another approach would be to hold the treatment effect constant but vary the degree of uncertainty (e.g., by increasing or decreasing the range; see Benjamin & Budescu, 2018)). Alternatively, varying the topic of uncertainty (e.g., environmental, health risks) could explore how the results of the present study generalise or differ across domains, particularly as prior studies have found responses to uncertainty information can vary by topic (e.g., see Jensen & Hurley, 2012).

The results of our study should be interpreted in light of the following limitations. First, the primary outcome was a numerical estimate of the treatment effect, which is a very cognitive approach to evaluating responses to uncertainty information. Subjective measures, such as risk perceptions or other emotional or affective responses may have revealed more subtle differences between uncertainty types or display formats. A recent review of research on uncertainty communications concluded that there is less work on emotional or affective reactions to uncertainty information than for cognitive responses, and results of existing studies are inconsistent and may depend on how such responses are defined, measured or represented (van der Bles et al., 2019). Future work would benefit from incorporating more affective measures to help close this important research gap. Second, the high recall scores across conditions were likely a consequence of the short delay between the presentation of outcomes and recall questions. Increasing this delay may affect recall scores and reveal subtle differences between the formats.

Third, we focused on two medications which may have been too few to examples to detect small effects. Incorporating a greater number of scenarios, varying both degree of uncertainty and treatment effect size would be an even more robust test of the effects of uncertainty type and display formats on outcomes. Also, future studies should explore whether results hold given non-hypothetical medical scenarios with greater salience and personal relevance to participants. Finally, the study may have been underpowered to detect small effects between display formats or uncertainty types, and future research with larger sample sizes is needed to confirm these null findings.

7 CONCLUSION

We have shown that participants responded to different types of uncertainty in communications about medical evidence in similar ways. These findings suggest that people may not be adversely affected by communications about uncertainty in medical evidence, such as imprecision or conflicting information (McDowell, Rebitschek, Gigerenzer, & Wegwarth, 2016). Rather, the absence of any clear differences between the point estimate, imprecision, and conflicting estimates conditions suggests that it would not be detrimental for risk communicators to communicate these scientific uncertainties to the general public. However, caution should be taken when communicating a lack of evidence, where treatment effects were estimated to be large in the absence of clear data. Future research should assess how to communicate lack of evidence, without treatment effects being overestimated by target audiences. Further, three different display formats adapted to incorporate uncertainty information, namely tables, bar graphs, and icon arrays, all show great promise for communicating these different types of uncertainty. Additional research on specific display formats or uncertainty types and with larger sample sizes would be needed to detect small effects. Nevertheless, the results of the present study suggest that communications that adopt these formats may promote awareness and understanding of uncertainties in scientific evidence and enable people to make better informed decisions about their health.

ACKNOWLEDGEMENTS

The work was completed while the corresponding author was at the Max Planck Institute for Human Development. We would like to thank Clara Schirren for helping to implement the visual designs and Mirta Galesic for the many helpful comments on previous versions of the manuscript.

References

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, *13*(6), 608-18. doi: 10.1197/jamia.M2115

Bansback, N., Harrison, M., & Marra, C. (2016). Does introducing imprecision around probabilities for benefit and harm influence the way people value treatments? *Medical Decision Making*, *36*(4), 490-502. doi: 10.1177/0272989x15600708

Benjamin, D. M., & Budescu, D. V. (2018). The role of type and source of uncertainty on the processing of climate models projections. *Frontiers in Psychology*, *9*(403). doi: 10.3389/fpsyg.2018.00403

Bisantz, A. M., Stone, R. T., Pfautz, J., Fouse, A., Farry, M., Roth, E., ... Thomas, G. (2009). Visual representations of meta-information. *Journal of Cognitive Engineering and Decision Making*, *3*(1), 67-91. doi: 10.1518/155534309x433726

Bruine de Bruin, W., van der Klaauw, W., van Rooij, M., Teppa, F., & de Vos, K. (2017). Measuring expectations of inflation: Effects of survey mode, wording, and opportunities to revise. *Journal of Economic Psychology*, *59*, 45-58. doi: 10.1016/j.joep.2017.01.011

Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of ipcc probabilistic statements around the world. *Nature Climate Change*, *4*, 508. doi: 10.1038/nclimate2194

Cabantous, L., Hilton, D., Kunreuther, H., & Michel-Kerjan, E. (2011). Is imprecise knowledge better than conflicting expertise? evidence from insurers? decisions in the united states. *Journal of Risk and Uncertainty*, *42*(3), 211-232. doi: 10.1007/s11166-011-9117-1

Chalmers, I. (2004). Well informed uncertainties about the effects of treatments. *BMJ*, *328*(7438), 475-476. doi: 10.1136/bmj.328.7438.475

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25-47.

