

# How Difference Tasks Are Affected by Probability Format, Part 1: A Making Numbers Meaningful Systematic Review

Natalie C. Benda , Brian J. Zikmund-Fisher , Mohit M. Sharma, Stephen B. Johnson, Michelle Demetres, Diana Delgado, and Jessica S. Ancker 

## Abstract

**Background.** To develop guidance on the effect of data presentation format on communication of health probabilities, the Making Numbers Meaningful project undertook a systematic review. **Purpose.** This article, one in a series, covers evidence about “difference tasks,” in which a reader examines a stimulus to evaluate differences between probabilities, such as the effect of a risk factor or therapy on the chance of a disease. This article covers the effect of format on 4 outcomes: 1) identifying a probability difference (identification) or recalling it (recall), 2) identifying the largest or smallest of a set of probability differences (contrast outcome), 3) placing a probability difference into a category such as “elevated” or “below average” (categorization outcome), and 4) performing computations (computation outcome). **Data Sources.** MEDLINE, Embase, CINAHL, the Cochrane Library, PsycINFO, ERIC, ACM Digital Library; hand search of 4 journals. **Finding Selection.** Pairwise screening to identify experimental/quasi-experimental research comparing 2 or more formats for quantitative health information. This article reports on 53 findings derived from 35 unique studies reported in 32 papers. **Data Extraction.** Pairwise extraction of information on stimulus (data in a data presentation format), cognitive task, and perceptual, affective, cognitive, or behavioral outcomes. **Data Synthesis.** Most evidence involving outcomes of difference-level cognitive tasks was weak or insufficient. Evidence was strong that 1) computations involving differences are easier with rates per  $10^n$  than with percentages or 1 in X rates and 2) adding graphics to numbers makes it easier to perform difference-level computations. **Limitations.** A granular level of evidence syntheses leads to narrow guidance rather than broad statements. **Conclusions.** Although many studies examined differences between probabilities, few were comparable enough to generate strong evidence.

## Highlights

- Most evidence about the effect of format on ability to evaluate differences in probabilities was weak or insufficient because of too few comparable studies.
- Strong evidence showed that computations relevant to differences in probabilities are easier with rates per  $10^n$  than with 1 in X rates.
- Adding graphics to probabilities helps readers compute differences between probabilities.

## Keywords

Systematic Reviews, Evidence Synthesis, Decision Aids, Risk Communication, Risk Perception, Shared Decision Making, Numeracy, Health Literacy, Physician-Patient Communication

Date received: February 29, 2024; accepted: September 9, 2024

## Corresponding Author

Jessica S. Ancker, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave, Suite 1475, Nashville, TN 37203, USA; (jessica.s.ancker@vumc.org).



This Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

**Table A** Current Articles' Scope within the Making Numbers Meaningful Systematic Review

Outcomes	Section <sup>a</sup>	Probability				
		Tasks				
		Point	Difference	Trend	Synthesis	Synthesis Bayesian
Identification or recall	1		This article			
Categorization	2		This article			
Contrast	3		This article			
Computation	4		This article			
Probability perceptions or feelings	5					
Effectiveness perceptions or feelings	6					
Behavioral intention or behavior	7					
Trust	8					
Preference	9					
Discrimination	10					

<sup>a</sup>This standardized numbering system has been used for results subheadings in this article and across all Making Numbers Meaningful results articles to ensure that readers can find comparable information in all articles. Gray cells represent combinations that are not possible according to the definitions presented in Ancker et al.<sup>1</sup>

Patients need numbers to make informed decisions on the basis of the probabilities of health and disease. One important communication challenge is how to express the difference between 2 probabilities, such as the effect of a risk factor or a therapy. The difference between 2 probabilities—the effect size<sup>3</sup> is key information that can help patients choose between therapies, determine whether to avoid risk factors or exposures, and make other decisions involving interpreting effect sizes. The difference between 2 probabilities can be formatted in multiple ways. Common approaches include a pair of individual probabilities (an increase from a 3% risk to a 4.5% risk), a relative difference between probabilities (a

50% relative increase, a relative risk of 1.5), an absolute difference between probabilities (a 1.5-percentage-point absolute increase), or a combination of several of these formats.

As described previously,<sup>1,2</sup> our systematic literature review collected evidence on how to communicate health-related numbers across data types and across different data presentation formats. We organized the literature according to a conceptual model of communication in which a reader views a stimulus, performs cognitive tasks to make sense of it, and experiences cognitive, affective, perceptual, or behavioral responses that are measured with outcome measures.

This article presents the subset of evidence pertaining to probability data and difference cognitive tasks requiring audience members to evaluate differences between 2 or more probabilities. Difference cognitive tasks are important for understanding the effects of therapies or risk factors. Readers may perform difference-level tasks upon absolute probabilities (for example, a pair of absolute probabilities expressing the chance of disease before and after vaccination) or upon precalculated probability comparisons (e.g., an absolute probability difference, a relative risk, or relative risk reduction [RRR]). (Additional articles in this series, as elaborated in Table A, cover point tasks, in which readers seek information about individual probabilities, synthesis tasks, in which the reader integrates several probabilities such as the set of risks and benefits for a therapeutic option, synthesis-Bayesian tasks involving interpreting probability information to estimate Bayesian posterior probabilities, and

Columbia University School of Nursing, New York, NY, USA (NCB); Department of Health Behavior and Health Education, University of Michigan, Ann Arbor, MI, USA (BJZ-F); Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA (BJZ-F); Center for Bioethics and Social Sciences in Medicine, University of Michigan, Ann Arbor, MI, USA (BJZ-F); Department of Population Health Sciences, Weill Cornell Medical College, New York, NY, USA (MMS); Department of Population Health, New York University Langone Health, New York, NY, USA (SBJ); Samuel J Wood Medical Library, Weill Cornell Medical College, New York, NY, USA (MD, DD); Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA (JSA). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support for this study was provided entirely by a grant from the National Library of Medicine (R01 LM012964, Ancker PI). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the reports.

time-trend tasks, in which readers examine stimuli to evaluate patterns over time.)

To keep article length manageable, the difference task research has been divided into a pair of articles. The current article (part 1) presents evidence on 4 outcomes: 1) identifying a probability difference in the stimulus (termed *identification*) or recalling it (termed *recall*), 2) identifying the largest or smallest of a set of probability differences (termed *contrast*), 3) placing a probability difference into a meaningful category such as “elevated” or “below average” (termed *categorization*), and 4) performing computations on probability differences such as subtracting one probability from another or converting a percentage into a number out of 100 (termed *computation*). Although the first 2 of these (identification and recall) would appear to be different, we grouped them (as described below) because of a frequent lack of clarity in the research about which was being measured. (A companion article, part 2, presents evidence on the remaining outcomes.) Here, we include evidence on the effects of all data presentation formats—numbers, graphics, and verbal probabilities—on these outcomes.

## Methods

Methods for the literature search, screening, risk-of-bias evaluation, data extraction, credibility evaluation of findings, and organization into evidence tables are reported in detail in our companion methods article.<sup>2</sup> In brief, a broad search was performed to find experimental or quasi-experimental (nonrandomized) research comparing 2 or more ways of presenting quantitative health-related data to patients or other lay audiences. The search was performed on MEDLINE, Embase, CINAHL, the Cochrane Library, PsycINFO, ERIC, and ACM Digital Library, and we conducted hand searches of tables of contents of *Medical Decision Making*, *Patient Education and Counseling*, *Risk Analysis*, and *Journal of Health Communication*. Substudies in the same article were extracted separately. All instruments used (search strategy, data extraction instrument, and study risk of bias rubric) are available at the Making Numbers Meaningful Project at the Open Science Framework site (<https://osf.io/rvxf2/>).

