How Difference Tasks Are Affected by Probability Format, Part 2: 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 evaluate the effect of data presentation format on communication of health probabilities, the Making Numbers Meaningful team undertook a systematic review. **Purpose.** This article presents evidence about difference tasks, in which a reader examines information to evaluate differences between probabilities, such as the effect of a therapy on the chance of recurrence. This article covers the effect of format on 5 outcomes: 1) perceptions of or feelings about effectiveness, 2) behavioral intentions or behaviors, 3) trust, 4) preference for the format, and 5) discrimination. **Data Sources.** MEDLINE, Embase, CINAHL, the Cochrane Library, PsycINFO, ERIC, ACM Digital Library; hand search. **Finding Selection.** Experimental/quasi-experimental studies comparing 2 or more formats for presenting quantitative health information. This article covers 205 findings from 101 unique studies reported in 84 articles. **Data Extraction.** Dual extraction of information on stimulus, task, and perceptual, affective, cognitive, and behavioral outcomes. **Data Synthesis.** Evidence is moderate to strong that behavioral intention is affected more by relative differences than absolute ones, by numerator-only graphics than part-to-whole graphics, by messages with anecdotes than without, and by information about what others chose. Evidence is strong that perceived and felt effectiveness is affected more by relative differences than by absolute ones and more by numerator-only graphics rather than part-to-whole graphics. For graphic preferences, bar charts were preferred to icon arrays and graphics with data labels to graphics without. Other comparisons had weak or insufficient evidence. **Limitations.** The detailed approach to evidence syntheses provides narrowly targeted evidence rather than broad statements. **Conclusions.** Moderate to strong evidence can be derived on effects of probability difference format on behavioral intention, perceived or felt effectiveness, and preference for format.

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).

Highlights

- Communicating relative risk differences as opposed to absolute risk differences, using numerator-only instead of part-to-whole graphics, and including anecdotes or information about others' decisions will all increase intentions to engage in a behavior.
- Relative risks (rather than absolute risk differences) and numerator-only graphics (rather than part-to-whole) will also increase felt and perceived effectiveness.
- To illustrate probability differences, people tend to prefer bar charts over icon arrays and graphics with labels over those without.
- All findings regarding the impact of different presentation formats for probability differences on trust produced insufficient evidence.

Numerical information is essential for patients to understand probabilities of health and disease. In particular, differences between 2 probabilities are used to express the effects of risk factors or therapies on the probability of disease, health, or other outcomes. The patient may have to assess the size of the effect to select therapeutic options, make behavior changes to reduce risk, or make other health-related decisions. Such differences may be formatted in a number of ways: as 2 individual probabilities (e.g., a change from a 3% risk to a 4.5% risk), the

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, 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.

relative difference between those probabilities (a 50% relative increase, a relative risk of 1.5), the absolute difference between them (a 1.5-percentage-point absolute increase), some combination of these, or some other format. Although all these formats express the same information, it is well established that the format can influence the reader's perceptions and decisions.

To develop evidence about the effects of different formats on patients and the public, we conducted a large systematic literature review on the communication of numbers in health, across data types and across different data presentation formats.[1,2] We organized the literature according to a conceptual model of communication in which a reader views a stimulus (consisting of data in some data presentation format), performs cognitive tasks to make sense of it, and experiences cognitive, perceptual, or behavioral responses that are measured with outcome measures. In this model, there are several types of tasks. *Point* tasks involve evaluating information about individual probabilities. *Time-trend* tasks are performed to assess probability patterns over time, and *synthesis* tasks are performed when multiple probabilities are integrated into a single judgment, such as when a patient evaluates the risks and benefits of a therapy. *Difference* tasks—the focus of the current article—assess differences between probabilities, such as relative or absolute risk differences.

Findings of the systematic review were too large to present in a single article and thus are being presented as

Table A Current Article's Scope within the Making Numbers Meaningful Systematic Review

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

[a]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.[1]

a series (Table A). The current article presents the evidence pertaining to probability data, difference tasks, and the following 5 outcomes: 1) effectiveness perception (perception of the size of a probability difference) or feeling (perception assessed on a scale of worry, concern, or other affective response), 2) behavioral intentions (intended or planned behavior) or health behavior, 3) trust (perceived credibility of the information as presented), 4) preference or positive perception of the data presentation format (e.g., perceived helpfulness and other related constructs), and 5) discrimination between probability differences of different sizes. A separate article labeled "Part 1" presents 5 additional outcomes regarding communication of probability differences.

Methods

As described in more detail in our companion methodology article,[1] the literature review sought experimental (randomized) and quasi-experimental (prospectively collected questionnaires or surveys that lacked random assignment) research comparing 2 or more ways of presenting quantitative health-related data to lay, nonmedical audiences. We followed systematic methods for the literature search, screening, risk-of-bias evaluation, data extraction, credibility evaluation of findings, and organization into evidence tables.[1] The search was performed on MEDLINE, Embase, CINAHL, the Cochrane Library, PsycINFO, ERIC, 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*. All instruments used (search strategy, data extraction instrument, and study risk of bias or S-ROB rubric) are available at the Making Numbers Meaningful Project at the Open Science Framework site (https://osf.io/rvxf2/).

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. We created a brief free-text description of sample demographics, which is available with the rest of the data at the OSF repository (https://osf.io/rvxf2/). 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*, *effectiveness 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*). Although we categorized the outcomes in this way, it is important to recognize that the actual measures used in the different studies varied.

During data extraction, each unique combination of task, format comparison, and outcome from a single study was termed a *finding*. For example, a single study could provide 1 finding about the effect of relative versus

absolute risk difference on perceived effectiveness and another finding about the effect of relative versus absolute risk difference on health behavior. Each finding was rated for credibility on a scale from 1 to 10 by pairs of authors (N.C.B., J.S.A., B.J.Z.-F.) holistically assessing a list of finding and study characteristics (sample size, statistical methods, validity of stimulus design, comparison, outcome measures, and covariates, plus the S-ROB for the study from which the finding came). Findings from the same study often varied in credibility; for example, primary analyses with good statistical power might have high credibility, and unplanned secondary or subset analyses with small samples might have lower credibility. A credibility of 7 or higher was considered high, 4.5 to 6.5 moderate, and 4 or below low. 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. After grouping relevant findings together, we graded the strength of evidence as follows:

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

For the current article, it is important to explain a few terms. Among probabilities, we distinguished between rates formatted as 1 in X (such as "1 in 5") and those formatted as a rate per 10^n (such as "20 in 100"). We used the term *natural frequency* only to describe a series of joint probabilities and conditional probabilities computed from the same pool of patients, in the context of Bayes' theorem.[3,4] By doing so, we intended to match the original definition of the term[4] and distinguish between findings that otherwise might appear contradictory.[3]

The literature review identified 316 articles addressing probability communication, of which 84 (representing 101 unique studies) involved difference tasks with probability data and 1 or more of the outcomes for this article.

Results

As outlined in Table B, each "Results" subsection summarizes evidence on the following comparisons in order: comparisons among number formats, among graphics formats, between number and graphic formats, between number and verbal formats, between different types of contextual elements, between different framings, effect of manipulations of denominators, effect of animation or interactivity, and manipulations of time frames. No articles examined the effect of representations of uncertainty. Within subsections, evidence is arranged from strongest to weakest.

Effects of Different Formats for Probability Differences on Effectiveness Perceptions and Feelings: Section 6

The way that a probability difference (the representation of the effect of a risk factor or therapy) is formatted has the potential to influence the perception of how large or important that effect is. When the perception was measured on a scale from small to large, we considered it a measure of "effectiveness perception," and when it was measured on an affective scale such as "not at all concerned" to "very concerned," we considered it "effectiveness feelings." Although these 2 outcomes were extracted separately, they are similar and so we group them into section 6.

It is important to note that we report throughout whether a format was associated with increases or decreases in perceived effectiveness, not whether it was associated with increases or decreases in accuracy. The concept of accuracy—that is, the ability to correctly identify or recall a specific number or quantity—is covered in our companion part 1 article, where it appears under section 1, the so-called identification/recall outcome.

Comparisons between numerical formats for probability differences in their effect on effectiveness perceptions and feelings (subsection 6A). EFFECT OF SHOWING RELATIVE DIFFERENCES: A high-credibility finding showed that perceived effectiveness was higher when a relative risk reduction as a percentage was added to a

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

Subsection		Section					
Data Presentation Format Comparison	**Section Number/ Subsection Letter[a,b]**	**Effectiveness Perceptions and Effectiveness Feelings**	**Behavior and Behavioral Intention**	**Trust**	**Preference**	**Discrimination**	**Total**
		6	7	8	9	10	
Comparisons between numerical formats	A	6A ($n = 9$)	7A ($n = 41$)	8A ($n = 2$)	9A ($n = 13$)	10A ($n = 1$)	66
Comparisons between graphical formats	B	6B ($n = 13$)	7B ($n = 16$)	8B ($n = 0$)	9B ($n = 23$)	10B ($n = 0$)	52
Comparisons between numerical and graphical formats	C	6C ($n = 21$)	7C ($n = 21$)	8C ($n = 1$)	9C ($n = 12$)	10C ($n = 1$)	56
Comparisons between numerical and verbal probabilities	D	6D ($n = 0$)	7D ($n = 2$)	8D ($n = 0$)	9D ($n = 0$)	10D ($n = 0$)	2
Comparisons of elements added for context	E	6E ($n = 3$)	7E ($n = 7$)	8E ($n = 0$)	9E ($n = 1$)	10E ($n = 2$)	13
Comparisons of frames (gain, loss, combination)	F	6F ($n = 0$)	7F ($n = 6$)	8F ($n = 0$)	9F ($n = 1$)	10F ($n = 0$)	7
Comparisons of larger or smaller denominators	H	6H ($n = 3$)	7H ($n = 0$)	8H ($n = 0$)	9H ($n = 0$)	10H ($n = 0$)	3
Comparisons of animation and interactivity	I	6I ($n = 0$)	7I ($n = 0$)	8I ($n = 1$)	9I ($n = 2$)	10I ($n = 0$)	3
Comparisons of shorter or longer time periods	J	6J ($n = 1$)	7J ($n = 1$)	8J ($n = 0$)	9J ($n = 1$)	10J ($n = 0$)	3
Total findings per outcome		50	94	4	53	4	205

[a]No findings pertained to the representation of uncertainty (row G).
[b]The subhead numbering system in Table B is standard across all Making Numbers Meaningful results articles. The standard numbers ensure that, for example, studies of the effects of gain-loss framing manipulations on health behavior are always placed in a subhead labeled section 7F (whether or not that article contains sections 1 through 6). The goal is for readers to be able to use this system to locate similar sections across articles.

pair of pre and post percentages.[5] Similarly, a moderate credibility finding[6] found higher perceived effectiveness with relative outcome rates than absolute probabilities.

INCREMENTAL RISK PLUS BASELINE RISK VERSUS PRE/POST ABSOLUTE RATES: High- and moderate-credibility findings (Zikmund-Fisher et al.[7] substudy 1, Zikmund-Fisher et al.[8]) showed less concern about side effects when the effects were presented as baseline rates per 10^n plus absolute difference than when they were shown as a pair of pre/post absolute rates per 10^n.

TABLES VERSUS TEXT: A high-credibility finding by Tait et al.[9] showed no difference in effectiveness perception when rate per 10^n numbers were presented in tables or in text.