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142-51. doi: 10.1109/tvcg.2014.2346298

Dieckmann, N. F., Peters, E., & Gregory, R. (2015). At home on the range? lay interpretations of numerical uncertainty ranges. *Risk Analysis*, *35*(7), 1281-1295. doi: 10.1111/risa.12358

Feldman-Stewart, D., Brundage, M. D., & Zotov, V. (2007). Further insight into the perception of quantitative information: judgments of gist in treatment decisions. *Medical Decision Making*, *27*(1), 34-43. doi: 10.1177/0272989x06297101

Fischhoff, B., & Davis, A. L. (2014). Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences*, *111*(Supplement 4), 13664.

Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics. *Human Factors*, *59*(4), 582-627. doi: 10.1177/0018720817690634

Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Medical Decision Making*, *30*(6), 672-684. doi: 10.1177/0272989x10369000

Gershon, N. (1998). Visualization of an imperfect world. *Computer Graphics and Applications, IEEE*, *18*(4), 43-45. doi: 10.1109/38.689662

Glenton, C., Santesso, N., Rosenbaum, S., Nilsen, E. S., Rader, T., Ciapponi, A., & Dilkes, H. (2010). Presenting the results of cochrane systematic reviews to a consumer audience: A qualitative study. *Medical Decision Making*, *30*(5), 566-577. doi: 10.1177/0272989x10375853

Gustafson, A., & Rice, R. E. (2019). The effects of uncertainty frames in three science communication topics. *Science Communication*, *41*(6), 679–706. doi: 10.1177/1075547019870811

Gustafson, A., & Rice, R. E. (2020). A review of the effects of uncertainty in public science communication. *Public Understanding of Science*, *29*(6), 614–633. doi: 10.1177/0963662520942122

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., ... Schünemann, H. J. (2011). Grade guidelines: 1. introduction?grade evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, *64*(4), 383-394. doi: http://dx.doi.org/10.1016/j.jclinepi.2010.04.026

Han, P. K. J. (2013). Conceptual, methodological, and ethical problems in communicating uncertainty in clinical evidence. *Medical care research and review*, *70*(1 Suppl), 14s-36s. doi: 10.1177/1077558712459361

Han, P. K. J., Klein, W. M. P., & Arora, N. K. (2011). Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, *31*(6), 828-838. doi: 10.1177/0272989X11393976

Han, P. K. J., Klein, W. M. P., Killam, B., Lehman, T., Massett, H., & Freedman, A. N. (2012). Representing randomness in the communication of individualized cancer risk estimates: Effects on cancer risk perceptions, worry, and subjective uncertainty about risk. *Patient Education and Counseling*, *86*(1), 106-113. doi: 10.1016/j.pec.2011.01.033

Han, P. K. J., Klein, W. M. P., Lehman, T. C., Massett, H., Lee, S. C., & Freedman, A. N. (2009). Laypersons' responses to the communication of uncertainty regarding cancer risk estimates. *Medical Decision Making*, *29*(3), 391-403.

Hawley, S. T., Zikmund-Fisher, B., Ubel, P., Jancovic, A., Lucas, T., & Fagerlin, A. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Education and Counseling*, *73*(3), 448-455. doi: 10.1016/j.pec.2008.07.023

Highhouse, S. (1994). A verbal protocol analysis of choice under ambiguity. *Journal of Economic Psychology*, *15*, 621-635. doi: 10.1016/0167-4870(94)90014-0

Jensen, J. D., & Hurley, R. J. (2012). Conflicting stories about public scientific controversies: Effects of news convergence and divergence on scientists' credibility. *Public Understanding of Science*, *21*, 689–704. doi: 10.1177/0963662510387759

Johnson, B. B., & Slovic, P. (1995). Presenting uncertainty in health risk assessment: initial studies of its effects on risk perception and trust. *Risk Analysis*, *15*(4), 485-94.