As described in detail elsewhere,<sup>2</sup> our literature review identified 316 articles, of which 32 (representing 35 unique studies) involved difference-level tasks with probability data that reported 1 or more of the 4 above outcomes.

We assigned each included study a study risk-of-bias (S-ROB) score according to a rubric developed for this

project, which considered sample representativeness, randomization, protocol deviations, presence/absence of demographic and covariate information, missing data, and other potential biases. Within each included study, we extracted information about task, stimulus (data and data presentation format), and outcome. The outcomes were informed by behavioral and risk communication theory (behavior or behavioral intention, probability perceptions or feelings, recall) or empirically on the basis of what was frequently measured by the research included in our review (trust, preference for a format), particularly measures used to measure comprehension (identification, contrast, computation, categorization, discrimination).

Each unique combination of task, format comparison, and outcome from a single study was termed a *finding*. We grouped findings by task and outcome, rated their credibility, and synthesized them into guidance statements. Credibility for each finding was assessed on a scale from 1 to 10 by pairs of authors (N.C.B., J.S.A., B.J.Z.-F.) using a holistic assessment of sample size, statistical methods, validity of stimulus design, comparison, outcome measures, and covariates (either criterion validity or face validity), plus the S-ROB for the study from which the finding came. Credibility of 7 or higher was considered high, 4.5 to 6.5 moderate, and 4 or lower low. Using the group of relevant findings, we then applied a standard rubric to grade the strength of evidence for each guidance statement according to finding credibility and consistency. Consistency was considered high if all findings were significant in the same direction or if a large majority were significant in one direction with a few lacking in significance, moderate if findings showed a small majority of significant effects in one direction with the remainder lacking significance, and low if the findings showed significant effects in different directions. Findings with high credibility (7 or higher on a scale of 1 to 10) and moderate credibility (4.5–6.5) are discussed below. Findings with lower credibility (4 or lower) are mentioned below, counted in Table B, and listed in our Findings tables, but they do not contribute to the evidence summaries or the statements in the evidence tables.

- **Strong:** High consistency within group of 2 or more high-credibility findings or a mix of high- and moderate-credibility findings
- **Moderate:** a) High consistency within a group of 2 or more moderate-credibility findings or b) moderate consistency within 2 or more findings of which at least 1 was high credibility and the others moderate credibility

**Table B** Section Headings for Each Subset of Outcome Evidence Included in This Article and the Number of Included Findings

Subsections  Data Presentation Format Comparison	Section Number/Subsection Letter	Sections			Total Findings per Data Presentation Format Comparison
		Identification or Recall	Contrast	Computation	
		1	2	4	
Comparisons between numerical formats	A	1A ( <i>n</i> = 8)	2A ( <i>n</i> = 7)	4A ( <i>n</i> = 9)	24
Comparisons between graphical formats	B	1B ( <i>n</i> = 3)	2B ( <i>n</i> = 5)	4B ( <i>n</i> = 4)	12
Comparisons between numerical and graphical formats	C	1C ( <i>n</i> = 4)	2C ( <i>n</i> = 3)	4C ( <i>n</i> = 3)	10
Comparisons of elements added for context	E	1E ( <i>n</i> = 1)	2E ( <i>n</i> = 1)	4E ( <i>n</i> = 0)	2
Comparisons of frames (gain, loss, or combination)	F	1F ( <i>n</i> = 1)	2F ( <i>n</i> = 1)	4F ( <i>n</i> = 1)	3
Comparisons of animation or interactivity	I	1I ( <i>n</i> = 1)	2I ( <i>n</i> = 1)	4I ( <i>n</i> = 0)	2
Total findings per outcome		18	18	17	53

1. No relevant findings for the following comparisons: numbers versus verbal probabilities (row D), uncertainty (row G), larger or smaller denominators (row H), longer or shorter time periods (row J).

2. No relevant findings for the following outcomes: categorization (section 3).

3. The standardized numbering system in Table B has been used for the subheadings of all Making Numbers Meaningful results articles. The numbers ensure that, for example, studies of the effects of gain-loss framing manipulations on computation are always placed in a subhead labeled subsection 4F (whether or not that article contains sections 1 through 3). Our goal is to ensure that readers can use this subhead system to locate similar sections across articles.

- **Weak:** Moderate consistency within group of 2 or more moderate-credibility findings or only a single high-credibility finding
- **Insufficient evidence—too few findings:** a) Only low-credibility findings available or b) only 1 moderate-credibility finding
- **Insufficient evidence—conflicting findings:** Any case in which evidence consistency was low

We have described our terminology in our previous articles, but for the current article, it is important to note several terms. For probabilities, we distinguished between 2 types of rates: those formatted as 1 in X (examples include “1 in 5” and “1 of every 25”) and those formatted as a rate per 10<sup>n</sup> (such as “12 in 100” or “2.5 per 1,000”). Also, we reserved the *natural frequency* label for a series of joint probabilities and conditional probabilities computed from the same pool of patients, in the context of Bayes’ theorem.<sup>3,4</sup> This definition matches the original term definition,<sup>4</sup> and using the term only for this purpose helps clarify otherwise apparently contradictory findings.<sup>3</sup>

## Results

No articles measured *categorization* outcomes related to probability differences. We grouped studies examining *identification and recall* outcomes because it was often unclear whether study participants could see the relevant stimuli when answering these questions.

We therefore summarize the findings related to *identification/recall*, *contrast*, and *computation* outcomes below (Table B). The full spreadsheet of difference task findings is available in the “Probability Findings” folder in the Making Numbers Meaningful Project at OSF (<https://osf.io/rvxf2/>).

Each results subsection summarizes evidence on the following comparisons in order: comparisons among number formats, among graphics formats, between number and graphic formats, between different types of contextual elements, between different framings, and effect of animation or interactivity. Within subsections, evidence is arranged from strongest to weakest. No studies examining the communication of probability differences examined comparisons between number and verbal formats, effect of representations of uncertainty, effect of

**Table 1A** Evidence-Based Guidance for Effects of Numerical Formats on Identification/Recall of Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Rate per 10 <sup>n</sup> v. percentages	Weak ( <i>n</i> = 1)	It may be easier to answer questions about the difference between 25% and 30% than about the difference between 25-in-100 and 30-in-100 or the difference between 25% (25-in-100) and 30% (30-in-100).	People may have an easier time answering questions about probability differences when given pairs of percentages alone rather than rates per 10 <sup>n</sup> (with either constant or variable denominators) or combinations of rates and percentages.
Adding rate per 10 <sup>n</sup> to difference	Weak ( <i>n</i> = 1)	When told their probability of disease is 5% higher than average, people's ability to answer questions about the probability difference may not be affected by also telling them their probability is 30% (30-in-100) as compared with the average probability of 25% (25-in-100).	The ability to answer questions about of probability differences may not be affected by adding pairs of absolute rates per 10 <sup>n</sup> to the absolute difference.
Tables v. text	Insufficient evidence— inconsistent findings ( <i>n</i> = 3)	It is not clear whether the ability to answer questions about probability differences is better with pairs of rates per 10 <sup>n</sup> in tables or embedded in text.	
Life expectancy	Insufficient evidence— too few findings ( <i>n</i> = 1)	It is not clear whether recall of probability differences is different for information presented as life expectancies than presented as a baseline percentage and absolute risk difference as a percentage.	
Instructions	Insufficient evidence— too few findings ( <i>n</i> = 1)	It is not clear whether identification of probability differences presented in table formats is affected by adding “how to read this table” instructions.	

manipulations of denominators, or manipulations of time periods.