EFFECT OF SHOWING ARITHMETIC DIFFERENCES: A moderate-credibility finding[10] showed effectiveness perception was higher when

Table 6A Evidence-Based Guidance for Effects of Numerical Formats for Probability Differences on Effectiveness Perceptions and Feelings

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	Guidance
Effect of showing relative differences	Strong ($n = 2$)	People perceive effectiveness to be higher if it is presented as, "Taking drug A reduces the chance of problems by 50% compared to taking drug B," than if it is presented as "30% of patients taking drug B and 15% of patients taking drug A experienced problems."	Presenting effectiveness as a relative difference (with or without the pre and post probabilities) increases effectiveness perceptions,
Incremental risk plus baseline risk v. pre/post absolute rates	Moderate ($n = 2$)	Concern about a probability difference is lower if it is shown as 3 more people out of 100 (on top of the baseline probability of 10 out of 100) than when it is shown as 13 out of 100 versus 10 out of 100.	Presenting a probability difference as baseline probability plus an absolute difference reduces concern compared to presenting it as a pair of probabilities.
Tables v. text	Weak ($n = 1$)	When showing that a drug reduces risk from 5 in 100 to 1 in 100, it may not matter whether a table is used or a sentence.	When probability differences are shown as rates per 10^n, putting them into a table rather than in text may not affect effectiveness feelings.
Effect of showing arithmetic differences	Insufficient—too few studies ($n = 1$)	It is not clear whether supplementing the arithmetic difference with the pre and post absolute rates leads to different perceptions of effectiveness than showing the absolute difference alone.	
Percentage probability v. life expectancy	Insufficient—too few studies ($n = 1$)	It is not clear whether presenting the effect of a risk factor as percentages or as impact on life expectancy affects feelings about the effects.	

arithmetic differences were accompanied by absolute rates than when the differences were presented alone; however, the effect was only in a high-benefit condition but not for a similarly designed low-benefit condition.

PERCENTAGE PROBABILITY VERSUS LIFE EXPECTANCY: A moderate-credibility finding by Galesic and Garcia-Retamero[11] showed perceived differences caused by a risk factor to be larger (as measured by lower perceived desirability) when post–risk factor estimates plus absolute difference statistics were framed in life expectancy terms rather than percentages.

Two low-credibility findings were not summarized (Blalock et al.[10] due to inconsistent findings across conditions with high and low benefit, and Wadhwa and Zhang[12] due to a small sample size, which reduced confidence in nonsignificant findings).

Comparisons between graphic formats for probability differences in their effect on effectiveness perceptions and feelings (subsection 6B). PART-WHOLE VERSUS FOREGROUND-ONLY GRAPHICS: Five high and moderate-credibility findings involve comparing part-to-whole graphics to numerator-only graphics, consistently finding that people perceived larger probability differences with numerator-only formats. A part-to-whole graphic portrays both the numerator and denominator of a probability, such as an icon array representing a 10% risk with 10 colored icons and 90 white icons. A numerator-only graphic shows only the numerator (e.g., 10 affected individuals alone). These include research on icon arrays (graphics that portray the probability of an event with colored human figures or abstract icons; Okan et al.[13] substudy 2, Stone et al.[14] substudy 2), bar charts (Okan et al.,[15] Stone et al.[16] substudy 2), and a comparison between numerator-only icon arrays and pie charts (Stone et al.[16] substudy 1). The latter 2 findings showed the same impact on effectiveness feelings.

BASELINE PLUS ABSOLUTE DIFFERENCE VERSUS PRE/POST ABSOLUTE RATES PORT-

Table 6B Evidence-Based Guidance for Effects of Graphical Formats for Probability Differences on Effectiveness Perceptions and Feelings

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Part-to-whole graphics v. foreground-only graphics	Strong ($n = 7$)	A probability difference of 5 percentage points will appear larger, and may feel larger, with icon arrays that display the numerators of the probability only (e.g., 10 affected individuals and 5 affected individuals), than with icon arrays that show the numerators and the denominators (an array of 10 affected and 90 unaffected icons, and an array of 5 affected and 95 unaffected ones).	Graphics that display numerators only (foreground-only design) make a difference in probability appear larger than part-to-whole graphics. There is weak evidence for a similar effect on emotional reactions.
Baseline plus absolute difference v. pre/post absolute rates portrayed as icon arrays	Moderate ($n = 2$)	Concern about a probability difference of 5 percentage points is lower if it is portrayed in an icon array showing the baseline probability plus incremental difference (10 affected individuals without treatment, 5 affected individuals with treatment) than when shown as a pair of icon arrays (one showing 10 in 100, the other 5 in 100).	Presenting a probability difference visually with graphics showing baseline probability plus an absolute difference reduces concern compared with presenting it as a pair of probabilities.
Use of number labels	Insufficient evidence—inconsistent findings ($n = 2$)	It is not clear whether adding rate per 10^n numbers to risk graphics affects perceptions of probability differences.	
Adding postintervention probability to baseline probability	Insufficient—too few findings ($n = 1$)	It is not clear whether adding postintervention probabilities to preintervention probabilities in risk graphics affects the perception of intervention effectiveness.	
Absolute risk differences v. risk ratios	Insufficient—too few findings ($n = 1$)	It is not clear whether effectiveness feelings about policies to reduce population disparities are affected by whether disparities are presented visually as absolute differences or as risk ratios.	

RAYED AS ICON ARRAYS: High- and moderate-credibility findings (Zikmund-Fisher et al.[7] substudy 1, Zikmund-Fisher et al.[8]) showed less concern about side effects when the effects were shown in icon array graphics as baseline probability plus absolute difference than when they were shown as a pair of pre/post absolute probabilities.

USE OF NUMBER LABELS: A moderate credibility finding[15] found lower perceived effectiveness of a drug therapy when rate per 10^n data labels was added to bar charts. However, a moderate-credibility finding[17] showed that adding number labels to vertical number lines with interpretive labels did not affect the perception of the effectiveness of exercise.

ADDING POSTINTERVENTION PROBABILITY TO BASELINE PROBABILITY: Janssen et al.[17] also included a moderate-credibility finding that adding postintervention probabilities (thereby visually illustrating the absolute risk difference) to vertical number lines did not affect perception of the effectiveness of exercise.

ABSOLUTE RISK DIFFERENCES VERSUS RISK RATIOS: A moderate-credibility finding[18] showed greater changes in effectiveness feelings for policies to reduce population-level risk disparities when grouped

Table 6C　Evidence-Based Guidance for Contrasts between Numerical and Graphical Formats for Probability Differences, and Combinations of Numerical and Graphical Formats, on Effectiveness Perceptions and Feelings

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	Guidance
Combining graphics with pairs of probabilities	Moderate ($n = 7$)	When a message describes the effect of a risk factor as "raising chance of disease from 10% to 15%," adding bar charts or icon arrays to the message may not change perceptions.	Adding part-to-whole graphics (icon array, bar chart) to pairs of probabilities (rates per 10^n, percentages) does not affect perceptions of probability differences.
Numerator-only graphics v. absolute probabilities or probability differences	Moderate ($n = 6$)	Describing the effect of a risk factor as "raising the chance from 10 in 100 to 15 in 100" may make it appear smaller than illustrating it as a line of 5 icons representing the additional 5 people in 100 affected.	Effectiveness perceptions are higher if a probability difference is presented as foreground-only graphics than as a pair of rates per 10^n.
Part-to-whole graphics v. absolute probability differences	Weak ($n = 5$)	Describing the effect of a risk factor as "raising chance of disease from 10% to 15%" may make the effect appear larger than showing it in a bar chart.	Effectiveness perceptions and feelings may be higher if a risk difference is presented as a pair of rates per 10^n than as part-to-whole graphics (pie charts or bar charts).
Other combinations	Insufficient—too few findings ($n = 1$)	It is unclear whether effectiveness perceptions were affected by a) part-to-whole icon arrays with rates per 10^n versus b) 1 in X plus relative risk reduction.	

Table 6D　Evidence-Based Guidance on Effects of Contextual Elements for Probability Differences on Effectiveness Perceptions and Feelings.

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Visual emphasis on denominator	Weak ($n = 1$)	Readers' effectiveness perceptions may not change if an icon array showing 5 additional people affected out of 100 is accompanied by the number **100** in boldface red font.	Using bolding and color to highlight the denominator of a rate per 10^n may not affect perceived effectiveness.
Adding average chance to message about individual chance	Insufficient-inconsistent findings ($n = 2$)	Whether providing the average person's risk in additional to personal risk affects perceived effectiveness of risk-reducing interventions.	

bar charts showed risk ratios instead of absolute differences.

Comparisons between numerical and graphical formats for probability differences in their effect on effectiveness perceptions and feelings (subsection 6C). COMBINING GRAPHICS WITH PAIRS OF PROBABILITIES (PERCENTAGE OR RATE PER 10^n): A high-credibility finding[5] showed no effect of adding part-to-

whole icon arrays to pre-post percentage with or without RRR as percentage. Both a high-credibility finding (Zikmund-Fisher et al.[7] substudy 1) and a moderate-credibility finding[8] showed no effect on concern about a set of side effects of adding part-to-whole icon arrays to pairs of rate per 10^n numbers, regardless of whether the information was pre/post absolute rates or baseline values plus arithmetic differences. Similarly, a moderate-credibility finding (Dragicevic and Jansen[19] substudy 2) suggested no differences between risk difference

Table 6H Evidence-Based Guidance on Effect of Manipulating Denominators of Probability Differences on Effectiveness Perceptions and Feelings

Comparison	Evidence Strength	General Guidance
Larger v. smaller denominators	Insufficient—conflicting studies ($n = 3$)	It is not clear whether adding an icon array to a pair of rates per 10^n reduces the effects of denominator neglect in effectiveness perception.

Table 6J Evidence-Based Guidance on Effect of Varying the Time Period for Probability Differences on Effectiveness Perceptions and Feelings

Comparison	Evidence Strength	General Guidance
Longer v. shorter time periods	Insufficient—too few studies ($n = 1$)	It is not clear whether presenting risk differences across different time frames (e.g., 1 mo v. 12 mo) affects effectiveness perceptions or feelings.

presented as pair of percentages with or without a part-to-whole bar chart. Also, Martin et al.[20] (a moderate-credibility finding) demonstrated no differences between presenting risk difference as arithmetic risk difference percentage alone or with the addition of icon arrays, speedometer graphics, or photos.

However, a high-credibility finding[7] demonstrated that in risk-benefit communication with pairs of pre-post probabilities (rate per 10^n and percent) plus arithmetic differences, effectiveness perception of the drug was lowest with a data table alone and highest when the numbers were supplemented with integrated icon arrays for each harm and benefit.

A moderate-credibility finding[21] showed that the effects of varying denominators when presenting treatment effectiveness as rates per 10^n could be significantly reduced by adding pairs of part-to-whole icon arrays.

NUMERATOR-ONLY GRAPHICS VERSUS ABSOLUTE PROBABILITIES OR PROBABILITY DIFFERENCES (RATE PER 10^n): In high- and moderate-credibility findings (Stone et al.[14] substudy 1 and Stone et al.[14] substudy 2), each communicating benefit from improved product as pair of pre-post rate per 10^n, a combined measure of effectiveness perception and probability of harm was higher when benefit was shown as a pair of foreground-only icon arrays than with the rate per 10^n only. (Note that the outcome measure combines perception of the difference between probabilities with perception of individual probabilities, reducing confidence in conclusions about either alone.) A moderate-credibility finding also demonstrated that foreground/numerator-only icon arrays resulted in higher effectiveness feelings than pairs of rate per 10^n numbers (Stone et al.[16] substudy 1 perceptions outcome), although effectiveness perception did not differ significantly (Stone et al.[16] substudy 1 feelings outcome). A moderate-credibility finding similarly demonstrated that foreground/numerator-only bar charts led to higher effectiveness perception and effectiveness feelings as compared with rate per 10^n (Stone et al.[16] substudy 2 perception outcome and Stone et al.[16] substudy 2 feelings outcome). These findings assessed the effect only for relatively rare events.