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126-40. doi: 10.1037/a0025185

Kasper, J., Heesen, C., Köpke, S., Mühlhauser, I., & Lenz, M. (2011). Why not? - communicating stochastic information by use of unsorted frequency pictograms - a randomised controlled trial. *Psycho-Social Medicine*, *8*, Doc08. doi: 10.3205/psm000077

Kinkeldey, C., MacEachren, A. M., & Schiewe, J. (2014). How to assess visual communication of uncertainty? a systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, *51*(4), 372-386. doi: 10.1179/1743277414Y.0000000099

Kuhn, K. M. (2000). Message format and audience values: Interactive effects of uncertainty information and environmental attitudes on perceived risk [Journal Article]. *Journal of Environmental Psychology*, *20*(1), 41-51. doi: 10.1006/jevp.1999.0145

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. *Medical Decision Making*, *27*(5), 696-713. doi: 10.1177/0272989x07307271

Markon, M.-P. L., & Lemyre, L. (2013). Public reactions to risk messages communicating different sources of uncertainty: An experimental test. *Human and Ecological Risk Assessment: An International Journal*, *19*(4), 1102-1126. doi: 10.1080/10807039.2012.702015

McDowell, M., Gigerenzer, G., Wegwarth, O., & Rebitschek, F. G. (2019). Effect of tabular and icon fact box formats on comprehension of benefits and harms of prostate cancer screening: A randomized trial. *Medical Decision Making*, *39*(1), 41-56. doi: 10.1177/0272989X18818166

McDowell, M., Rebitschek, F. G., Gigerenzer, G., & Wegwarth, O. (2016). A simple tool for communicating the benefits and harms of health interventions. *MDM Policy & Practice*, *1*(1), 1-10. doi: doi:10.1177/2381468316665365

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*(4), 601-607. doi: 10.3758/s13423-012-0247-5

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2018). Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy. *Quarterly Journal of Experimental Psychology*, *71*(12), 2506-2519. doi: 10.1177/1747021817744546

Okan, Y., Garcia-Retamero, R., Galesic, M., & Cokely, E. T. (2012). When higher bars are not larger quantities: On individual differences in the use of spatial information in graph comprehension. *Spatial Cognition & Computation*, *12*(2-3), 195-218. doi: 10.1080/13875868.2012.659302

Okan, Y., Janssen, E., Galesic, M., & Waters, E. A. (2019). Using the short graph literacy scale to predict precursors of health behavior change. *Medical Decision Making*, *39*(3), 183-195. doi: 10.1177/0272989X19829728

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411-419.

Pollock, A., Gray, C., Culham, E., Durward, B., & Langhorne, P. (2014). Interventions for improving sit-to-stand ability following stroke. *Cochrane Database of Systematic Reviews*(5). doi: 10.1002/14651858.CD007232.pub4

Schünemann, H. J., Best, D., Vist, G., Oxman, A. D., & Group, f. T. G. W. (2003). Letters, numbers, symbols and words: how to communicate grades of evidence and recommenda-

tions. *Canadian Medical Association Journal*, *169*(7), 677-680.

Schwartz, L. M., & Woloshin, S. (2011). Communicating uncertainties about prescription drugs to the public: a national randomized trial. *Arch Intern Med*, *171*(16), 1463-8. doi: 101001/archinternmed2011396

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2009). Using a drug facts box to communicate drug benefits and harms: two randomized trials. *Annals of Internal Medicine*, *150*(8), 516-27. doi: 10.7326/0003-4819-150-8-200904210-00106

Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, *4*(1), 31-60. doi: 10.1146/annurev-statistics-010814-020148

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393-1400. doi: 10.1126/science.1191181

Stacey, D., Légaré, F., Col, N. F., Bennett, C. L., Barry, M. J., Eden, K. B., ... Wu, J. H. (2014). Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews*, *CD001431*. doi: 10.1002/14651858.CD001431.pub4

Statista. (2020). *Total population in the united states by gender from 2010 to 2024.* `https://www.statista.com/statistics/737923/us-population-by-gender/`. (Accessed: 23 March, 2020)

Trevena, L. J., Zikmund-Fisher, B. J., Edwards, A., Gaissmaier, W., Galesic, M., Han, P. K. J., ... Woloshin, S. (2013). Presenting quantitative information about decision outcomes: A risk communication primer for patient decision aid developers. *BMC Medical Informatics and Decision Making*, *13*(SUPPL. 2), S7. doi: 10.1186/1472-6947-13-S2-S7

US Census Bureau. (2018). *Census data tables, 2018.* `https://data.census.gov/cedsci/?` (Accessed: 23 March, 2020)

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, *6*(5), 181870. doi: doi:10.1098/rsos.181870

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, *117*(14), 7672-7683. doi: 10.1073/pnas.1913678117

Supplementary Material

Definitions of coding categories for estimation strategies

Descriptions of estimation strategies were classified according to a pre-determined coding scheme developed on the basis of pilot data and supplemented with additional categories developed inductively using a subsample of strategy descriptions from the main study (Table S1)[13]. A second coder double-coded one third of the strategies for reliability (Cohen's κ=0.73). Estimation strategies are summarised descriptively.