### *Effects of Different Formats for Probability Trends on Ability to Identify or Recall Information (Identification/Recall Outcome): Section 1*

Researchers often assessed comprehension by asking questions about probability differences in the stimulus. However, published studies were frequently unclear about whether the stimulus was available or removed when the questions were presented and thus whether the outcome was assessing ability to *identify* the numbers or ability to *recall* them. Given this ambiguity, we combined these 2 outcomes in this section. However, for studies that had clear distinctions between identify and recall outcomes, we reflect the distinction in the summaries and guidance statements.

*Comparisons between number formats on identification/recall of probability differences (subsection 1A). RATES*

**PER 10<sup>n</sup> VERSUS PERCENTAGES:** In a high-credibility finding from a large study, Woloshin and Schwartz<sup>5</sup> found the ability to answer questions about probability differences was better when pairs of probabilities were presented as percentages in a drug facts box table than when as rates per 10<sup>n</sup> (either fixed denominator or variable denominator) or combinations of percentages and rates.

**ADDING RATE PER 10<sup>n</sup> TO DIFFERENCE:** In a high-credibility finding, Sullivan et al.<sup>6</sup> found no difference in ability to answer questions about numbers when pairs of rates (percentages plus rates per 10<sup>n</sup>) were added to arithmetic differences in a table, although the presence of verbal labels in addition to the absolute difference complicates interpretation of this negative finding.

**TABLES VERSUS TEXT:** A high-credibility finding (Tait et al.<sup>7</sup>) showed no significant differences in ability to answer questions about differences in numbers of people affected by a drug when pairs of rates per 10<sup>n</sup> were presented in table or text format. However, a moderate-credibility finding from a different study by the same author team (Tait et al.<sup>8</sup>) found that ability to answer questions about differences was better with rates per 10<sup>n</sup>

**Table 1B** Evidence-Based Guidance for Effects of Graphical Formats on Identification/Recall of Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Icon arrays	Weak ( $n = 1$ )	People's ability to answer questions about a probability difference in an incremental icon array (one that shows the arithmetic difference between 2 treatments) may not be affected by whether the array highlights and labels the number of people who survived and the number who died from more than 1 cause or highlights and labels only the number of people who survived.	Ability to answer questions about probability differences may be similar for combination-framed icon arrays showing incremental differences in multiple outcomes and gain-framed icon arrays highlighting only incremental survival outcomes.
Other graphics v. pie chart	Insufficient evidence—too few findings ( $n = 1$ )	It is not clear whether specific graphics affect people's ability to identify or recall probability differences.	

in a table versus rates per 10<sup>n</sup> in text. Mühlbauer et al.<sup>9</sup> found mixed findings when comparing text to table.

**LIFE EXPECTANCY:** In a moderate-credibility finding (Galesic and Garcia Retamero<sup>10</sup>), short- and long-term recall of the effect of a risk factor was higher when information was presented in life expectancy terms than with a baseline percentage and absolute risk difference as a percentage.

**INSTRUCTIONS:** A moderate-credibility finding (Mühlbauer et al.<sup>9</sup>) showed that identification of probability differences was improved when a drug facts box table format was accompanied by a “how to read this table” instructions.

An additional recall finding was not summarized due to a floor effect, specifically very poor recall at 1 mo across formats.<sup>11</sup>

*Comparisons between graphic formats on identification/recall of probability differences (subsection 1B).* **ICON ARRAYS:** A high-credibility finding from Zikmund-Fisher et al.<sup>12</sup> substudy 1 found no differences in the ability to answer questions about probability differences between combination-framed icon arrays showing incremental differences in survival and mortality outcomes versus gain-framed icon arrays only highlighting incremental survival outcomes.

**OTHER GRAPHICS VERSUS PIE CHART:** A moderate-credibility finding (Hawley et al.<sup>13</sup>) showed that certain graphical formats (icon array, bar chart, number line) appeared to help people answer questions about the number of people affected or probability differences better than pie charts did (with or without

circular axis labels). However, this study mixed questions about individual probabilities and probability differences, making it impossible to assess a unique effect for differences.

A low-credibility finding from Schonlau and Peters<sup>14</sup> substudy 2 was not summarized due to methodological concerns, specifically the inability to make comparisons between graphical formats and determine statistically significant differences based on the results presented.

*Comparisons between numerical and graphical formats, and combinations of numerical and graphical formats, on identification/recall of probability differences (subsection 1C).* **VARIOUS GRAPHICS:** A high-credibility finding (Tait et al.<sup>7</sup>) found that icon arrays led to better ability to answer questions about numbers of people and differences in numbers of people affected by a drug than rates per 10<sup>n</sup> in table or text format. However, a moderate-credibility finding by the same author team (Tait et al.<sup>8</sup>) found that performance was better with rates per 10<sup>n</sup> in a table versus icon arrays or rates per 10<sup>n</sup> in text. Another moderate-credibility finding (Hawley et al.<sup>13</sup>) also found that a table of rates per 10<sup>n</sup> helped people answer questions about the number of people affected more effectively than a variety of graphical formats (pie charts were the worst). All of these studies mixed questions about individual probabilities versus probability differences, making it impossible to assess a unique effect for differences.

A low-credibility finding<sup>14</sup> was not summarized due to methodological concerns, specifically, the inability to identify statistically significant differences based on results presented.

**Table 1C** Evidence-Based Guidance for Contrasts between Numerical and Graphical Formats, and Combinations of Numerical and Graphical Formats, on Identification/Recall of Probability Differences

Comparison	Evidence Strength	General Guidance
Various graphics	Insufficient evidence— inconsistent findings ( $n = 3$ )	It is not clear whether numbers and graphics have different effects on people's ability to identify or recall probability effects.

**Table 1E** Evidence-Based Guidance for Effect of Adding Context on Identification/Recall of Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Difference labels	Weak ( $n = 1$ )	Adding the descriptor “more people are affected with drug A than drug B” to “drug A has an 8% probability and drug B has a 5% probability” may not improve the ability to answer questions about probability differences.	Adding labels to numbers describing which option has higher rates of each outcome may not affect people's ability to answer questions about probability differences.

*Comparisons of elements added for context on identification/recall of probability differences (subsection 1E).* **DIFFERENCE LABELS:** A high-credibility finding (Sullivan et al.<sup>6</sup>) found no difference in ability to answer questions about probability differences in a table resembling a drug facts box including pairs of absolute rates per 10<sup>n</sup> and/or absolute probability differences by whether the table did or did not include labels describing which drug had the higher rate of each outcome.

*Comparisons of frames (gain, loss, combination) on identification/recall of probability differences (subsection 1F).* **COMBINATION FRAME VERSUS GAIN FRAME:** A high-credibility finding (Zikmund-Fisher et al.<sup>12</sup> substudy 1) found no differences in ability to answer questions about probability differences between combination-framed icon arrays showing survival and mortality outcomes versus gain-framed icon arrays only highlighting survival outcomes.

*Comparisons of animation or interactivity on identification/recall of probability differences (subsection 1I).* **STATIC VERSUS ANIMATED GRAPHICS:** A moderate-credibility finding (Housten et al.<sup>15</sup>) did not find any differences in the ability to answer questions about probabilities by format (static icon arrays, icon arrays animated to draw attention to subsets, or icon

arrays with additional animation effects to emphasize randomness) but was limited by small sample size.

### *Effects of Different Formats on Ability to Identify Largest or Smallest of a Set of Numbers (Contrast Outcome): Section 2*

Responses to questions about identifying the largest or smallest in a list of differences or ranking the differences in order of size were considered *contrast* outcomes.

*Comparisons between numerical formats on ability to contrast probability differences (subsection 2A).* **PERCENTAGE PAIRS VERSUS RELATIVE RISK DIFFERENCE:** One high-credibility finding (Perneger and Agoritsas<sup>16</sup>) shows that ability to select the option (based on size of risk reduction) was better when the participant received the RRR as a raw singular number versus either positively or negatively framed pre/post percentage values or pre/post values plus the RRR.