PART-TO-WHOLE GRAPHICS VERSUS ABSOLUTE PROBABILITY DIFFERENCES (RATE PER 10^n): Two moderate-credibility findings from a single article compared rate per 10^n numeric formats to graphical formats. Stone et al.[16] substudy 1 (perception and feelings outcomes) showed that pairs of rates per 10^n resulted in higher perceived and effectiveness feelings than pie charts did, while Stone et al.[16] substudy 2 demonstrated the same effects of rate per 10^n compared with horizontal part-to-whole bar charts (perception and feelings outcomes). Both findings assessed the effect only for relatively rare events, decreasing confidence in the generalizability of this finding for larger probabilities. However, a small moderate-credibility finding[20] showed that in communicating the effect of a hypothetical anti–rheumatoid arthritis drug, there was no difference in effectiveness perception by format (arithmetic difference in percentage, arithmetic difference in percentage plus graphic photos, pair of icon arrays, or pair of speedometer illustrations).

OTHER COMBINATIONS: One moderate-credibility finding (Vogt and Marteau[22] substudy 1)

Table 7A Evidence-Based Guidance for Effects of Numerical Formats for Probability Differences on Health Behavior and Behavioral Intention

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Relative risk reduction v. pairs of pre-post values	Strong ($n = 17$)	People are more likely to choose treatment A if they are told that it reduces their risk by half than if they are told it reduces their risk from 10% to 5%.	A relative risk reduction (alone or in combination with other statistics) has a greater effect on behavioral intention than pairs of pre-post probabilities or the arithmetic difference between them.
NNT v. other probability formats	Strong ($n = 7$)	People are more likely to choose treatment A if they are told that it reduces their risk by half than if they are told that 20 people need to take treatment A for 1 death to be prevented.	Showing a risk difference in familiar probability and probability difference formats (such as rates per 10^n, 1 in X, percentage) is more effective at prompting behavioral intentions than showing it as NNT (number needed to treat).
Effective age/ heart age v. relative risk reduction	Moderate ($n = 2$)	People may be more likely to choose treatment A if told it reduces their risk by half than if they are told that it would change their "heart age" from 60 to 58 y.	Relative risk reductions have stronger effects on behavioral intention than "heart age" or "effective age" formats.
Effective age/ heart age v. absolute percentages	Weak ($n = 1$)	People may not have different intentions about treatment A whether they are told it reduces their heart age from 60 to 58 y or whether they are told that it reduces their risk from 10% to 5%.	There may be no difference in behavioral intentions whether effects are described in "heart age" terms or pairs of absolute percentages.
Delay of event v. probability formats	Insufficient— inconsistent findings ($n = 5$)	It is not clear whether showing a probability difference as years of prolonged life or delay of event has consistent effects on behavioral intention compared with relative risk reduction, pre/post absolute probabilities, or arithmetic difference between probabilities.	
Round numbers	Weak ($n = 3$)	People may be more likely to engage in risk-reducing behaviors when the effect is described as a reduction of 33.00% versus a reduction of 33.33%.	Showing a probability reduction in round numbers v. nonround numbers may increase behavioral intentions.
1 in X v. percentages	Insufficient—too few findings ($n = 1$)	It is not clear whether showing a probability difference as in 1 in X or with pairs of percentages affects behavioral intentions or behaviors.	
Rate per 10^n v. percentages	Insufficient—too few findings ($n = 2$)	It is not clear whether showing a probability difference as rates per 10^n or with pairs of percentages affects behavioral intentions or behaviors.	
Rate per 10^n v. decimal values	Insufficient—too few findings ($n = 1$)	It is not clear whether showing a probability difference as rates per 10^n or with pairs of decimals affects behavioral intentions or behaviors.	

demonstrated that presenting a risk difference as rate per 10^n plus icon array was associated with higher effectiveness perceptions compared with 1-in-X number formats plus relative risk reduction. Effectiveness feelings were not assessed.

Two low-credibility findings (Dragicevic and Jansen[19] substudies 1 and 3) were not summarized due to small sample sizes and in 1 case ceiling effects.

Comparisons of elements added for context on effectiveness perceptions and feelings (subsection 6D). VISUAL EMPHASIS ON A DENOMINATOR: A high-credibility finding (Stone et al.[14] substudy 1) demonstrated no difference in perception of differences when a foreground-only icon array accompanied by the numerical denominator had salience increased by making the denominator number bold and red.

Table 7B Evidence-Based Guidance for Effectiveness of Graphical Formats for Probability Differences on Health Behavior and Behavioral Intention

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Numerator-only v. part-to-whole graphics	Strong ($n = 4$)	Intention to take a risk-reducing action that brings an 8% absolute benefit is stronger when people see its effect portrayed as 8 icons alone (portraying the 8 additional people who will benefit) rather than 8 colored icons in a field of 92 white icons.	Showing an effect as a numerator-only icon array or bar chart, rather than as a part-to-whole graphic, is associated with stronger intention to act.
Different graphical formats	Moderate ($n = 7$)	People's intention to act may not be affected in a consistent manner by whether a 50% risk reduction from (say) vaccination is shown as a bar chart, an icon array, or another form of graphic.	No specific form of risk graphic appears to consistently result in stronger or weaker behavioral intentions. Evidence is insufficient for behaviors.
Numerical labels	Weak ($n = 2$)	Intention to take a risk-reducing action may not change when the number lines showing current and postintervention probabilities include labels such as "current risk: 8 out of 100" and "risk with medication: 5 out of 100."	Adding rate per 10^n labels to bar charts or number lines may not affect behavioral intentions.
Random v. grouped icons in icon array	Insufficient—too few findings ($n = 1$)	It is not clear whether using grouped versus randomly arranged icons in an icon array affects behavioral intention.	
Absolute probabilities v. relative risks	Insufficient—too few findings ($n = 1$)	It is not clear whether support for policies to reduce population disparities is affected by whether disparities are presented visually as absolute differences or risk ratios.	
Postintervention probability data	Insufficient—too few findings ($n = 1$)	It is not clear whether adding postintervention probability data (thereby visually showing the absolute probability difference) affects behavioral intentions.	

ADDING AVERAGE CHANCE TO INDIVIDUAL CHANCE: A high-credibility finding[23] demonstrated that people rated the effectiveness of a drug as more important when they were told that they were at higher-than-average risk than when they were told they were at lower-than-average risk (when the actual quantitative risk was held constant). However, a moderate-credibility finding[17] demonstrated that when effects of exercise on risk were shown in a risk ladder, effectiveness perceptions were not affected by adding a second vertical number line showing the "average person's risk."

Comparisons of larger versus smaller denominators for probability differences on effectiveness perceptions and feelings (subsection 6H). LARGER VERSUS SMALLER DENOMINATORS: One moderate-credibility finding[21] demonstrated that denominator neglect in perceived effectiveness was reduced by supplementing rates per 10^n with different denominators with pairs of icon arrays. The attenuation of denominator neglect was particularly marked for those who were reading information not in their native language. A high-credibility finding (Zikmund-Fisher et al.)[7] found reduced concern about rarer side effects when presented as rates per 10^n with a denominator of 100 versus 1,000, regardless of whether icon array displays were added. Furthermore, in a moderate-credibility finding (Zikmund-Fisher et al.),[8] concern about side effects was not affected by varying the denominator of rates per 10^n and/or icon arrays between 100 and 1,000.

Comparisons of longer versus shorter time period for probability differences on effectiveness perceptions and feelings (subsection 6J). LONGER VERSUS SHORTER TIME PERIOD: One moderate-credibility finding (Vogt and Marteau[22] substudy 2) demonstrated that effectiveness perceptions were higher for effects presented over a

Table 7C Evidence-Based Guidance for Comparisons of Numerical and Graphical Formats for Probability Differences, and Combinations of Numerical and Graphical Formats, on Health Behavior and Behavioral Intention

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Numerator-only icon array v. numbers alone	Strong ($n = 6$)	People are more likely to act on an 8-percentage-point risk difference if it is shown as a set of 8 icons (portraying 8 people in 100) rather than describing it as an 8% increase or decrease.	Showing a probability difference as a numerator-only icon array leads to stronger intentions to act than showing it as numbers alone.
Part-to-whole icon arrays v. numbers alone	Moderate ($n = 8$)	People may be more likely to act on an 8-percentage-point risk difference if it is shown as 2 icon arrays (one showing a 90 in 100 chance and the other showing an 82 in 100 chance) rather than describing it as an 8% increase or decrease.	Showing a risk difference as a pair of side-by-side, part-to-whole icon arrays leads to stronger intentions to act than showing it as numbers alone.
Bar charts v. numbers alone	Moderate ($n = 4$)	It doesn't seem to matter for behavioral intention if a risk difference is shown as 8 in 100 or as a bar chart.	Showing a risk difference in bar charts does not appear to lead to any increase or decrease in behavioral intention compared with showing numbers alone.
Incremental risk icon array v. numbers alone	Insufficient— inconsistent findings ($n = 2$)	It is not clear whether showing a probability difference as a single icon array illustrating incremental risk, as compared with numbers alone, affects intentions to act.	

1-mo time frame than over a 12-mo time frame (rates per 10^n plus icon arrays). However, in this case of a smoking cessation intervention, the chance of relapse accumulated over time so that the stated probability of being smoke free was lower in the long term than the short term, which likely affected perceptions.

Effects of Different Formats for Probability Differences on Health Behaviors and Behavioral Intentions: Section 7

Promoting specific health behaviors is an important goal of health communication. In many studies we found, behavioral intentions were measured rather than actual behaviors. Because of the relative rarity of communication studies tracking actual behavior, and because of the theoretical grounding (in the health belief model and other health behavior theories) suggesting that intentions are correlated with behaviors, we have grouped intention and behavior in this section.

Comparisons between number formats for probability differences on health behavior and behavioral intention (subsection 7A). RELATIVE RISK DIFFERENCE VERSUS PAIRS OF PRE-POST PROBABILITIES: A large majority of findings including many high-credibility ones shows that presenting a risk difference as a relative risk difference alone or in combination with other statistics prompts stronger behavioral intention than presenting the difference as pairs of pre-post values or the arithmetic difference between them. Fourteen findings showed an effect (Goodyear-Smith et al.[24] substudy 1, Gyrd-Hanson et al.,[25] Berglund et al.,[26] Malenka et al.[27] substudy 1, Misselbrook and Armstrong,[28] Stovring et al.,[29] Sinsky et al.,[30] Sarfati et al.,[31] Heard et al.[32] substudy 1, Carling et al.,[33] Stone et al.[34] substudy 1 and 2, Berry et al.,[35] Hux and Naylor[36]). However, 3 findings[37–39] did not show a statistically significant effect.

NNT VERSUS PROBABILITY FORMATS: Four high-credibility (Goodyear-Smith et al.[24] substudy 1, Misselbrook and Armstrong,[28] Stovring et al.,[29] Sarfati et al.[31]) and 2 moderate-credibility[33,39] findings show that the number needed to treat (NNT) format is less effective at prompting behavioral intention than most other probability comparison formats. However, 1 high-credibility finding, Gyrd-Hansen et al.,[38] did not find this effect.