Table S1: Definitions of coding strategies for estimates of treatment and placebo effects.

Estimation Strategy	Definition
When one number shown (e.g., point estimate condition)	
Numbers as shown	Provide the exact number shown.
When two numbers shown (e.g., imprecision or conflicting estimates)	
Upper number shown	Provide the exact upper number shown.
Lower number shown	Provide the exact lower number shown
Median estimate	Provide the median (average) of the two numbers shown.
Any case	
Anchoring	Make a small adjustment – within +/- 5 range – to the a) lower or b) upper of two values, or c) given or median of two values.
10+ below/above lowest/ highest number	Provide an estimate that is 10+ points above or below the lowest or highest number shown (or placebo for the *lack of evidence*).
Half of placebo	Provide an estimate that is half the placebo value.
Placebo for treatment	Used placebo as treatment estimate.
Guessed or not codeable	Strategy did not fit any other coding category or was unclear
*Additional strategies**	
Average placebo and treatment	Averaged the placebo and treatment value
Average all studies	Averaged two treatment and placebo values
50 percent	Stated that they gave a 50/50 estimate
Anchor on placebo	Make a small adjustment – within +/-5 range – to the placebo value
Strategy did not match	The stated strategy did not match the number provided

Note: A category *Half lowest number shown* was removed in place of the categories 10+ below/above lowest/highest estimate.; *Strategies were added to the coding scheme after sub-sample of coding was completed.

Distributional perceptions.

At the end of the study, participants in the *imprecision* and *conflicting estimates* condition completed a multiple choice question on whether they perceived the numbers in the range [two different study estimates] to be equally likely, or values in the middle of the range [middle of the two estimates], at the low end of the range [closer to the lower estimate], or high end of the range [closer to the higher estimate] to be more likely. These perceptions relate to uniform, normal, positively or negatively skewed distributions, respectively. The item was adapted from Dieckmann et al. (2015).

[13]As it was expected that not all participants would provide clear descriptions that could be easily coded (e.g., "*I took it from the graph*"), numerical estimates were also classified according to the coding categories. Results were largely consistent, except for a higher proportion of "not codeable" responses in the *lack of evidence* condition

The majority of participants in the *imprecision* and *conflicting estimates* conditions interpreted the numbers they received as consistent with a normal (63.2% and 62.3%, respectively) or a uniform distribution (20.1% and 23.9%). The remaining participants perceived values closer to the lower estimate (10.5% and 10.1%) as slightly more likely than values closer to the higher estimate (5.7% and 3.0%), consistent with a positively and negatively skewed distribution, respectively. Responses were largely consistent across display formats (see S1).