**SUPPLEMENTING ABSOLUTE RATES PER 10<sup>n</sup> WITH DIFFERENCES:** In a moderate-credibility finding from Covey<sup>17</sup> substudy 1, pairs of rates per 10<sup>n</sup> supplemented with absolute risk reduction (ARR) or RRR increased the ability to select the largest difference as compared with pairs of percentages also supplemented with ARR and RRR. There were no significant differences between supplementing with ARR or with RRR.

**Table 1F** Evidence-Based Guidance for Effect of Framing on Identification/Recall of Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Combination frame v. gain frame	Weak ( $n = 1$ )	People's ability to answer questions about a probability difference in an incremental icon array (one that highlights the arithmetic difference between 2 treatments) may not be affected by whether the array highlights and labels the number of people who survived and the number who died from more than 1 cause or highlights and labels only the number of people who survived.	Ability to answer questions about probability differences may be similar for combination-framed icon arrays showing multiple outcomes versus gain-framed icon arrays highlighting only survival outcomes.

**Table 1I** Evidence-Based Guidance for Effect of Animation or Interactivity on Identification/Recall of Probability Differences

Comparison	Applied Example of Evidence-Based Communication	General Guidance
Static v. animated graphics	Insufficient evidence— too few findings ( $n = 1$ )	It is not clear whether static or animated graphics have different effects on identification or recall of probability differences.

Covey<sup>17</sup> substudy 2 produced the finding that rates per  $10^n$  supplemented with ARR information improved the ability to select the largest difference as compared with RRR as a percentage alone. For both of these studies, small samples of students reduce confidence in findings.

**RELATIVE VERSUS ABSOLUTE DIFFERENCE:** One high-credibility finding from Sheridan et al.<sup>18</sup> was that conveying RRR as a percentage led to an improved ability to select the largest difference as compared with conveying ARR as a percentage, number needed to treat (NNT), or a combination of all three.

**BASELINE PLUS ABSOLUTE DIFFERENCE VERSUS BASELINE PLUS RELATIVE DIFFERENCE:** One moderate-credibility finding, Lavallie et al.,<sup>19</sup> found that conveying baseline probability and ARR as a common denominator rate per  $10^n$  was superior to baseline and RRR as common denominator rate per  $10^n$  or to baseline common denominator rate per  $10^n$  with NNT.

**PERCENTAGES VERSUS RATES PER  $10^n$ :** One moderate-credibility finding (Wolfe et al.<sup>20</sup> substudy 1) found that using pairs of percentages (versus pairs of common denominator rate per  $10^n$ ) improved women's ability to identify 2 similar probabilities as "approximately equal."

A low-credibility finding<sup>11</sup> was not summarized because the single question assessing contrast outcomes from difference tasks was not reported separately from an aggregate knowledge measure.

*Comparisons between graphical formats on ability to contrast probability differences (subsection 2B).* **TYPES OF BAR CHARTS:** One high-credibility finding (Okan et al.<sup>21</sup> substudy 1) determined that selecting the option with the largest effect in a vertical bar chart improved when showing absolute difference as positive bars (percentage point improvement) as opposed to negative bars (percentage point reduction).

**ICON ARRAYS VERSUS PIE CHARTS VERSUS BAR CHARTS:** Tolbert et al.,<sup>22</sup> in a high-credibility finding, found that pairs of icon arrays and pie charts improved participants' ability to select the option with the largest effect over bar charts. However, the bar charts used in the study had more visual elements (3 bars), which may have made the interpretation task using the bar chart more cognitively complex. A moderate-credibility finding by Waters et al.<sup>23</sup> showed that in a description of the chance of benefit of a drug, the ability to recognize that the drug reduced the total probability



**Table 2A** Evidence-Based Guidance for Effects of Numerical Formats on Ability to Contrast Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Percentage pairs v. relative difference	Weak ( <i>n</i> = 1)	Telling people, “Taking drug A over drug B led to a 50% reduction in adverse events” may improve their ability to select the best option as compared with “10% experienced an adverse event with drug A, while 20% experienced an adverse event with drug B.”	People may be better able to contrast multiple treatment options with the relative risk difference than with pairs of pre/post percentages using positive, negative, or combination framing.
Supplementing absolute rates per 10 <sup>n</sup> with differences	Weak ( <i>n</i> = 2)	Telling people “4 in 10 of untreated persons will get a disease. Of those taking drug A, 2 in 10 will get the disease (a 2 in 10 risk reduction), while 1 in 10 of those taking drug B will get the disease (a 3 in 10 risk reduction)” may improve their ability to select the most effective option as compared with telling them “40% of untreated persons will get a disease. Of those taking drug A, 20% will get the disease (a 20% risk reduction), while 10% of those taking drug B will get the disease (a 30% risk reduction).”	People may be better able to contrast treatment options with pairs of common denominator rates per 10 <sup>n</sup> than with percentages. This has been demonstrated specifically for pre-post information supplemented with absolute risk reduction or relative risk reduction.
Relative v. absolute difference	Weak ( <i>n</i> = 1)	Telling people “drug A results in a 25% risk reduction; drug B results in a 50% risk reduction” may improve their ability to select the best option as compared with “10 patients would have to receive drug A for 1 additional patient to NOT contract the disease; 5 patients would have to receive drug B for 1 additional patient to NOT contract the disease” OR “drug A results in an improvement of 10 percentage points while drug B results in an improvement of 20 percentage points.”	People may be better able to contrast treatment options with a relative risk difference as a percentage than with an absolute risk difference as a percentage, number needed to treat, or a combination of the 3 formats.
Baseline plus absolute difference v. baseline plus relative difference	Insufficient evidence—too few findings ( <i>n</i> = 1)	It is not clear whether providing different statistics (baseline + absolute risk reduction, baseline + relative risk reduction, or baseline + number needed to treat) affects people’s ability to select a normatively dominant treatment option.	
Percentages v. rates per 10 <sup>n</sup>	Insufficient evidence—too few findings ( <i>n</i> = 1)	It is not clear whether using pairs of percentages versus pairs of rates per 10 <sup>n</sup> affects people’s ability to identify 2 similar probabilities as approximately equal.	

of disease was similar with icon arrays and with bar charts.

**GROUPED VERSUS RANDOM ICON ARRAYS:** A moderate-credibility finding (Wright et al.<sup>24</sup>) found that the ability to select the largest or smallest effects of smoking on disease outcomes shown in icon array graphics was similar whether the icon array was grouped or random.

**BAR CHARTS ILLUSTRATING ABSOLUTE DIFFERENCES OR RISK RATIOS:** A moderate-

credibility finding (Harper et al.<sup>25</sup>) found a higher ability to determine whether population-level disparities increased, decreased, or stayed the same when grouped bar charts showed risk ratios instead of absolute differences.

*Comparisons between numerical and graphical formats, and combinations of numerical and graphical format, on ability to contrast probability differences (subsection 2C). BAR CHART WITH OR WITHOUT A*

**Table 2B** Evidence-Based Guidance for Effects of Graphical Formats on Ability to Identify the Dominant Option

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Bar charts	Weak ( $n = 1$ )	For the difference between a 10% probability and a 20% probability of disease, people may have an easier time selecting the biggest difference with a bar chart showing a 10% absolute improvement than a bar chart showing a 10% absolute reduction.	People's ability to contrast effect sizes may be better with a bar chart of percentage point improvement than a bar chart of percentage point reduction.
Icon arrays v. pie charts v. bar charts	Insufficient evidence—conflicting findings ( $n = 2$ )	It is not clear whether icon arrays, pie charts, or bar charts are better for helping people identify the biggest difference.	
Grouped v. random icon arrays	Insufficient evidence—too few findings ( $n = 1$ )	It is not clear whether grouped or random icon arrays are better for people's ability to identify the largest or smallest probability difference.	
Bar charts illustrating absolute differences or risk ratios	Insufficient evidence—too few findings ( $n = 1$ )	It is not clear whether it is easier to determine whether a difference increased, decreased, or remained the same with bar charts of risk ratios or bar charts of absolute differences.	