EFFECTIVE AGE VERSUS PROBABILITY FORMATS: Three findings about "heart age" or

Table 7D Evidence-Based Guidance for Comparisons of Numerical and Verbal Formats for Probability Differences on Health Behavior and Behavioral Intention

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Numbers v. verbal probabilities	Moderate ($n = 2$)	Changing the description of the effect of screening from numbers (a reduction of 1 in 10, 10%, or 10 in 100) to verbal probabilities ("reduced risk," etc.) does not appear to affect behavioral intention.	The decision about whether to describe probability differences as numbers alone or verbal terms alone does not appear to affect behavioral intentions in the context of screening.

Table 7E Evidence-Based Guidance for Effects of Context for Probability Differences on Health Behavior and Behavioral Intention

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Anecdotes	Strong ($n = 2$)	People are more likely to choose option A if numerical information is supplemented with anecdotes about people who experienced a benefit from it, such as a story about Annette who was successfully treated.	A numerical message about a therapeutic option is associated with stronger intention to act if it is supplemented with anecdotes about people who have experienced a benefit from it.
Normative message	Strong ($n = 2$)	People may be more likely to choose option A if numerical information is supplemented with a so-called normative message such as, "Most people like you choose option A."	A numerical message about options is associated with stronger intention to act if it is supplemented with a normative message about what others have chosen (when most others have chosen an option).
Verbal explanation of causality	Insufficient—too few findings ($n = 1$)	It is not clear whether adding a verbal description of causality (e.g., how a gene affects cancer risk) affects behavioral intentions (e.g., to get tested).	
Interpretive labels	Insufficient—too few findings ($n = 1$)	It is not clear whether adding verbal interpretive labels describing the size of an effect changes people's behavioral intentions.	
Visual emphasis	Insufficient—too few findings ($n = 1$)	It is not clear whether using cues such as bolding or red color to highlight numerical information about probability differences affects behavioral intentions.	

"effective age" as an alternative to probabilities do not strongly support the use of this concept to promote behavioral intention. In 2 high-credibility findings, Heard et al.[32] substudies 1 and 2 demonstrated that "effective age" prompted weaker behavioral intentions than relative risk reduction or as time lost from life (years or hours).

A single high-credibility finding, Bonner et al.,[40] showed no difference in behavioral intention between heart age and pairs of pre-post percentage.

DELAY OF EVENT VERSUS PROBABILITY FORMATS: Five high-credibility findings examined the related concepts of delay of event or prolongation/shortening of life. Although Berglund et al.[26] (high credibility) demonstrated that conveying a risk difference as years of delay of event led to behavioral intentions nearly as strong as with relative risk reduction, Stovring et al.[29] (high credibility) demonstrated that supplementing visuals with prolongation of life information was less effective in changing behavioral intention than adding relative risk reduction. Harmsen et al.[41] (high credibility) demonstrated that prolongation of life led to lower behavioral intentions than the pre-post probabilities or arithmetic difference between them. Heard et al.[32] substudy 1

Table 7F Evidence-Based Guidance for Framing of Probability Differences on Health Behavior and Behavioral Intention

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Loss frame v. gain frame for benefits	Strong ($n = 3$)	People will be more likely to choose an option if the message about its impact focuses on the negative outcome rather than the positive one. For example, therapy A will be more attractive with a statement such as, "This therapy reduces the chance of death from 10% to 5%" rather than a message such as, "This therapy increases the chance of survival from 90% to 95%."	With a communication about the beneficial effect of a potential therapy, more people will choose it when the information is loss-framed (stated in terms of effect on the chance of negative outcome) than when it is gain-framed information (stated in terms of effect on the chance of a beneficial outcome).
Loss frame v. gain frame for sure and risky options	Moderate ($n = 2$)	Sure option: Everybody survives radiation and lives on average for another year. A – Risky option in gain frame: 20% of people will survive surgery and will live an average of 5 y. B – Risky option in loss frame: 80% of people will die during surgery but the remainder will not die until an average of 5 y. The risky option (surgery) will seem more attractive when the options are in a gain frame (A) versus a loss frame (B).	In risky choice situations, when an option with an uncertain outcome is compared with a certain negative outcome, the uncertain therapy is more attractive in the gain frame. However, when an uncertain therapy is compared with a certain positive outcome, the uncertain therapy is more attractive in the loss frame.

Table 7J Evidence-Based Guidance on Effect of Varying the Time Period for Probability Differences on Health Behavior and Behavioral Intention

Comparison	Evidence Strength	General Guidance
Time frame	Insufficient—too few findings ($n = 1$)	It is unclear whether changing the time frames used to show treatment differences affects behavioral intentions.

(high credibility) demonstrated that "years of life lost" or "hours per day lost" resulted in stronger behavioral intentions than "effective age" but was no different from the percentage format. Yet 1 high-credibility finding, Gyrd-Hansen et al.,[38] did not find any differences in in willingness to accept treatment between prolongation of life and pairs of rates per 10^n, relative risk reduction, or NNT formats.

ROUND NUMBERS: Several moderate-credibility findings (Wadhwa and Zhang[12] substudies 1, 3 [behavior outcome], and 4) found small and borderline statistically significant increases in behavioral intention/behavior when relative risk reduction was presented in round numbers instead of nonround numbers. However, all numbers were presented to a precision of the nearest 100th

(e.g., 33.00 was considered a "round" number and 33.33 a "nonround" number).

OTHER FORMATS: Strong conclusions about the exact ranking of other probability or probability comparison formats (e.g., posttreatment probabilities, pairs of pre-post probabilities, arithmetic difference between them) do not appear possible from the existing evidence.

A high-credibility finding[42] demonstrated stronger behavioral intentions associated with 1 in X than with pairs of percentages. In 2 moderate-credibility findings, Ruiz et al.[43] demonstrated no difference in behavioral intention or self-reported behavior change whether risk-reduction information was shown as pairs of rate per 10^n or pairs of percentages. In a moderate-credibility finding,[44] willingness to pay for an improved drug was higher

Table 8A Evidence-Based Guidance on the Effect of Numerical Formats for Probability Differences on Trust

Comparison	Evidence Strength	General Guidance
Heart age or life expectancy v. probability	Insufficient—inconsistent findings ($n = 2$)	It is not clear whether different number formats (percentage probability, heart age, or life expectancy) for conveying probability differences affect trust in the information in consistent ways.

Table 8I Evidence-Based Guidance on the Effect of Animated or Interactive Formats for Probability Differences on Trust

Comparison	Evidence Strength	General Guidance
Animation/ interactivity	Insufficient—too few findings ($n = 1$)	It is not clear whether animation or interactivity affects trust in information on treatment effects or risk differences.

Table 9A Evidence-Based Guidance for Preferences for Numerical Formats for Probability Differences

Comparison	Evidence Strength	General Guidance
Number formats	Insufficient—inconsistent findings ($n = 6$)	It is not clear whether people consistently prefer a particular numerical format over others for presenting risk differences.
Years of life lost v. relative risk reduction	Insufficient—inconsistent findings ($n = 2$) evidence	It is not clear whether people prefer risk reduction information in terms of years of life lost or relative risk reduction.
Presence of baseline probability	Insufficient—too few findings ($n = 1$)	It is not clear whether people prefer the presence or absence of baseline probability information with risk difference communications.

with pairs of rate per 10^n than for pairs of decimals but only for 1 of 2 probability levels tested.

Lower-credibility findings (Wadhwa and Zhang[12] substudy 3 [behavioral intentions outcome] and Hembroff et al.[45]) are not synthesized due to the small sample size and lack of statistical significance testing, respectively.

Comparisons between graphical formats for probability differences in effect on health behavior and behavioral intention (subsection 7B). NUMERATOR-ONLY VERSUS PART-TO-WHOLE GRAPHICS: One high-credibility finding (Stone et al.[14] substudy 2) and 2 moderate-credibility findings (Okan et al.[13] substudy 1 and 2) demonstrated that behavioral intention around a risk-reducing action was stronger when benefit was shown as a numerator-only icon array rather than a part-to-whole icon array. One high-credibility finding[15] demonstrated a similar part-to-whole graphic effect but using bar charts with truncated *y*-axes and full *y*-axes.

DIFFERENT GRAPHICAL FORMATS: Two high-credibility findings[46,47] contrasted icon arrays with other graphical formats and demonstrated no effect on

behavioral intentions. A moderate-credibility finding[48] similarly compared icon arrays and bar charts, finding small differences in behavioral intention (likely not significant, but no specific testing was reported). Adarkwah et al.[49] similarly demonstrated no effect of different graphical formats on actual adherence to lifestyle change or medication at 3 mo. In a moderate-credibility finding, Stone et al.[50] substudy 2 showed no difference in willingness to pay for "improved" products by whether the improvement was depicted in foreground-only icon arrays or a pair of bars in a bar chart. Another moderate credibility finding[51] demonstrated no difference between risk reductions presented in consistent denominator part-whole icon arrays versus an NNT format with separate different denominator arrays for those treated or not treated. However, 1 high-credibility finding[52] demonstrated that showing the benefit of a drug as an integrated icon array showing the incremental risk difference led to lower intention to take a drug than a pair of bar charts did. Two of these findings[47,49] also demonstrated no difference between number lines, bar charts, icon arrays, or percentages.

Table 9B Evidence-Based Guidance for Preferences for Graphical Formats for Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Bar charts v. icon arrays	Strong (n = 8)	When evaluating a risk reduction from 10% to 8%, people will like a bar chart with bars representing 10% and 8% more than they will like a pair of icon arrays, one of which shows 10 affected individuals and the other of which shows 8.	People prefer bar charts over icon arrays for presenting probability differences.
Data labels	Moderate (n = 3)	People may prefer an icon array labeled with the probability difference (e.g., "an increase of 8 chances in 100") to an unlabeled icon array showing the same information.	Graphics with data labels (e.g., rate per 10^n numbers) are preferred to unlabeled graphics.
Pie charts v. bar charts	Weak (n = 2)	People may like pie charts more than bar charts for presenting differences in probability.	People may prefer pie charts over bar charts for presenting probability differences.
Icon arrays v. other graphics	Insufficient—too few findings (n = 2)	It is unclear whether people tend to prefer icon arrays to other sorts of graphics.	
Part-to-whole graphics v. foreground-only graphics	Insufficient— inconsistent findings (n = 3)	It is unclear whether people prefer part-to-whole versus foreground-only icon arrays for showing risk differences.	
Icon arrays that highlight one v. multiple outcomes	Insufficient too few findings (n = 1)	It is unclear whether people prefer icon arrays that highlight only 1 outcome over those that highlight multiple outcomes for showing probability differences.	
Postintervention probability data	Insufficient—too few findings (n = 1)	It is unclear whether adding postintervention probability data (thereby illustrating probability differences) affects preference.	

NUMERICAL LABELS: A moderate-credibility finding[17] found no change to behavioral intentions when number were labeled with rate per 10^n in addition to category labels. Similarly, Okan et al.[15] (moderate credibility) also did not find any differences in behavioral intention using bar charts with and without number labels in rate per 10^n format.

RANDOM VERSUS GROUPED ICONS IN ICON ARRAY: A moderate-credibility finding[53] showed no difference in intention to quit smoking by whether icons colored to represent the event were randomly arranged throughout the array or grouped in a block.

ABSOLUTE DIFFERENCES VERSUS RISK RATIOS: A moderate-credibility finding[18] demonstrated no consistent difference in support for policies to reduce population-level risk disparities when grouped bar charts showed absolute differences versus risk ratios.