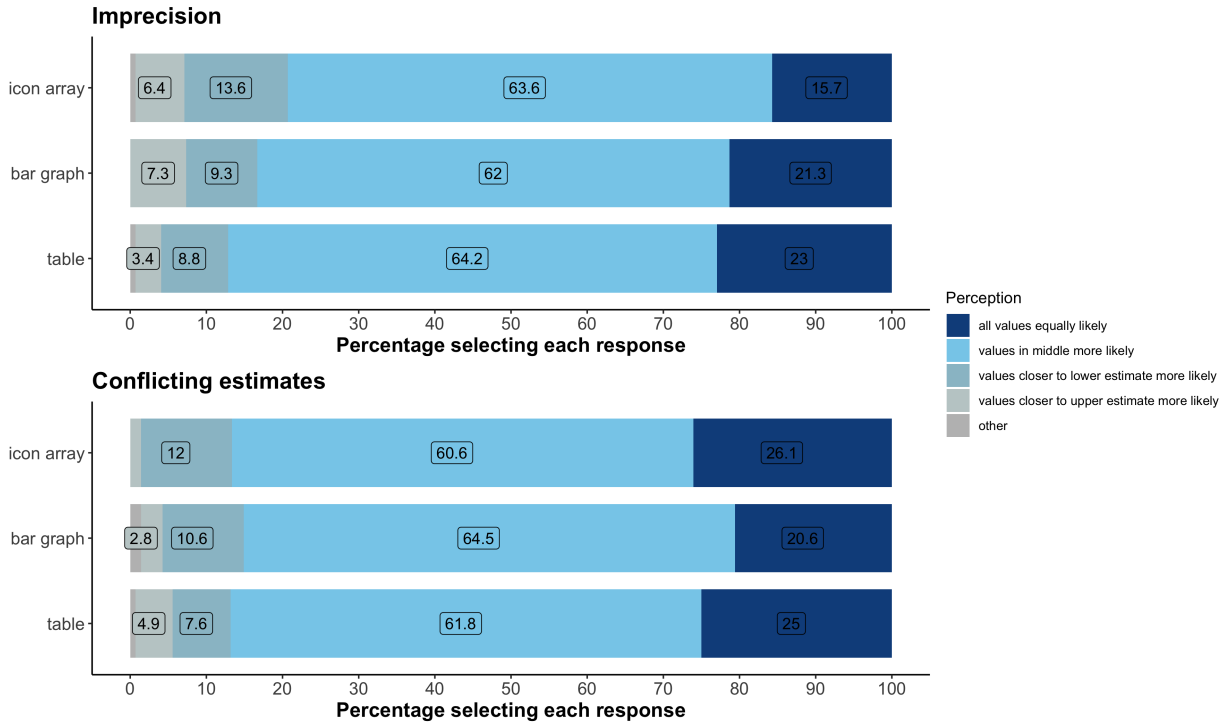


Figure S1: Distribution perceptions for *imprecision* and *conflicting estimates* for each display format. Numbers in bars refer to percentages. Percentages for "other" responses were small and are not shown in numbers.

Estimated Marginal Means: Uncertainty Type

Table S2: Estimated marginal mean differences for uncertainty type for each outcome measure.

	contrast	difference estimate	SE	p.value	lower 95%CI	upper 95%CI	effect size
	Variation						
1	point estimate - imprecision	-0.95	0.49	0.20	-2.20	0.29	-0.17
2	point estimate - conflicting	-0.83	0.49	0.32	-2.09	0.42	-0.15
3	point estimate - lack of evidence	-12.40	0.49	0.00	-13.66	-11.14	-2.20
4	imprecision - conflicting	0.12	0.49	0.99	-1.13	1.37	0.02
5	imprecision - lack of evidence	-11.45	0.49	0.00	-12.70	-10.19	-2.03
6	conflicting - lack of evidence	-11.57	0.49	0.00	-12.83	-10.30	-2.06
	Subjective evaluations						
1	point estimate - imprecision	0.00	0.05	1.00	-0.12	0.13	0.01
2	point estimate - conflicting	0.06	0.05	0.57	-0.06	0.19	0.16
3	point estimate - lack of evidence	0.54	0.05	0.00	0.41	0.66	1.38
4	imprecision - conflicting	0.06	0.05	0.61	-0.06	0.18	0.15
5	imprecision - lack of evidence	0.54	0.05	0.00	0.41	0.66	1.37
6	conflicting - lack of evidence	0.48	0.05	0.00	0.35	0.60	1.22
	Certainty perceptions						
1	point estimate - imprecision	-0.02	0.06	0.99	-0.17	0.14	-0.04
2	point estimate - conflicting	0.10	0.06	0.32	-0.05	0.26	0.26
3	point estimate - lack of evidence	0.40	0.06	0.00	0.24	0.55	1.01
4	imprecision - conflicting	0.12	0.06	0.19	-0.03	0.27	0.30
5	imprecision - lack of evidence	0.41	0.06	0.00	0.26	0.57	1.05
6	conflicting - lack of evidence	0.30	0.06	0.00	0.14	0.45	0.75
	Trustworthiness						
1	point estimate - imprecision	-0.01	0.06	1.00	-0.15	0.14	-0.01
2	point estimate - conflicting	0.13	0.06	0.09	-0.01	0.28	0.29
3	point estimate - lack of evidence	0.29	0.06	0.00	0.15	0.44	0.65
4	imprecision - conflicting	0.14	0.06	0.07	-0.01	0.28	0.31
5	imprecision - lack of evidence	0.30	0.06	0.00	0.16	0.45	0.67
6	conflicting - lack of evidence	0.16	0.06	0.02	0.02	0.31	0.36
	Behavioral intentions						
1	point estimate - imprecision	0.26	0.16	0.37	-0.15	0.67	0.13
2	point estimate - conflicting	0.34	0.16	0.16	-0.08	0.76	0.17
3	point estimate - lack of evidence	-0.12	0.16	0.88	-0.54	0.30	-0.06
4	imprecision - conflicting	0.08	0.16	0.96	-0.34	0.49	0.04
5	imprecision - lack of evidence	-0.38	0.16	0.09	-0.80	0.04	-0.19
6	conflicting - lack of evidence	-0.46	0.16	0.03	-0.88	-0.04	-0.23
		Odds Ratio	SE	p.value	lower 95%CI	upper 95%CI	
	Recall						
1	point estimate / imprecision	1.41	0.39	0.60	0.69	2.87	
2	point estimate / conflicting	1.69	0.47	0.22	0.84	3.43	
3	point estimate / lack of evidence	3.18	0.87	0.00	1.58	6.41	
4	imprecision / conflicting	1.20	0.32	0.90	0.60	2.39	
5	imprecision / lack of evidence	2.25	0.60	0.01	1.14	4.46	
6	conflicting / lack of evidence	1.88	0.49	0.08	0.96	3.68	