**Table 2C** Evidence-Based Guidance for Contrasts between Numerical and Graphical Formats, and Combinations of Numerical and Graphical Formats, on Ability to Contrast Probability Differences

Comparison	Evidence Strength	General Guidance
Bar chart with or without a percentage	Insufficient evidence—too few findings ( $n = 1$ )	It is not clear whether adding graphics to numbers affects ability to identify 2 similar probabilities as approximately equal.
Icon array v. numbers alone	Insufficient evidence—too few findings ( $n = 1$ )	It is not clear whether icon arrays or numbers are better for helping people identify the larger of 2 probability differences.

**Table 2E** Evidence-Based Guidance for Effect of Adding Context on Ability to Contrast Probability Differences

Comparison	Evidence Strength	General Guidance
Instructions	Insufficient evidence—too few findings ( $n = 1$ )	It is not clear whether adding instructions that contain the gist of the message affects ability to identify 2 similar probabilities as approximately equal.

**PERCENTAGE:** A moderate-credibility finding (Wolfe et al.<sup>20</sup> substudy 2) found that adding a bar chart to a percentage did not improve women's ability to identify 2 similar probabilities as "approximately equal."

**ICON ARRAYS VERSUS NUMBERS:** A moderate-credibility finding by Waters et al.<sup>23</sup> showed that in a description of the chance of benefit of a drug, the ability to recognize that the drug reduced the total probability of disease was better with icon arrays than with percentages alone.<sup>23</sup>

One finding by Silk and Parrott<sup>26</sup> not synthesized had lower credibility due to limited comparability of the information in the graphical and numerical arms.

*Comparisons of elements added for context on ability to contrast probability differences (subsection 2E).* **INSTRUCTIONS:** A moderate-credibility finding (Wolfe et al.<sup>20</sup> substudy 2) found that adding "gist-evoking instructions" to percentages (with or without a bar

**Table 2F** Evidence-Based Guidance for Effect of Framing on Ability to Contrast Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Framing percentages	Weak ( <i>n</i> = 1)	It may improve people's ability to select the dominant option if you tell them "20% of people died with drug A and 10% of people died with drug B" versus telling them "80% of people survived with drug A and 90% of people survived with drug B."	People may be better able to contrast multiple options with pairs of negatively framed percentages (with or without positively framed percentages) than with pairs of positively framed ones alone.

**Table 2I** Evidence-Based Guidance for Effect of Animation or Interactivity on Ability to Contrast Probability Differences

Comparison	Evidence Strength	General Guidance
Animated v. static icon arrays	Insufficient evidence—too few studies ( <i>n</i> = 1)	It is not clear whether animation or interactivity in presentation formats affects people's ability to select the normatively dominant option from among multiple options.

**Table 4A** Evidence-Based Guidance for Effect of Numerical Format on Ability to Perform Computations on Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Rates per 10 <sup>n</sup> v. other numeric formats	Strong ( <i>n</i> = 5)	People have an easier time computing a probability difference with a statement such as "the baseline probability for disease A is 25 in 100, but if you take treatment B, the probability is 5 in 100" instead of "the baseline probability for disease A is 25%, but if you take treatment B, the probability is 5%" OR "the baseline probability for disease A is 1 in 4, but if you take treatment B, the probability is 1 in 20."	Providing effect sizes and baseline probabilities as rates per 10 <sup>n</sup> , as compared with other formats such as percentages or 1 in X, may improve peoples' ability to compute relationships between numbers. However, the effect probably depends on what computation they are asked to perform, and it is preferable not to ask people to perform computations.
Absolute v. relative probability	Weak ( <i>n</i> = 2)	People may be able to determine the absolute risk difference more accurately if you tell them, "10% of people die of this condition, but the treatment reduces the probability to 5%," instead of "10% of people die of this condition, but the treatment reduces the probability by 50%."	Providing a pair of pre-post percentages rather than the relative risk reduction as a percentage may improve peoples' ability to compute certain relationships between the numbers. However, the effect probably depends on what computation they are asked to perform, and it is preferable not to ask people to perform computations.
Number needed to treat	Insufficient evidence—too few findings ( <i>n</i> = 1)	It is not clear whether providing number needed to treat in combination with other effect measures affects people's ability to perform computations.	
Table v. text	Insufficient evidence—too few findings ( <i>n</i> = 1)	It is not clear whether presenting risks and benefits as pre/post pairs of rates per 10 <sup>n</sup> in a drug fact box versus in sentences affects people's ability to perform difference computations.	

**Table 4B** Evidence-Based Guidance for Effect of Graphical Format on Ability to Perform Computations on Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Icon arrays v. bar charts	Weak ( <i>n</i> = 2)	If people are expected to compute probability differences, it may not matter if they are given a pair of part-to-whole icon arrays (one showing 12 affected individuals out of 100, and the other showing 8 affected individuals) or a pair of part-to-whole bars (one showing a 12% probability of disease and the other showing 8%).	Ability to compute probability differences may be similar with pairs of part-to-whole icon arrays and part-to-whole bar charts.
Part-to-whole v. foreground-only graphics	Weak ( <i>n</i> = 1)	People may have an easier time computing the difference between a 4% and a 10% probability with a pair of icon arrays showing 4 in 100 and 10 in 100 than with a pair of icon arrays showing the 4 people affected in the first group and the 10 affected in the second.	People may have an easier time computing probability difference with graphics depicting both the numerators and denominators of the probabilities than with graphics that show only the numerators (numbers affected).

chart) improved women's ability to identify 2 similar probabilities as "approximately equal." The "gist-evoking" instructions were labeled, "When is a difference really a difference?" and provided examples of small differences that were meaningful (a millimeter movement during a surgical procedures) and were not (a millimeter difference in 2 peoples' heights).

*Comparisons of frames (gain, loss, combination) on ability to contrast probability differences (subsection 2F).* **FRAMING PERCENTAGES:** One high-credibility finding (Perneger and Agoritsas<sup>16</sup>) found that ability to select the normatively dominant option was better when the difference comparisons were presented either as a pair of negatively framed percentages or a combination frame (percentage dying plus percentage surviving plus RRR) versus as a pair of positively framed percentages.

*Comparisons of animation or interactivity on ability to contrast probability differences (subsection 2I).* **ANIMATED VERSUS STATIC ICON ARRAYS:** One moderate-credibility finding (Housten et al.<sup>15</sup>) found no effect of animation (of icon arrays specifically) in participants' ability to detect the lowest or highest probability difference value. Interactivity was not assessed. The small sample reduces the confidence in the negative finding.

#### *Effects of Different Formats on Ability to Perform Computations on Probability Differences (Computation Outcome): Section 4*

Although it is not generally desirable to make readers perform computations, and best practice in health literacy is to perform computations for the reader,<sup>27</sup> there are some situations in which the correct calculation cannot be performed for every reader. Researchers therefore sometimes assess participants' ability to perform computations as a measure of comprehension. Although the ability to perform the computations is influenced by numeracy, it also provides information about the clarity and ease of use of the data presentation format.