POSTINTERVENTION PROBABILITY DATA: One study used a visual indicator to portray the probability of disease without treatment on a vertical number line as well as a second indicator showing the probability of disease with treatment. A moderate-credibility finding

found that adding the second indicator did not affect behavioral intentions.[17]

Comparisons between numerical and graphic formats for probability differences in effect on health behavior and behavioral intention (subsection 7C). NUMERATOR-ONLY ICON ARRAYS VERSUS NUMBERS ALONE: One high- (Stone et al.[14] substudy 2) and 5 moderate-credibility findings (Stone et al.[14] substudy 1, Stone et al.[50] substudies 1 and 2, and 2 findings from Schirillo and Stone[54] substudies 1 and 2) demonstrated that numerator-only icon arrays lead to higher willingness to pay than pairs of rate per 10^n.

PART-TO-WHOLE ICON ARRAYS VERSUS NUMBERS ALONE: Two high-credibility findings and 1 moderate-credibility finding suggest that icon array pairs alone or in combination with numbers appear to have stronger effects on behavioral intention than numbers alone (Cox et al.,[55] Vogt and Marteau[22] substudy 1, Timmermans et al.[56] in certain scenarios). Similarly, in a high-credibility finding, Chen et al.[42] demonstrated that

Table 9C Evidence-Based Guidance for Preferences for Numerical versus Graphical Formats for Probability Differences

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Numbers plus icon arrays v. numbers alone	Weak ($n = 5$)	For a probability difference of 8 percentage points, people may prefer the combination of "an increase from 4 per 100 to 12 per 100 increase" combined with an illustrative icon array over the numbers alone.	Numbers supplemented with icon arrays may be preferred over numbers alone for presenting risk differences.
Numbers alone v. graphics alone	Insufficient evidence—inconsistent findings ($n = 3$)	It is not clear whether numbers alone or graphics alone are preferred for portraying risk differences, in part because different sorts of graphics and combinations have been tested.	

Table 9E Preferences for Added Context for Probability Differences

Comparison	Evidence Strength	Guidance
Social comparison	Insufficient—too few findings ($n = 1$)	It is not clear whether adding social comparisons affects preferences for information about probability differences.

Table 9I Preferences about Animation and Interactivity

Comparison	Evidence Strength	Guidance
Animation	Insufficient—too few findings ($n = 1$)	It is not clear whether animation improves satisfaction with risk difference information. Any effects likely depend on type of animation.
Interactivity	Insufficient—too few findings ($n = 1$)	It is not clear whether interactivity improves satisfaction with risk difference information. Any effects likely depend on type of interactivity.

Table 9J Preferences about Time Period for Probability Differences

Comparison	Evidence Strength	General Guidance
Long v. short time periods	Insufficient—too few findings ($n = 1$)	It is not clear whether people prefer to compare risks expressed over a lifetime or over shorter time frames such as 10 y or 20 y.

an icon array pair led to stronger intention than a percentage but had weaker effects than 1 in X. A high-credibility finding and 2 moderate-credibility findings demonstrated no difference between a pair of icon arrays and numerical formats, but both studies had some limitations (small sample size for Ruiz et al.[43] [both behavioral intentions and behavior] and ceiling effects in a setting with high behavioral intention at baseline for Cameron et al.[57]). Another moderate-credibility finding[48] compared percentage with icon arrays and also did not show any

large effects, although there were no statistical tests of this comparison.

BAR CHARTS VERSUS NUMBERS ALONE: One high[52]- and 3 moderate-credibility[48,58,59] findings have no clear evidence of effect on behavioral intention.

INCREMENTAL-RISK ICON ARRAYS VERSUS NUMBERS ALONE: Two findings compared numbers to part-to-whole icon arrays that displayed incremental risk, with no clear consensus. A high-credibility finding[52] showed that an incremental icon array led to lower

intention to take a drug than a pair of bar charts or percentage. However, a moderate-credibility finding[60] used a variation of the incremental icon array (an image of a football stadium with dots representing people), finding that it did not increase intention above just presenting the numbers of people affected.

A lower-credibility finding by Goodyear-Smith et al.[39] was not synthesized due to concerns about small sample size.

Comparisons between numerical and verbal formats for probability differences in effect on health behavior and behavioral intention (subsection 7D). NUMBERS VERSUS VERBAL PROBABILITIES: Schwartz et al.[61] had 1 finding of no difference to colorectal cancer screening uptake within 6 mo (high-credibility finding) and a second finding for overall colorectal cancer screening intention (moderate-credibility finding) with quantitative risk and benefit information (percentage and rate per 10^n) versus ordinal verbal probabilities. However, with this second finding, there was a higher intention to pursue fecal immunochemical testing specifically.

Comparisons of elements added for context to probability differences on health behavior and behavioral intention (subsection 7E). ANECDOTES: Two high-credibility findings looked at the impact of anecdotes, both showing that adding anecdotes to numerical information affected behavioral intention. Ubel et al.[62] study 1 showed that additional anecdotes describing 1 of 2 therapeutic options increased intention to choose that option. Ubel et al.[62] study 2 showed that with no anecdotes, people tended to choose the option with the highest chance of success and that adding anecdotes reduced the intention to choose the most effective option.

NORMATIVE MESSAGE: Two high-credibility findings examined the impact of providing a normative message (what people "like you" chose), both finding an effect under different circumstances. Zikmund-Fisher et al.[63] substudies 1 and 2 showed that the normative message influenced behavioral intention when it was a large numerical proportion but not when it was a small one and when the proportion was not numerical but qualitative ("most women").

VERBAL EXPLANATION OF CAUSALITY, INTERPRETIVE LABELS, VISUAL EMPHASIS: Three moderate-credibility findings examined a variety of different types of information meant to provide context: verbal explanation of causality,[57] interpretive labels,[17] and visual emphasis including color coding and

boldface font (Stone et al.[14] substudy 1). All of these showed no impact, but credibility is reduced by modest sample sizes.

Comparisons of frames (gain, loss, comparison) for probability differences on health behavior and behavioral intention (subsection 7F). Gain-framing or loss-framing the same probability information strongly affects behavioral intention and intended decisions.

LOSS FRAME VERSUS GAIN FRAME FOR BENEFITS: When given information about a medical therapy, a high-credibility finding[64] showed the therapy was more attractive when the benefit outcome was in the loss frame (chance of getting disease) than in the gain frame (chance of avoiding disease). Similarly, in a moderate-credibility finding,[58] when given the choice between 2 products, the safer product was more attractive when the improvement was presented as a pair of negatively framed numbers (chance of disease). However, a moderate-credibility finding[51] showed no effect of framing on intent to take a medication, although the small sample size lowers confidence in this negative result.

LOSS FRAME VERSUS GAIN FRAME FOR SURE AND RISKY OPTIONS: One high-credibility finding (Wilson et al.[65] study 1) looked at choices between a sure option and a risky option, finding that the risky option was more attractive in the gain (survival) frame. In a related moderate-credibility finding (Wilson et al.[65] study 3), when given the option to remove a patient in special care to the regular floor to make room for another needy patient, the status quo (presumably seen as a positive thing for the current patient) was more attractive when probabilities were loss framed (mortality).

A lower-credibility finding by Goodyear-Smith et al.[39] is not summarized here due to a small sample size and lack of statistical testing.

7J. Effect of varying the time period on health behavior and behavioral intention. One moderate-credibility finding (Vogt and Marteau[22] study 2) compared the effect of different time frames for probability difference information on behavioral intention, finding no effect, but internal inconsistencies between different outcome measures reduce ability to draw broader conclusions.

Effects of Different Formats on Trust in the Message: Section 8

We found multiple instances of researchers assessing whether format affected the perceived credibility or

trustworthiness of the information, under the assumption that high-credibility messages were more likely to promote changes in beliefs or behaviors.

Comparisons between number formats for probability differences in effect on trust (subsection 8A). HEART AGE OR LIFE EXPECTANCY VERSUS PROBABILITY: One high-credibility finding[40] showed that showing a probability difference with pre/post percentages was associated with higher trust than showing it as "heart age" plus the arithmetic difference between heart age and actual age. However, a moderate-credibility finding suggested a different effect: Galesic and Garcia-Retamero[11] showed that the "imaginability" of the outcome was higher for life expectancy (plus the difference in life expectancy (". . . 73 years, or 60 months shorter than average person . . .") than percentage probability (plus a difference in percentage from the average person).

Comparisons between numerical and graphical formats for probability differences in their effects on trust (subsection 8C). One potentially relevant finding is not summarized due to problems with the graphic design of the graphic comparator (a bar chart depicting relative risks as bars with a baseline of 0, rather than in comparison to a baseline of 1).[66]

Comparisons of animated or interactive formats for probability differences on trust (subsection 8I). One moderate-credibility finding[67] showed no effect of various forms of animation in icon arrays on credibility of information.

Preferences about Formats for Portraying Probability Differences (Section 9)

Much research has examined the information recipient's perceptions of the attractiveness of or helpfulness of a format for probability difference as a primary or secondary measure because it seems possible that a patient might be more attentive to or receptive of information presented in a preferred format.

Preferences for different number formats for probability differences (subsection 9A). NUMBER FORMATS: There is no coherent message from the available findings comparing different numerical formats for communicating probability differences. A high-credibility finding[68]

showed a preference for pairs of 1 in X over pairs of percentages for very small risk differences associated with genetic testing. A moderate-credibility finding (Goodyear-Smith et al.[24] substudy 1) showed relative risk reduction as percentage was preferred to a pair of pre/post percentages. However, 4 additional moderate- and high-credibility findings showed no differences between different numerical risk reduction formats.[5,26,33,41]

YEARS OF LIFE LOST VERSUS RELATIVE RISK REDUCTION: In a moderate-credibility finding, people preferred years lost of life to relative risk reduction as percentage or change in effective age (Heard et al.[32] substudy 1). However, relative risk reduction as percentage was preferred to effective age in another moderate-credibility finding in the same study (Heard et al.[32] substudy 2)

PRESENCE OF BASELINE PROBABILITY: In a moderate-credibility finding, Berry et al.[35] showed that in communication of probability differences, people preferred the presence of baseline probability to its absence.

Findings from Miron-Shatz et al.[69] were considered lower credibility because of the small sample and differences between the denominators in the stimuli. Both Goodyear-Smith et al.[39] and Hill et al.[70] had small samples and lack of statistical testing, and Selinger et al.[71] had a very small sample.

Preferences for different graphic formats for portraying probability differences (subsection 9B). BAR CHARTS VERSUS ICON ARRAYS: Eight findings compared the portrayal of probability differences in bar charts and icon arrays,[47,72–78] with 7 of these (2 high-credibility and 5 moderate) showing various degrees of preference for bar charts.

DATA LABELS: One high-credibility finding[15] showed bar charts with data labels (rate per 10^n format) were preferred to those that did not. A moderate-credibility finding[17] similarly showed that number lines were rated as easier to understand when labeled with rate per 10^n in addition to category labels. Similarly, another moderate-credibility finding[79] determined that icon arrays were evaluated more positively when supplemented with rate per 10^n labels.

PIE CHARTS VERSUS BAR CHARTS: A high-credibility and a moderate-credibility finding[73,74] compared pie charts to bar charts, in both cases finding a preference for pie charts. Comparisons to number lines, flow charts, and other graphic formats produced less consistent findings.