.
Results are averaged over the levels of: display format. Degrees-of-freedom method: asymptotic. P value adjustment: Tukey method for comparing a family of 4 estimates. Effect size = Cohen's *d*.

Estimated Marginal Means: Display Format

Table S3: Estimated marginal mean differences for display format for each outcome measure.

	contrast	difference estimate	SE	p.value	lower 95%CI	upper 95%CI	effect size
	Variation						
1	table - bar graph	0.08	0.42	0.98	-0.91	1.07	0.01
2	table - icon array	0.12	0.42	0.95	-0.86	1.11	0.02
3	bar graph - icon array	0.05	0.42	0.99	-0.95	1.04	0.01
	Subjective evaluations						
1	table - bar graph	0.04	0.04	0.59	-0.06	0.14	0.10
2	table - icon array	-0.01	0.04	0.98	-0.10	0.09	-0.02
3	bar graph - icon array	-0.05	0.04	0.49	-0.15	0.05	-0.12
	Certainty perceptions						
1	table - bar graph	0.07	0.05	0.40	-0.05	0.19	0.17
2	table - icon array	0.01	0.05	0.98	-0.11	0.13	0.03
3	bar graph - icon array	-0.06	0.05	0.54	-0.18	0.07	-0.14
	Trustworthiness						
1	table - bar graph	0.07	0.05	0.31	-0.04	0.18	0.16
2	table - icon array	0.02	0.05	0.94	-0.10	0.13	0.04
3	bar graph - icon array	-0.05	0.05	0.50	-0.17	0.06	-0.12
	Behavioral intentions						
1	table - bar graph	0.22	0.14	0.26	-0.11	0.55	0.11
2	table - icon array	0.01	0.14	1.00	-0.32	0.33	0.00
3	bar graph - icon array	-0.21	0.14	0.29	-0.54	0.12	-0.11
		Odds Ratio	SE	p.value	lower 95%CI	upper 95%CI	
	Recall						
1	table / bar graph	0.91	0.21	0.92	0.53	1.58	
2	table / icon array	1.42	0.33	0.28	0.83	2.44	
3	bar graph / icon array	1.55	0.36	0.14	0.90	2.69	

.

Results are averaged over the levels of: uncertainty type. Degrees-of-freedom method: asymptotic. P value adjustment: Tukey method for comparing a family of 4 estimates. Effect size = Cohen's *d*.