*Comparisons of numerical formats on ability to perform computations on probability differences (subsection 4A).* **RATES PER 10<sup>n</sup> VERSUS OTHER NUMERIC FORMATS:** Five findings contrasted rates per 10<sup>n</sup> and other number formats. In 4 high-credibility findings, rates of correct responses were higher when baseline probability or all numbers were given as rates per 10<sup>n</sup> rather than percentages or 1 in X (Cuite et al.<sup>28</sup> substudy 2, Bodemer et al.<sup>29</sup> substudy 1 and 2, Schwartz et al.<sup>30</sup>). However, 1 moderate-credibility finding from a small study found no difference between rates per 10<sup>n</sup> and percentages (Koo et al.<sup>31</sup>).

**ABSOLUTE VERSUS RELATIVE PROBABILITY:** Two moderate- to high-credibility findings showed that the accuracy of computing a probability

**Table 4C** Evidence-Based Guidance for Contrasts between Graphical and Numerical Formats, and Combinations of Numerical and Graphical Formats, on Ability to Perform Computations on Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Numbers with or without graphics	Strong ( $n = 3$ )	People can compute a probability difference more accurately when provided with a pair of icon arrays as well as a statement (e.g., 75% of untreated people recover and 90% of people who took drug A recover) than with either graphics alone or numbers alone.	Combining graphics and numbers to present probability differences, as compared with providing either graphics or numbers alone, improves people's ability to perform calculations of the effect size or difference. However, the effect probably depends on what computation they are asked to perform, and it is preferable not to ask people to perform computations.

difference was better with an absolute difference (pair of pre-post percentages) than RRR as percentage (Garcia Retamero et al.,<sup>32</sup> Schwartz et al.<sup>30</sup>).

NNT: Rates of correct answers were lower when NNT was provided alone or in combination with other metrics of effect size (Sheridan et al.,<sup>18</sup> moderate credibility).

TABLE VERSUS TEXT: One moderate-credibility finding by Brick et al.<sup>11</sup> was that computation ability improved when the risks and benefits were presented as pairs of pre-post rates per 10<sup>n</sup> in a drug facts box rather than as pairs of rates per 10<sup>n</sup> in sentence text.

*Comparisons between graphical formats on ability to perform computations on probability differences (subsection 4B).* ICON ARRAYS VERSUS BAR CHARTS: Two moderate-credibility findings (Garcia-Retamero et al.,<sup>33</sup> Garcia-Retamero et al.<sup>34</sup>) found the ability to perform computations to estimate differences was similar with pairs of part-to-whole icon arrays and pairs of part-to-whole bar charts.

PART-TO-WHOLE VERSUS FOREGROUND-ONLY GRAPHICS: A high-credibility finding showed that the ability to perform computations to estimate differences was higher with pairs of part-to-whole graphics (icon arrays or bar charts) than foreground-only ones (Garcia-Retamero and Galesic<sup>35</sup>).

A lower-credibility finding (Price et al.<sup>36</sup>) compared different icon array fill patterns but was insufficiently powered.

*Comparisons between numerical and graphical formats, and combinations of numerical and graphical formats, on ability to perform computations on probability differences (subsection 4C).* NUMBERS WITH OR WITHOUT GRAPHICS: Three moderate to high-credibility findings

from different studies by the same author team suggest that the combination of graphics and numbers is generally superior to numbers alone in supporting computations. In one of these (high credibility), part-to-whole icon arrays or bar charts plus the arithmetic difference were superior to numbers alone as well as to several other graphics (Garcia-Retamero and Galesic<sup>35</sup>). The remaining 2 (moderate- to high-credibility findings) (Garcia-Retamero et al.,<sup>34</sup> Garcia-Retamero et al.<sup>32</sup>) also found that icon arrays or bar charts performed better than numbers alone, but the effect may be strongest among or limited to those with high graph literacy.

*Comparisons of frames (gain, loss, or combination) on ability to perform computations on probability differences (subsection 4F).* SURVIVAL VERSUS MORTALITY CURVES: A moderate-credibility finding (Armstrong et al.<sup>37</sup>) found a greater ability to calculate the difference between 2 groups at a point in time when the data were presented in survival curves instead of mortality curves.

### Summary of Evidence

Both of the 2 **strong** evidence findings from this part of our systematic review pertained to the *computation* outcome:

- Computations related to probability differences are easier with pairs of rates per 10<sup>n</sup> than with pairs of percentages or 1 in X rates (subsection 4A: *compute* outcome, numerical formats comparison).
- Adding graphics (icon arrays and bar charts) to numeric communications about probability differences makes it easier to perform computations than providing either numbers alone or graphics alone

**Table 4F** Evidence-Based Guidance for Effect of Framing on Ability to Perform Computations on Probability Differences

Comparison	Evidence Strength	General Guidance
Survival v. mortality curves	Insufficient evidence— too few findings ( $n = 1$ )	It is not clear whether using survival curves versus mortality curves affects ability to compute differences between 2 groups at a single point in time.

(subsection 4C: *compute* outcome, numerical and graphical format comparison).

The available **weak** evidence suggests that there may be different effects for formats such that:

- People may do better answering questions about risk reductions with a pair of percentages than with a pair of rates per 10<sup>n</sup> (subsection 1A, *identification/recall* outcome, numerical format comparison).
- People may be better able to contrast several options with relative risk differences than with pairs of absolute percentages (subsection 2A, *contrast* outcome, numerical format comparison).
- Showing risk reduction as pairs of rates per 10<sup>n</sup> with the same denominator rather than pairs of percentages may help people identify larger versus smaller differences (subsection 2A, *contrast* outcome, numerical format comparison).
- It may be easier to contrast treatment options with a relative risk difference than with either absolute risk differences, NNT, or a combination of these formats (subsection 2A, *contrast* outcome, numerical format comparison).
- People's ability to contrast effect sizes may be better with a bar chart of percentage point improvement than a bar chart of percentage point reduction (subsection 2B, *contrast* outcome, graphical format comparison).
- People may be better able to contrast probability differences associated with multiple options with pairs of negatively framed percentages (with or without positively framed percentages) than with pairs of positively framed ones alone (subsection 2F, *contrast* outcome, numerical format comparison).
- Using a) pre-post percentages instead of RRR (subsection 4A: numerical formats comparison) or b) graphics with numerators and denominators instead of only numerators (subsection 4B: graphical formats comparison) may help people perform certain computations relevant to probability differences (*computation* outcome).

Further **weak** evidence indicates there may **not** be any effect for the following outcome-format comparison combinations:

- The ability to answer questions about probability differences may not be affected by a) adding pairs of absolute rates per 10<sup>n</sup> to the absolute difference (subsection 1A: *identification/recall* outcome, numerical format comparison), b) adding labels to numbers saying which is greater (subsection 1E: *identification/recall* outcome, context), or c) using combination-framed icon arrays showing incremental differences in multiple outcomes versus gain-framed icon arrays highlighting only incremental survival outcomes (subsection 1A: *identification/recall* outcome, graphic format comparison; gain-loss framing).
- The ability to perform computations relevant to probability differences may not be affected by using icon arrays versus bar charts (subsection 4B, *compute* outcome, graphical format comparisons).

## Discussion

This literature review synthesizes the evidence pertaining to the impact of data presentation formats on multiple outcomes (identify/recall, contrast, categorization, and computation) when performing difference cognitive tasks on probability information. The task of assessing the difference between 2 probabilities is a critical one for patients seeking to understand the size of a treatment benefit, the size of a potential harm from a risk factor, and other risk reduction or risk increase messages such as those included in US Preventive Services Task Force guidance<sup>38</sup> or decision aids.<sup>39</sup>

Even when focusing on those questions for which research evidence exists (see Table B), it is notable that we classified most of the evidence from this synthesis as insufficient or weak, mostly due to limited numbers of relevant studies that could be directly compared. As a result, we provide only a few strong guidance statements and no moderate statements.