ICON ARRAYS VERSUS OTHER GRAPHICS: Both a high-credibility finding[80] and a moderate-credibility finding[20] showed no differences in preference

Table 10A Evidence-Based Guidance about the Effects of Numerical Formats for Probability Differences on Discrimination

Comparison	Evidence Strength	General Guidance
Percentages v. rates per 10^n	Insufficient evidence—too few findings ($n = 1$)	It is not clear whether rates per 10^n or percentage formats make people more sensitive to small probability differences.

between different graphical formats, including icon arrays versus time-to-event bars[80] and icon arrays versus speedometer graphics.[20]

PART-TO-WHOLE VERSUS FOREGROUND-ONLY GRAPHICS: A high-credibility finding showed preference for side-by-side part-to-whole icon arrays rather than foreground-only icon arrays (Okan et al.[13] substudy 2). However, 2 similarly designed, high-credibility findings[15,81] showed no difference in preference for part-to-whole horizontal bars versus foreground-only horizontal bars.

ICON ARRAY HIGHLIGHTING ONE VERSUS MULTIPLE OUTCOMES: A moderate-credibility finding showed preference for an integrated multioutcome icon array that highlighted risk differences over pairs of icon arrays.[82] The same finding showed a preference for icon arrays with background shading versus those without shading.

POSTINTERVENTION PROBABILITY DATA: An additional moderate-credibility finding found lower perceived understanding/believability when postintervention probabilities (thereby visually illustrating the absolute risk difference) were added to vertical number lines.[17]

Two low-credibility findings were not included in the synthesis (Dolan et al.[83] and Feldman-Stewart et al.[84] substudy 1) due to small sample sizes. Another relevant finding from Emmons et al.[85] was considered lower credibility because of the inconsistency between 2 measures of preference.

Preferences for numerical versus graphical formats for probability differences (subsection 9C). NUMBERS PLUS ICON ARRAYS VERSUS NUMBERS ALONE: A high-credibility finding[9] showed that rates per 10^n plus icon arrays were preferred over number-only formats. However, a high-credibility finding from a very similar study using a different population[86] showed that a pair of rates per 10^n in table format was preferred to a pair of rates per 10^n plus icon arrays. Also, 2 moderate-credibility findings showed that people preferred pairs of numbers supplemented with icon arrays to numbers alone (pre-post rates per 10^n for Cox et al.,[60] pre-post

percentages for Garcia-Retamero et al.[5]). However, another finding showed no difference between rate per 10^n numbers alone and numbers supplemented with icon arrays.[55]

NUMBERS ALONE VERSUS GRAPHICS ALONE: A moderate-credibility finding[77] showed that numbers (rate per 10^n) or a number-based flowchart were preferred to icon arrays, number lines, or bar charts. van Weert et al.[74] (moderate credibility) showed a preference for pie charts with radial clocklike labels over a table containing rate per 10^n numbers (and other graphics including icon arrays), and another moderate-credibility finding comparing multiple formats showed no preference difference by format.[20]

Lower-credibility findings from Goodyear-Smith et al., Hill et al., Sellinger et al., and Feldman-Stewart et al. were not included in the synthesis due to small sample sizes and lack of statistical testing.[39,70,71,84]

Preferences for added elements for context for probability differences (subsection 9E). SOCIAL COMPARISON: A moderate-credibility finding[17] showed adding social comparison information (e.g., "somewhat higher than average") actually reduced ease of understanding/believability.

Preferences for gain-loss framing of probability differences (subsection 9F). One finding from Goodyear-Smith et al.[39] was relevant but is not summarized due to the small sample size.

Preferences for animation or interactivity for probability differences (subsection 9I). Two findings, both moderate credibility and with different interventions, are available examining the effects of animation or interactivity on preferences for communication of probability differences.

ANIMATION: Okan et al.[79] showed that animated icon arrays were evaluated more positively than static ones.

INTERACTIVITY: Okan also showed that prompting interaction with the information via reflective

Table 10C Evidence-Based Guidance about Effects of Numerical and Graphical Formats on Discrimination

Comparison	Evidence Strength	General Guidance
Bar charts plus numbers v. numbers alone	Insufficient—too few findings ($n = 1$)	It is unclear whether numbers or numbers combined with graphics are better for helping people distinguish small differences between 2 probabilities.

Table 10E Evidence-Based Guidance about Effects of Context on Discrimination

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Category labels	Weak ($n = 1$)	Labeling a probability difference as "positive/abnormal" increases people's ability to detect the difference. However, this may cause an overreaction.	Adding interpretive labels to numerical effect information may increase sensitivity to changes in probability.
Explanations and examples	Insufficient—too few findings ($n = 1$)	It is unclear whether providing explanations regarding when to consider numbers as approximately equal changes sensitivity to probability effects.	

questions lead to higher evaluations but only among less graph-literate respondents.

Preferences for long versus short time periods for probability differences (subsection 9J). LONG VERSUS SHORT TIME PERIODS: A moderate-credibility finding[78] studying communication of treatment effects showed a preference for lifetime risk over 10-y or 20-y risk statistics.

Discrimination of Probability Differences: Section 10

Discrimination is assessed by determining whether the format enables small differences in the probability difference to influence responses to the message.

Comparison between numerical formats for probability differences in effects on discrimination (subsection 10A). PERCENTAGES VERSUS RATES PER 10^n: Only 1 moderate-credibility finding (Wolfe et al.[87] substudy 1) looked at the effect of format on people's judgments of probabilities as "approximately equal." Women were more likely to judge a probability difference as "approximately equal" when it was portrayed as 2 percentages than as rates per 10^n, suggesting that discrimination was higher with rate per 10^n.

Comparison between numerical and graphical formats for probability differences in effects on discrimination (subsection 10C). BAR CHARTS PLUS NUMBERS VERSUS NUMBERS ALONE: Only 1 moderate-credibility finding (Wolfe et al.[87] study 2) examined the impact of numbers versus graphics on people's ability to make a distinction between 2 probabilities. The finding showed that adding a bar chart to the numbers made no difference. No other graphical formats were assessed, making it difficult to draw any broad conclusions about graphics.

Comparisons of elements added for context on sensitivity to deviation (subsection 10E). CATEGORY LABELS: A high-credibility finding[88] showed that adding an interpretive label of "positive/abnormal" or "negative/normal" to a rates per 10^n produced an interaction such that people had stronger reactions to whether genetic test results had increased or lowered their risk. This increased discrimination but in fact created a bias not found with numbers-only communications.

EXPLANATIONS AND EXAMPLES: A moderate-credibility finding (Wolfe et al.[87] study 2) showed that providing interpretive instructions that provided examples of small differences that were nevertheless "about equal" reduced sensitivity to small differences between numbers.

Summary of Evidence

Strong or moderate evidence about *perceptions* of probability difference communications and *feelings about effectiveness* includes:

- Relative differences (with or without the baseline and post–risk factor risks) increase perceptions of effectiveness.
- Numerator-only graphics (foreground-only design) make a difference in probability appear larger than part-to-whole graphics.
- Baseline probability plus an absolute difference reduces concern compared with presenting it as a pair of absolute rates.
- Adding part-to-whole graphics (icon array, bar chart) to pairs of probabilities (rate per 10^n, percentage) may not affect perceptions of probability differences.
- Showing a foreground-only risk graphic instead of pairs of rates per 10^n increases effectiveness perceptions.

Strong or moderate evidence about the effects of probability difference formats on *behavioral intention* included:

- Relative differences (alone or in combination with other statistics) have a greater effect on behavioral intention than pairs of pre-post probabilities or the arithmetic difference between them.
- Showing a probability difference in familiar formats (such as rates per 10^n, 1 in X, or percentage) is more effective at prompting behavioral intentions than showing it as NNT.
- Relative risk reductions have stronger effects on behavioral intention than "heart age" or "effective age" formats.
- Showing a probability difference effect as a numerator-only icon array or bar chart, rather than as a part-to-whole graphic, is associated with stronger intention to act.
- No specific form of probability difference graphic appears to consistently result in stronger or weaker behavioral intentions.
- Numerator-only icon arrays to illustrate probability differences lead to stronger intentions to act than numbers alone do.
- A pair of side-by-side, part-to-whole icon arrays to illustrate probability differences leads to stronger intentions to act than numbers alone do.

- Bar charts to illustrate probability differences do not appear to lead to any increase or decrease in behavioral intention compared with numbers alone.
- Probability differences related to screening tests as pairs of numbers versus verbal terms does not appear to affect screening behavioral intentions/behavior.
- A numerical probability difference message about a therapeutic option is associated with stronger intention to act if it is supplemented with anecdotes about people who have experienced a benefit from it.
- A numerical probability difference message about options is associated with stronger intention to act if it is supplemented with a normative message about what most others have chosen.
- With a communication about the effect of a potential therapy, more people will choose it when the probability difference information is loss-framed (stated in terms of the effect on the chance of negative outcome) than when it is gain-framed information (stated in terms of the effect on the chance of a beneficial outcome).
- In risky choice situations with multiple options, each with differing chances of success, the framing of the probability difference associated with the "sure" option (the one with either 0% or 100% likelihood) is particularly important. When an uncertain therapy is compared with a certain negative outcome, the therapy is more attractive in the gain frame. However, when an uncertain therapy is compared with a certain positive outcome, the therapy is more attractive in the loss frame.

Strong or moderate evidence for people's preferences for different formats for communicating probability differences:

- People prefer bar charts over icon arrays for presenting probability differences.
- Graphics with data labels (e.g., rate per 10^n numbers) may be preferred to unlabeled graphics for communicating probability differences.

We are unable to present anything but insufficient evidence about the effects of probability difference formats on *trust* or *discrimination* outcomes.

Discussion

The current article focuses on the important task of helping patients evaluate probability differences, such as the

information used to express the effect of a risk factor, therapy, or exposure. Such probability difference information is key to making informed judgments about the selection of therapies, the choice of whether to undergo disease screening, and the decision to avoid certain behaviors or exposures. This article presents the evidence on the effects of numerical and graphical formats for probability differences on 5 important outcomes: effectiveness perceptions/feelings, behaviors/behavioral intentions, trust, preference, and discrimination. (The remaining outcomes of identification/recall, contrast, categorization, and computation appear in the part 1 article.)

Research on difference tasks has focused heavily on behaviors and behavioral intentions (94 findings). Also, a large amount of work has focused on peoples' preferences for graphical and numerical formats (53 findings) and perceptions of and feelings about effectiveness (50 findings). Very little work has been done on how different formats affect trust or discrimination. Almost all the available research has compared 2 or more numerical formats for communicating probability differences (66 findings), 2 or more graphics (52 findings), or graphics versus numbers (56 findings), rather than other sorts of features.

Unlike several of the other review articles from this systematic review project, the current article presents a relatively large amount of strong or moderately strong evidence about the effects of formats for communicating probability differences on *perceptions and feelings about effectiveness* and *behavioral intention*. The evidence confirms that relative risk numbers have a greater impact on perceived/felt effectiveness and on behavioral intentions than pairs of absolute probabilities, that numerator-only graphics have a greater impact on these outcomes than part-to-whole graphics, and that side-by-side icon arrays have a greater impact than numbers alone. Adding personal anecdotes or social norm information about the choices others have made similarly increases the impact on perceptions and behavioral intentions. Conversely, the impact on perceptions is smaller with the incremental difference (e.g., a 2-percentage-point increase from the baseline 5% chance) rather than the pre and post probabilities (e.g., an increase from 5% to 7%). Somewhat less familiar formats including the NNT and heart age or effective age have less effect on behavioral intention than relative risk reductions. Framing influences behavioral intention, specifically in that people are more likely to choose a therapy when its effect is loss framed or expressed in terms of its impact on the chance of a negative outcome than when it is gain framed or expressed in terms of its impact on the chance of a good outcome.