Secondary Analyses: Linear mixed effects models excluding *lack of evidence*

Table S4: Results of linear mixed effects models on variation, subjective evaluations, certainty perceptions, trust, and behavioral intentions. (excluding *lack of evidence* condition)

	Variation[a]	Subjective evaluations[b]	Certainty perceptions[b]	Trust[b]	Behavioral intentions[c]
	Estimate [95%CI]	Estimate [95%CI]	Estimate [95%CI]	Estimate [95%CI]	Estimate [95%CI]
Fixed effects					
Intercept	9.17 [7.96; 10.38]	4.25 [4.14; 4.37]	4.13 [3.98; 4.28]	4.07 [3.92; 4.22]	8.72 [8.30; 9.12]
Uncertainty type					
Point estimate	referent	referent	referent	referent	referent
Imprecision	0.97 [0.07; 1.85]	0.00 [−0.09; 0.09]	0.02 [−0.09; 0.14]	0.01 [−0.09; 0.13]	−0.25 [−0.55; 0.06]
Conflicting estimates	0.84 [−0.03; 1.71]	−0.05 [−0.15; 0.04]	−0.10 [−0.21; 0.01]	−0.13 [−0.24; −0.02]	−0.33 [−0.62; −0.01]
Display format					
Table	referent	referent	referent	referent	referent
Bar graph	−0.04 [−0.91; 0.78]	−0.06 [−0.14; 0.03]	−0.09 [−0.21; 0.02]	−0.11 [−0.22; 0.00]	−0.27 [−0.59; 0.03]
Icon array	0.25 [−0.65; 1.07]	−0.03 [−0.12; 0.06]	−0.03 [−0.14; 0.08]	−0.05 [−0.15; 0.07]	0.06 [−0.25; 0.38]
Covariates					
Graph literacy[d]	−1.99 [−2.31; −1.66]	0.02 [−0.01; 0.05]	−0.13 [−0.17; −0.08]	−0.08 [−0.12; −0.04]	−0.20 [−0.31; −0.08]
Numeracy[e]	−0.59 [−0.92; −0.28]	0.02 [−0.01; 0.06]	−0.01 [−0.06; 0.03]	0.04 [0.00; 0.08]	0.03 [−0.08; 0.15]
Treatment effect size					
Moderate effect	referent	referent	referent	referent	referent
Small effect	−0.60 [−0.98; −0.25]	−0.08 [−0.11; −0.05]	−0.10 [−0.13; −0.07]	−0.05 [−0.08; −0.02]	−1.67 [−1.81; −1.53]
Random effects					
Intercept σ^2	30.65	0.37	0.62	0.56	3.77
Residual	23.40	0.13	0.15	0.19	3.25

Note: referent = reference group; Estimates are unstandardized coefficients. Square brackets indicate bootstrapped confidence intervals. [a]Mean absolute deviations from each group's median estimate; [b]5-point Likert scales ranging from 'not at all'–'very'. [c]11-point Likert scale 'I would definitely not – definitely take [medication]. [d]Scores range 0-4; higher scores indicate greater graph literacy. [e]Scores range 1-4; higher scores indicate greater numeracy.

Recall analyses

Table S5: Results of general linear mixed effects models for recall, including the *lack of evidence* condition (left) and excluding *lack of evidence* condition (right)

	Recall[a]: including *lack of evidence* Odds Ratio [95%CI]	Recall[a]: excluding *lack of evidence** Odds Ratio [95%CI]
Fixed effects		
Intercept	5.32 [6.00; 59.18]	4.16 [8.11; 309.40]
Uncertainty type		
Point estimate	referent	referent
Imprecision	0.71 [0.42; 1.25]	0.70 [0.41; 1.26]
Conflicting estimates	0.59 [0.34; 1.09]	0.58 [0.36; 1.07]
Lack of evidence	0.31 [0.17; 0.54]	-
Display format		
Table	referent	referent
Bar graph	1.09 [0.64; 1.69]	1.30 [0.72; 2.12]
Icon array	0.70 [0.42; 1.12]	0.91 [0.53; 1.56]
Covariates		
Graph literacy[b]	3.30 [2.80; 4.28]	3.36 [2.12; 3.87]
Numeracy[c]	1.29 [1.03; 1.49]	1.35 [1.03; 1.54]
Recall value		
Recall placebo	referent	referent
Recall treatment	0.43 [0.30; 0.49]	0.61 [0.40; 0.73]
Random effects		
Intercept σ^2	8.19	9.72

Note: referent = reference group; Estimates are odds ratios and square brackets indicate bootstrapped confidence intervals. [a]Recall of treatment and placebo numbers for both medications were coded such that each participant received a binary score of correct (vs. incorrect) for each of the four numbers (two medications each with a treatment value and placebo value). [b]Scores range 0-4 with higher scores indicating greater graph literacy. [c]Scores range from 1-4 with higher scores indicating greater numeracy. *A model including treatment effect did not improve model fit ($\chi^2(1)=1.96$, $p=.161$).