Although there are 2 pieces of strong evidence about how to help readers perform computations (by using rates per  $10^n$  rather than other numbers and by adding graphics to numbers), we note that information designers should be cautious about requiring readers to do their own computations. Many best practices for inclusive communication<sup>40,41</sup> recommend against making readers perform computations because the cognitive effort and skill level required may be barriers to using the numbers. Instead, the designer of the information should perform the computation for the reader when possible. When computations are unavoidable, the evidence provided here may make them easier.

Weak evidence regarding probability difference communications was more common. Often, this was not because of an absence of high-credibility research but because only 1 high-credibility finding was identified for a particular format comparison, which did not meet our standard for moderate or strong evidence. Overall, these weak evidence statements reinforce the idea that a format beneficial for one outcome may not be ideal for another outcome. For example, showing RRRs (e.g., a 30% reduction) appear to help people *contrast* the effect sizes for several options, but pairs of pre-post percentages (e.g., a reduction from 40% to 10%) appear to be better for helping people perform *computations* around probability differences. This supports our general message that communicators should thoughtfully choose their goal for presenting information before deciding on the format.

Much of the evidence pertaining to the use of graphics to represent probability differences was weak and so specific that it may not be very helpful to professional communicators. For example, it appears that bar charts showing percentage increase are easier for contrast than bar charts of percentage reduction, but little other guidance is available about the design of bar charts. We note that there is more evidence available about the use of graphics to represent single probabilities (rather than probability differences), as summarized in our companion articles; the extent to which this evidence generalizes to representations of probabilities differences is not clear.

There were no findings related to the outcome of *categorization* of probability differences, although, in the context of tasks evaluating the difference between 2 or more values, this is not particularly surprising. Instead of asking participants to categorize differences, studies instead asked which difference was less or greater (*contrast* outcome), such as to help readers determine which options may be best.

Limitations for the Making Numbers Meaningful project include the possibility that studies were missed in

the search, the use of a small group of experts to evaluate study risk of bias and finding credibility, and the granular data extraction that meant that we prioritized narrow comparisons of highly comparable studies (studies with the same task, format comparison, and outcome) rather than more global assessments. We did not perform analyses by audience characteristic (such as numeracy or culture) because so few comparable articles segmented their results by the same characteristics.

An additional limitation is our decision to distinguish between research findings related to point tasks from those related to difference tasks. Because of this decision, we excluded research questions such as, “Do people have an easier time performing computations with absolute risk of disease or with risk relative to the average person’s risk?” In our approach, performing a cognitive task such as a computation with a single probability (such as absolute disease risk) is different from doing so with relative differences. That means that such comparisons did not meet our inclusion criterion of having the same information presented in different formats (see the Methods section).

As shown in Table A, this review article covers part of the research evidence about difference tasks, that is, situations in which an audience examines a stimulus to seek information about probability differences. Other tasks are covered in other articles focusing on point tasks (in which people look for information about individual probabilities), time-trend tasks (involving assessing patterns of probability over time), and synthesis tasks (involving aggregating information about multiple probabilities together, such as a set of risks and benefits of a therapeutic option). As a result, each article in this series presents an inherently incomplete snapshot of the effect of format on important outcomes. Findings and evidence presented here should be considered in the context of the companion articles filling out the evidence on the effects of format on important outcomes in probability communication.


In conclusion, difference tasks—tasks involving assessing or making decisions about a probability difference—are critical to informed decision making about risk-reducing therapies or risk-increasing exposures. The evidence presented here pertains to 3 specific outcomes that might map to communicators’ goals when designing information. When trying to help patients to *identify* or *remember* probability differences, weak evidence supports the superiority of pairs of percentages than pairs of rates per  $10^n$ . When trying to help patients contrast probability differences to select the biggest (or smallest), weak evidence favors relative risk differences or pairs of rates per  $10^n$  with the same denominator rather than pairs of percentages, absolute risk differences, or NNT;

negatively framed percentages rather than positively framed ones; and bar charts of percentage increase rather than bar charts of percentage decrease. Finally, several communication choices (rates per 10<sup>n</sup> rather than 1 in X or percentages; adding graphics to numbers) improve patients' ability to perform computations when the information designer cannot find a way to avoid computations altogether. Communicators who wish to influence other outcomes, such as perceived effectiveness or behavioral intention, should refer to the companion article in this series ("How Difference Tasks Are Affected by Probability Format, Part 2: A Making Numbers Meaningful Systematic Review").<sup>42</sup>


### Acknowledgments

We thank the Numeracy Expert Panel for contributions to conceptualizing the Making Numbers Meaningful project (Cynthia Baur, Sara Cjaza, Angela Fagerlin, Carolyn Petersen, Rima Rudd, Michael Wolf, and Steven Woloshin). We are grateful to Marianne Sharko, MD, MS, Andrew Z. Liu, MPH, and Lisa Grossman Liu, MD, PhD, for contributions to article screening and risk-of-bias assessment. We also thank Jordan Brutus for assisting with data management.

### ORCID iD

Natalie C. Benda  <https://orcid.org/0000-0002-3256-0243>

Brian J. Zikmund-Fisher  <https://orcid.org/0000-0002-1637-4176>

Jessica S. Ancker  <https://orcid.org/0000-0002-3859-9130>

### Data Availability Statement

Abstracted data are freely available in the online appendix referenced in the Results section of this article at <https://osf.io/rvxf2/>. Other methods and materials are available to other researchers upon request to Jessica S. Ancker.

### Supplemental Material

Supplementary material for this article is available online at <https://osf.io/rvxf2/>.

### References

- Ancker JS, Benda NC, Sharma MM, Johnson SB, Weiner S, Zikmund-Fisher BJ. Taxonomies for synthesizing the evidence on communicating numbers in health: goals, format, and structure. *Risk Anal.* 2022;42(12):2656–70. DOI: 10.1111/risa.13875
- Ancker JS, Benda NC, Sharma MM, et al. Scope, methods, and overview findings for the Making Numbers Meaningful evidence review of communicating probabilities in health: a systematic review. *MDM Policy & Pract.* 2025;10(1):23814683241255334. DOI: 10.1177/23814683241255334
- Gigerenzer G. What are natural frequencies? *BMJ.* 2011;343:d6386. DOI: 10.1136/bmj.d6386
- Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med.* 1998;73(5): 538–40. DOI: 10.1097/00001888-199805000-00024
- Woloshin S, Schwartz LM. Communicating data about the benefits and harms of treatment: a randomized trial. *Ann Intern Med.* 2011;155(2):87–96. DOI: 10.7326/0003-4819-155-2-201107190-00004
- Sullivan HW, O'Donoghue AC, Aikin KJ. Communicating benefit and risk information in direct-to-consumer print advertisements: a randomized study. *Ther Innovation Regul Sci.* 2015;49(4):493–502. DOI: 10.1177/2168479015572370
- Tait AR, Voepel-Lewis T, Zikmund-Fisher BJ, Fagerlin A. The effect of format on parents' understanding of the risks and benefits of clinical research: a comparison between text, tables, and graphics. *J Health Commun.* 2010;15(5): 487–501. DOI: 10.1080/10810730.2010.492560
- Tait AR, Voepel-Lewis T, Zikmund-Fisher BJ, Fagerlin A. Presenting research risks and benefits to parents: does format matter? *Anesth Analg.* 2010;111(3):718–23. DOI: 10.1213/ANE.0b013e3181e8570a
- Mühlbauer V, Prinz R, Mühlhauser I, Wegwarth O. Alternative package leaflets improve people's understanding of drug side effects—a randomized controlled exploratory survey. *PLoS One.* 2018;13(9):e0203800. DOI: 10.1371/journal.pone.0203800
- Galesic M, Garcia-Retamero R. Communicating consequences of risky behaviors: life expectancy versus risk of disease. *Patient Educ Couns.* 2011;82(1):30–5. DOI: 10.1016/j.pec.2010.02.008
- Brick C, McDowell M, Freeman ALJ. Risk communication in tables versus text: a registered report randomized trial on 'fact boxes.' *R Soc Open Sci.* 2020;7(3):190876. DOI: 10.1098/rsos.190876
- Zikmund-Fisher B, Fagerlin A, Ubel P. A demonstration of "less can be more" in risk graphics. *Med Decis Making.* 2010;30:661–71. DOI: 10.1177/0272989x10364244
- Hawley S, Zikmund-Fisher B, Ubel P, Jankovic A, Lucas T, Fagerlin A. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Educ Couns.* 2008;73: 448–55. DOI: 10.1016/j.pec.2008.07.023
- Schonlau M, Peters E. Comprehension of graphs and tables depend on the task: empirical evidence from two Web-based studies. *Stat Polit Policy.* 2012;3:1–35.
- Houston AJ, Kamath GR, Bevers TB, et al. Does animation improve comprehension of risk information in patients with low health literacy? A randomized trial. *Med Decis Making.* 2020;40(1):17–28. DOI: 10.1177/0272989X19890296
- Perneger TV, Agoritsas T. Doctors and patients' susceptibility to framing bias: a randomized trial. *J Gen Intern Med.* 2011;26(12):1411–7. DOI: 10.1007/s11606-011-1810-x