We do not draw conclusions here about whether one format is "better" or "worse" than another overall but rather recommend that communicators take these effects into consideration when they determine which format to select. (We also reiterate that the set of evidence presented here is about whether effectiveness is perceived as larger or smaller; the question of whether participants could accurately identify or recall a specific probability difference is covered in our companion part 1 article.)

This review, unlike some others in our series, revealed some moderate and strong evidence about *preferences*. People preferred bar charts to icon arrays, and graphics with labels to unlabeled graphics when evaluating probability differences. However, it is worth noting that part 1 of this review uncovered weak evidence that icon arrays or pie charts instead of bar charts may help people contrast probability differences to identify the treatment with the greatest effect. These contrasting findings provide yet another example of the fact that no probability communication format is optimal for all outcomes.

Although it seems plausible that trust in the information about a probability difference could affect the acceptance of it, there is little research on trust as an outcome. There is also little evidence about discrimination, that is, the extent to which people can distinguish between small differences in the size of probability differences. Finally, most of the research relevant to behavior measured behavioral intention rather than actual behavior, reducing the strength of the evidence relevant to health behavior.

This article is limited by the same factors that are limitations for the Making Numbers Meaningful project as a whole, including potential missing studies. Another limitation is the use of a small group of experts to evaluate finding credibility according to a holistic appraisal of a list of study and finding characteristics rather than a weighted quantitative checklist. Also, we conducted highly granular data extraction that enabled narrow comparisons of very similar studies (studies with the same task, outcome, and format comparison) rather than more global assessments. Although this approach enables more precise comparisons, it also prevents us from providing simple answers to very broad questions such as, "Are icon arrays better in general than bar charts?" Also, the studies did not collect the same sets of participant sociodemographics, use the same instruments to collect them, or use them in similar ways. That is, many studies that reported demographics did not use them in the analysis, while others used demographic variables to guide sampling, others controlled for them in the analysis, and a very few stratified analyses to obtain stratum-specific estimates. To handle this heterogeneity, we created free-

text summaries of demographic characteristics rather than structured abstractions, and as a result, we were not able to confirm that samples were sufficiently diverse to generalize to all potential populations. However, we recognize that the effects of data presentation format might differ by participant sociodemographic and cultural characteristics, and we encourage future researchers to extend this work by designing studies to assess these impacts among diverse populations. Future evidence synthesis work could then begin to systematically analyze findings by education level, age, numeracy, culture, or other potentially important factors.

In addition, we note that our decision to separate research findings related to point tasks from those related to difference tasks means that we do not present evidence on whether people prefer point versus difference tasks. That means we considered out of scope a research question such as, "Do people prefer to evaluate information about their absolute risk of disease or information about their risk relative to the average person's risk?" In the framework we presented above, assessing a single probability (such as absolute disease risk) is a very different task from assessing a relative difference between probabilities and involves different types of information. Such comparisons did not meet our inclusion criteria of the same information presented in different formats (see the "Methods" section). As a result, we excluded such comparisons from our review. An example is found in Emmons et al.[85]; many of the comparisons in this article did meet our inclusion criteria, but we excluded the comparison of preferences about current risk (a point task) and about current risk plus postintervention risk (a difference task).

In conclusion, a number of different formats have been demonstrated to affect perceived and felt effectiveness and behavioral intentions, and there is strong evidence for certain preferences for formats among patients and lay people. Evidence is weaker or lacking for the effect of format on behavior (rather than behavioral intention), trust in the information, and ability to discriminate between small differences in effects.

ORCID iDs

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

Abstracted data are freely available in the online appendix referenced in the "Results" section of this article. 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

1. 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

2. 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

3. Gigerenzer G. What are natural frequencies? *BMJ*. 2011;343:d6386. DOI: 10.1136/bmj.d6386

4. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med*. 1998;73(5):538–40.

5. 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.

6. Davis JJ, Isaacson SM. Responses to alternative forms of efficacy descriptions in direct-to-consumer pharmaceutical communications. *J Med Mark*. 2014;14(1):57–67. DOI: 10.1177/1745790414541576

7. Zikmund-Fisher BJ, Fagerlin A, Roberts TR, Derry HA, Ubel PA. Alternate methods of framing information about medication side effects: incremental risk versus total risk occurrence. *J Health Commun*. 2008;13(2):107–24.

8. Zikmund-Fisher BJ, Ubel PA, Smith DM, et al. Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. *Patient Educ Couns*. 2008;73(2):209–214.

9. 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

10. Blalock SJ, Sage A, Bitonti M, Patel P, Dickinson R, Knapp P. Communicating information concerning potential medication harms and benefits: what gist do numbers convey? *Patient Educ Couns*. 2016;99(12):1964–70. DOI: 10.1016/j.pec.2016.07.022

11. Galesic M, Garcia-Retamero R. Communicating consequences of risky behaviors: life expectancy versus risk of disease. *Patient Educ Couns*. 2011;82(1):30–35. DOI: 10.1016/j.pec.2010.02.008

12. Wadhwa M, Zhang K. When numbers make you feel: impact of round versus precise numbers on preventive health behaviors. *Organ Behav Hum Decis Process*. 2019;150:101–111.

13. Okan Y, Stone ER, Parillo J, Bruine de Bruin W, Parker AM. Probability size matters: the effect of foreground-only versus foreground + background graphs on risk aversion diminishes with larger probabilities. *Risk Anal*. 2020;40(4): 771–88. DOI: 10.1111/risa.13431

14. Stone ER, Reeder EC, Parillo J, Long C, Walb L. Salience versus proportional reasoning: rethinking the mechanism behind graphical display effects. *J Behav Decis Mak*. 2018;31(4):473–86. DOI: 10.1002/bdm.2051

15. Okan Y, Stone ER, Bruine de Bruin W. Designing graphs that promote both risk understanding and behavior change. *Risk Anal*. 2018;38(5):929–46. DOI: 10.1111/risa.12895

16. Stone ER, Sieck WR, Bull BE, Yates JF, Parks SC, Rush CJ. Foreground:background salience: explaining the effects of graphical displays on risk avoidance. *Organ Behav Hum Decis Process*. 2003;90(1):19–36. DOI: 10.1016/S0749-5978(03)00003-7

17. Janssen E, Ruiter RAC, Waters EA. Combining risk communication strategies to simultaneously convey the risks of four diseases associated with physical inactivity to sociodemographically diverse populations. *J Behav Med*. 2018;41(3):318–32. DOI: 10.1007/s10865-017-9894-3

18. 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

19. Dragicevic P, Jansen Y. Blinded with science or informed by charts? A replication study. *IEEE Trans Vis Comput Graph*. 2018;24(1):781–90. DOI: 10.1109/TVCG.2017.2744298

20. Martin RW, Brower ME, Geralds A, Gallagher PJ, Tellinghuisen DJ. An experimental evaluation of patient decision aid design to communicate the effects of medications on the rate of progression of structural joint damage in rheumatoid arthritis. *Patient Educ Couns*. 2012;86(3):329–34. DOI: 10.1016/j.pec.2011.06.001

21. Garcia-Retamero R, Dhami MK. Pictures speak louder than numbers: on communicating medical risks to immigrants with limited non-native language proficiency. *Health Expect*. 2011;14(suppl 1):46–57. DOI: 10.1111/j.1369-7625.2011.00670.x

22. Vogt F, Marteau TM. Perceived effectiveness of stop smoking interventions: impact of presenting evidence using numbers, visual displays, and different timeframes. *Nicotine Tob Res*. 2012;14(2):200–208.

23. Fagerlin A, Zikmund-Fisher BJ, Ubel PA. "If I'm better than average, then I'm ok?" Comparative information influences beliefs about risk and benefits. *Patient Educ Couns*. 2007;69(1–3):140–4. DOI: 10.1016/j.pec.2007.08.008

24. Goodyear-Smith F, Kenealy T, Wells S, Arroll B, Horsburgh M. Patients' preferences for ways to communicate benefits of cardiovascular medication. *Ann Fam Med*. 2011;9(2):121–7. DOI: 10.1370/afm.1193

25. Gyrd-Hansen D, Kristiansen IS, Nexøe J, Nielsen JB. How do individuals apply risk information when choosing among health care interventions? *Risk Anal*. 2003;23(4): 697–704. DOI: 10.1111/1539-6924.00348

26. Berglund E, Westerling R, Sundström J, Lytsy P. Treatment effect expressed as the novel delay of event measure is associated with high willingness to initiate preventive treatment—a randomized survey experiment comparing effect measures. *Patient Educ Couns*. 2016;99(12):2005–11. DOI: 10.1016/j.pec.2016.07.028

27. Malenka DJ, Baron JA, Johansen S, Wahrenberger JW, Ross JM. The framing effect of relative and absolute risk. *J Gen Intern Med*. 1993;8(10):543–8.

28. Misselbrook D, Armstrong D. Patients' responses to risk information about the benefits of treating hypertension. *Br J Gen Pract*. 2001;51(465):276–9.

29. Stovring H, Gyrd-Hansen D, Kristiansen IS, Nexoe J, Nielsen JB. Communicating effectiveness of intervention for chronic diseases: what single format can replace comprehensive information? *BMC Med Inform Decis Mak*. 2008;8: 25. DOI: 10.1186/1472-6947-8-25

30. Sinsky CA, Foreman-Hoffman V, Cram P. The impact of expressions of treatment efficacy and out-of-pocket expenses on patient and physician interest in osteoporosis treatment: implications for pay-for-performance programs. *J Gen Intern Med*. 2008;23(2):164–8. DOI: 10.1007/s11606-007-0490-z

31. Sarfati D, Howden-Chapman P, Woodward A, Salmond C. Does the frame affect the picture? A study into how attitudes to screening for cancer care are affected by the way benefits are expressed. *J Med Screen*. 1998;5(3):137–40. DOI: 10.1136/jms.5.3.137

32. Heard CL, Rakow T, Spiegelhalter D. Comparing comprehension and perception for alternative speed-of-ageing and standard hazard ratio formats. *Appl Cogn Psychol*. 2018;32(1):81–93. DOI: 10.1002/acp.3381

33. Carling C, Kristoffersen DT, Herrin J, et al. How should the impact of different presentations of treatment effects on patient choice be evaluated? A pilot randomized trial. *PLoS One*. 2008;3(11):e3693. DOI: 10.1371/journal.pone.0003693

34. Stone ER, Yates JF, Parker AM. Risk communication: absolute versus relative expressions of low-probability risks. *Organ Behav Hum Decis Process*. 1994;60(3):387–408. DOI: 10.1006/obhd.1994.1091

35. Berry DC, Knapp P, Raynor T. Expressing medicine side effects: assessing the effectiveness of absolute risk, relative risk, and number needed to harm, and the provision of baseline risk information. *Patient Educ Couns*. 2006; 63(1–2):89–96. DOI: 10.1016/j.pec.2005.09.003

36. Hux JE, Naylor CD. Communicating the benefits of chronic preventive therapy: does the format of efficacy data determine patients' acceptance of treatment? *Med Decis Making*. 1995;15(2):152–7. DOI: 10.1177/0272989X9501500208

37. O'Donoghue AC, Sullivan HW, Aikin KJ, Chowdhury D, Moultrie RR, Rupert DJ. Presenting efficacy information in direct-to-consumer prescription drug advertisements. *Patient Educ Couns*. 2014;95(2):271–80. DOI: 10.1016/j.pec.2013.12.010

38. Gyrd-Hansen D, Halvorsen P, Nexøe J, Nielsen J, Støvring H, Kristiansen I. Joint and separate evaluation of risk reduction: impact on sensitivity to risk reduction magnitude in the context of 4 different risk information formats. *Med Decis Making*. 2011;31(1):E1–10. DOI: 10.1177/0272989X10391268