17. Covey J. The effects of absolute risks, relative risks, frequencies, and probabilities on decision quality. *J Health Commun.* 2011;16(7):788–801. DOI: 10.1080/10810730.2011.561916
18. Sheridan SL, Pignone MP, Lewis CL. A randomized comparison of patients' understanding of number needed to treat and other common risk reduction formats. *J Gen Intern Med.* 2003;18(11):884–92. DOI: 10.1046/j.1525-1497.2003.21102.x
19. Lavallie DL, Wolf FM, Jacobsen C, Sprague D, Buchwald DS. Health numeracy and understanding of risk among older American Indians and Alaska natives. *J Health Commun.* 2012;17(3):294–302. DOI: 10.1080/10810730.2011.626497
20. Wolfe CR, Reyna VF, Smith RJ. On judgments of approximately equal. *J Behav Deci Making.* 2018;31(1):151–63. DOI: 10.1002/bdm.2061
21. Okan Y, Galesic M, Garcia-Retamero R. How people with low and high graph literacy process health graphs: evidence from eye-tracking. *J Behav Deci Making.* 2016;29(2-3):271–94. DOI: 10.1002/bdm.1891
22. Tolbert E, Brundage M, Bantug E, et al. In proportion: approaches for displaying patient-reported outcome research study results as percentages responding to treatment. *Qual Life Res.* 2019;28(3):609–20. DOI: 10.1007/s11136-018-2065-3
23. Waters EA, Weinstein ND, Colditz GA, Emmons KM. Reducing aversion to side effects in preventive medical treatment decisions. *J Exp Psychol Appl.* 2007;13(1):11–21. DOI: 10.1037/1076-898X.13.1.11
24. Wright AJ, Whitwell SCL, Takeichi C, Hankins M, Marteau TM. The impact of numeracy on reactions to different graphic risk presentation formats: an experimental analogue study. *Br J Health Psychol.* 2009;14(pt 1):107–25. DOI: 10.1348/135910708X304432
25. Harper S, King NB, Young ME. Impact of selective evidence presentation on judgments of health inequality trends: an experimental study. *PLoS One.* 2013;8(5):e63362. DOI: 10.1371/journal.pone.0063362
26. Silk KJ, Parrott RL. Math anxiety and exposure to statistics in messages about genetically modified foods: effects of numeracy, math self-efficacy, and form of presentation. *J Health Commun.* 2014;19(7):838–52. DOI: 10.1080/10810730.2013.837549
27. Shoemaker SJ, Wolf MS, Brach C. *The Patient Education Materials Assessment Tool (PEMAT) and User's Guide.* AHRQ Publication 14-0002-EF. Rockville (MD): Agency for Healthcare Research and Quality; 2013.
28. Cuite CL, Weinstein ND, Emmons K, Colditz G. A test of numeric formats for communicating risk probabilities. *Med Decis Making.* 2008;28(3):377–84. DOI: 10.1177/0272989X08315246
29. Bodemer N, Meder B, Gigerenzer G. Communicating relative risk changes with baseline risk: presentation format and numeracy matter. *Med Decis Making.* 2014;34(5):615–26. DOI: 10.1177/0272989X14526305
30. Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. *Ann Intern Med.* 1997;127(11):966–72. DOI: 10.7326/0003-4819-127-11-199712010-00003
31. Koo K, Brackett CD, Eisenberg E, Kieffer KA, Hyams ES. Impact of numeracy on understanding of prostate cancer risk reduction in PSA screening. *J Gen Intern Med.* 2017;12(12):e0190357. DOI: 10.1371/journal.pone.0190357
32. Garcia-Retamero R, Galesic M, Gigerenzer G. Enhancing understanding and recall of quantitative information about medical risks: a cross-cultural comparison between Germany and Spain. *Span J Psychol.* 2011;14(1):218–26. DOI: 10.5209/rev\_sjop.2011.v14.n1.19
33. Garcia-Retamero R, Galesic M, Gigerenzer G. Improving comprehension and communication of risks about health. *Psicothema.* 2011;23(4):599–605.
34. Garcia-Retamero R, Muñoz R. Cómo mejorar la comprensión de los riesgos médicos en personas mayores [How to improve comprehension of medical risks in older adults]. *Revista Latinoamericana de Psicología.* 2013;45(2):253–64. DOI: 10.14349/rlp.v45i2.1071
35. Garcia-Retamero R, Galesic M. Who profits from visual aids: overcoming challenges in people's understanding of risks [published correction appears in *Soc Sci Med.* 2010;70(12):2097]. *Soc Sci Med.* 2010;70(7):1019–25. DOI: 10.1016/j.socscimed.2009.11.031
36. Price M, Cameron R, Butow P. Communicating risk information: the influence of graphical display format on quantitative information perception—accuracy, comprehension and preferences. *Patient Educ Couns.* 2007;69(1–3):121–8. DOI: 10.1016/j.pec.2007.08.006
37. Armstrong K, Schwartz JS, Fitzgerald G, Putt M, Ubel PA. Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. *Med Decis Making.* 2002;22(1):76–83. DOI: 10.1177/0272989X0202200108
38. Nelson HD, Fu R, Zakher B, Pappas M, McDonagh M. Medication use for the risk reduction of primary breast cancer in women: updated evidence report and systematic review for the US preventive services task force. *JAMA.* 2019;322(9):868–86. DOI: 10.1001/jama.2019.5780
39. International Patient Decision Aid Standards (IPDAS) Collaboration. Available from: <http://ipdas.ohri.ca/>
40. Baur C, Prue C. The CDC clear communication index is a new evidence-based tool to prepare and review health information. *Health Promot Pract.* 2014;15(5):629–37. DOI: 10.1177/1524839914538969
41. Koh HK, Baur C, Brach C, Harris LM, Rowden JN. Toward a systems approach to health literacy research. *J Health Commun.* 2013;18(1):1–5. DOI: 10.1080/10810730.2013.759029
42. Benda NC, Zikmund-Fisher BJ, Sharma MM, et al. How difference tasks are affected by probability format, part 2: a Making Numbers Meaningful systematic review. *MDM Policy Pract.* 2025;10(1):23814683241310242. DOI: 10.1177/23814683241310242