39. Goodyear-Smith F, Arroll B, Chan L, Jackson R, Wells S, Kenealy T. Patients prefer pictures to numbers to express cardiovascular benefit from treatment. *Ann Fam Med*. 2008;6(3):213–7. DOI: 10.1370/afm.795

40. Bonner C, Jansen J, Newell BR, et al. Is the "heart age" concept helpful or harmful compared to absolute cardiovascular disease risk? An experimental study. *Med Decis Making*. 2015;35(8):967–78. DOI: 10.1177/0272989X15597224

41. Harmsen CG, Kristiansen IS, Larsen PV, et al. Communicating risk using absolute risk reduction or prolongation of life formats: cluster-randomised trial in general practice. *Br J Gen Pract*. 2014;64(621):e199–207. DOI: 10.3399/bjgp14X677824

42. Chen JC, Cooper RJ, Lopez-O'Sullivan A, Schriger DL. Measuring patient tolerance for future adverse events in low-risk emergency department chest pain patients. *Ann Emerg Med*. 2014;64(2):127–36.e3. DOI: 10.1016/j.annemergmed.2013.12.025

43. Ruiz JG, Andrade AD, Garcia-Retamero R, Anam R, Rodriguez R, Sharit J. Communicating global cardiovascular risk: are icon arrays better than numerical estimates in improving understanding, recall and perception of risk? *Patient Educ Couns*. 2013;93(3):394–402. DOI: 10.1016/j.pec.2013.06.026

44. Siegrist M. Communicating low risk magnitudes: incidence rates expressed as frequency versus rates expressed as probability. *Risk Anal*. 1997;17(4):507–510. DOI: 10.1111/j.1539-6924.1997.tb00891.x

45. Hembroff LA, Holmes-Rovner M, Wills CE. Treatment decision-making and the form of risk communication: results of a factorial survey. *BMC Med Inform Decis Mak*. 2004;4:20. DOI: 10.1186/1472-6947-4-20

46. Sullivan HW, O'Donoghue AC, Aikin KJ, Chowdhury D, Moultrie RR, Rupert DJ. Visual presentations of efficacy data in direct-to-consumer prescription drug print and television advertisements: a randomized study. *Patient Educ Couns*. 2016;99(5):790–9. DOI: 10.1016/j.pec.2015.12.015

47. Masson G, Mills K, Griffin SJ, et al. A randomised controlled trial of the effect of providing online risk information and lifestyle advice for the most common preventable cancers. *Prev Med*. 2020;138:106154. DOI: 10.1016/j.ypmed.2020.106154

48. 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

49. Adarkwah CC, Jegan N, Heinzel-Gutenbrunner M, et al. The Optimizing-Risk-Communication (OptRisk) randomized trial—impact of decision-aid-based consultation on adherence and perception of cardiovascular risk. *Patient Prefer Adherence*. 2019;13:441–52. DOI: 10.2147/PPA.S197545

50. Stone ER, Yates JF, Parker AM. Effects of numerical and graphical displays on professed risk-taking behavior. *J Exp Psychol Appl*. 1997;3(4):243–56. DOI: 10.1037/1076-898X.3.4.243

51. Kalluru R, Petrie KJ, Grey A, et al. Randomised trial assessing the impact of framing of fracture risk and osteoporosis treatment benefits in patients undergoing bone densitometry. *BMJ Open*. 2017;7(2):e013703. DOI: 10.1136/bmjopen-2016-013703

52. Navar AM, Wang TY, Mi X, et al. Influence of cardiovascular risk communication tools and presentation formats on patient perceptions and preferences. *JAMA Cardiol*. 2018;3(12):1192–9. DOI: 10.1001/jamacardio.2018.3680

53. Wright AJ, Takeichi C, Whitwell SCL, Hankins M, Marteau TM. The impact of genetic testing for Crohn's disease, risk magnitude and graphical format on motivation to stop smoking: an experimental analogue study. *Clin Genet*. 2008;73(4):306–314. DOI: 10.1111/j.1399-0004.2008.00964.x

54. Schirillo JA, Stone ER. The greater ability of graphical versus numerical displays to increase risk avoidance involves a common mechanism. *Risk Anal*. 2005;25(3):555–66.

55. Cox DS, Cox AD, Sturm L, Zimet G. Behavioral interventions to increase HPV vaccination acceptability among mothers of young girls. *Health Psychol*. 2010;29(1):29–39. DOI: 10.1037/a0016942

56. Timmermans DRM, Ockhuysen-Vermey CF, Henneman L. Presenting health risk information in different formats: the effect on participants' cognitive and emotional evaluation and decisions. *Patient Educ Couns*. 2008;73(3):443–7. DOI: 10.1016/j.pec.2008.07.013

57. Cameron LD, Marteau TM, Brown PM, Klein WM, Sherman KA. Communication strategies for enhancing understanding of the behavioral implications of genetic and

biomarker tests for disease risk: the role of coherence. *J Behav Med.* 2012;35(3):286–98. DOI: 10.1007/s10865-011-9361-5

58. Chua HF, Yates JF, Shah P. Risk avoidance: graphs versus numbers. *Mem Cognit.* 2006;34(2):399–410.

59. 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

60. Cox D, Sturm L, Cox AD. Effectiveness of asking anticipated regret in increasing HPV vaccination intention in mothers. *Health Psychol.* 2014;33(9):1074–83. DOI: 10.1037/hea0000071

61. Schwartz PH, Imperiale TF, Perkins SM, Schmidt KK, Althouse S, Rawl SM. Impact of including quantitative information in a decision aid for colorectal cancer screening: a randomized controlled trial. *Patient Educ Couns.* 2019;102(4):726–34. DOI: 10.1016/j.pec.2018.11.010

62. Ubel PA, Jepson C, Baron J. The inclusion of patient testimonials in decision aids: effects on treatment choices. *Med Decis Making.* 2001;21(1):60–68. DOI: 10.1177/0272989X0102100108

63. Zikmund-Fisher BJ, Windschitl PD, Exe N, Ubel PA. "I'll do what they did": social norm information and cancer treatment decisions. *Patient Educ Couns.* 2011;85(2):225–9. DOI: 10.1016/j.pec.2011.01.031

64. Carling CLL, Kristoffersen DT, Oxman AD, et al. The effect of how outcomes are framed on decisions about whether to take antihypertensive medication: a randomized trial. *PLoS One.* 2010;5(3):e9469. DOI: 10.1371/journal.pone.0009469

65. Wilson DK, Kaplan RM, Schneiderman LJ. Framing of decisions and selections of alternatives in health care. *Soc Behav.* 1987;2(1):51–59.

66. Parrott R, Silk K, Dorgan K, Condit C, Harris T. Risk comprehension and judgments of statistical evidentiary appeals: when a picture is not worth a thousand words. *Hum Commun Res.* 2005;31(3):423–52. DOI: 10.1093/hcr/31.3.423

67. Housten 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.

68. Nagle C, Hodges R, Wolfe R, Wallace E. Reporting Down syndrome screening results: women's understanding of risk. *Prenat Diagn.* 2009;29(3):234–9. DOI: 10.1002/pd.2210

69. Miron-Shatz T, Hanoch Y, Graef D, Sagi M. Presentation format affects comprehension and risk assessment: the case of prenatal screening. *J Health Commun.* 2009;14(5):439–50. DOI: 10.1080/10810730903032986

70. Hill S, Spink J, Cadilhac D, et al. Absolute risk representation in cardiovascular disease prevention: comprehension and preferences of health care consumers and general practitioners involved in a focus group study. *BMC Public Health.* 2010;10:108. DOI: 10.1186/1471-2458-10-108

71. Selinger CP, Kinjo Y, Jones DB, et al. Conveying medication benefits to ulcerative colitis patients and effects on patient attitudes regarding thresholds for adherence. *J Crohns Colitis.* 2013;7(8):e312–7. DOI: 10.1016/j.crohns.2012.11.006

72. Zikmund-Fisher B, Fagerlin A, Ubel P. Improving understanding of adjuvant therapy options by using simpler risk graphics. *Cancer.* 2008;113(12):3382–90. DOI: 10.1002/cncr.23959

73. 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

74. van Weert JCM, Alblas MC, van Dijk L, Jansen J. Preference for and understanding of graphs presenting health risk information. The role of age, health literacy, numeracy and graph literacy. *Patient Educ Couns.* 2021;104(1):109–117. DOI: 10.1016/j.pec.2020.06.031

75. McCaffery KJ, Dixon A, Hayen A, Jansen J, Smith S, Simpson JM. The influence of graphic display format on the interpretations of quantitative risk information among adults with lower education and literacy: a randomized experimental study. *Med Decis Making.* 2012;32(4):532–44. DOI: 10.1177/0272989X11424926

76. Carling CLL, Kristoffersen DT, Flottorp S, et al. The effect of alternative graphical displays used to present the benefits of antibiotics for sore throat on decisions about whether to seek treatment: a randomized trial. *PLoS Med.* 2009;6(8):e1000140. DOI: 10.1371/journal.pmed.1000140

77. Dolan JG, Qian F, Veazie PJ. How well do commonly used data presentation formats support comparative effectiveness evaluations? *Med Decis Making.* 2012;32(6):840–50. DOI: 10.1177/0272989X12445284

78. Fortin JM, Hirota LK, Bond BE, O'Connor AM, Col NF. Identifying patient preferences for communicating risk estimates: a descriptive pilot study. *BMC Med Inform Decis Mak.* 2001;1:2.

79. Okan Y, Garcia-Retamero R, Cokely ET, Maldonado A. Improving risk understanding across ability levels: encouraging active processing with dynamic icon arrays. *J Exp Psychol Appl.* 2015;21(2):178–94. DOI: 10.1037/xap0000045

80. Adarkwah CC, Jegan N, Heinzel-Gutenbrunner M, et al. Time-to-event versus ten-year-absolute-risk in cardiovascular risk prevention - does it make a difference? Results from the Optimizing-Risk-Communication (OptRisk) randomized-controlled trial. *BMC Med Inform Decis Mak.* 2016;16(1):152. DOI: 10.1186/s12911-016-0393-1

81. Stone ER, Bruine de Bruin W, Wilkins AM, Boker EM, MacDonald Gibson J. Designing graphs to communicate risks: understanding how the choice of graphical format influences decision making. *Risk Anal.* 2017;37(4):612–28. DOI: 10.1111/risa.12660

82. 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

83. Dolan JG, Iadarola S, Dolan JG, Iadarola S. Risk communication formats for low probability events: an exploratory study of patient preferences. *BMC Med Inform Decis Mak*. 2008;8(1):14. DOI: 10.1186/1472-6947-8-14

84. Feldman-Stewart D, Kocovski N, McConnell BA, Brundage MD, Mackillop WJ. Perception of quantitative information for treatment decisions. *Med Decis Making*. 2000;20(2):228–38.

85. Emmons KM, Wong M, Puleo E, Weinstein N, Fletcher R, Colditz G. Tailored computer-based cancer risk communication: correcting colorectal cancer risk perception.

J Health Commun. 2004;9(2):127–41. DOI: 10.1080/10810730490425295

86. 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

87. Wolfe CR, Reyna VF, Smith RJ. On judgments of approximately equal. *J Behav Decis Mak*. 2018;31(1):151–63. DOI: 10.1002/bdm.2061

88. Zikmund-Fisher BJ, Fagerlin A, Keeton K, Ubel PA. Does labeling prenatal screening test results as negative or positive affect a woman's responses? *Am J Obstet Gynecol*. 2007;197(5):528.e1–6.