

# How Point (Single-Probability) Tasks Are Affected by Probability Format, Part 1: A Making Numbers Meaningful Systematic Review

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

## Abstract

**Background.** To create guidance on the effect of data presentation format on communication of health numbers, the Making Numbers Meaningful project undertook a systematic review. **Purpose.** This article (one of a series) covers research studying so-called “point tasks,” in which a reader examines stimuli to obtain information about single probabilities. The current article presents the evidence on the effects of data presentation format on multiple outcomes: identification and recall, contrast, categorization, and computation. **Data Sources.** MEDLINE, Embase, CINAHL, the Cochrane Library, PsycINFO, ERIC, ACM Digital Library; hand search of 4 journals. **Finding Selection.** Manual pairwise screening to identify experimental and quasi-experimental research comparing 2 or more formats for quantitative health information for patients or other lay audiences. This article reports on 218 findings from 99 articles on single probability communication. **Data Extraction.** Pairwise extraction of data on stimulus (data in a data presentation format), task, and perceptual/affective/cognitive/behavioral outcomes. **Data Synthesis.** Most evidence on these outcomes was weak or insufficient. There was moderate to strong evidence that 1) recall was better with icon arrays with human figures than icon arrays with blocks, 2) survival curves make it easier to identify points of highest survival than mortality curves (contrast outcome), 3) adding an average population probability to a message about an individual probability may not affect recall, 4) computation performance is better with bar charts combined with data labels than with either numbers or graphics alone, 5) computation performance with rates is better when denominators match, and 6) framing strongly affects risky choices (contrast). **Limitations.** Heterogeneous study designs reduced the ability to develop strong evidence. **Conclusions.** Few findings assessing identification or recall, contrast, categorization, or computation outcomes for point tasks were comparable enough to each other to generate strong evidence.

## Highlights

- Many researchers have studied the effects of data presentation formats of single probabilities on different outcomes.
- However, few findings are comparable enough to allow for strong evidence-based conclusions about the impact on identification, recall, contrast, categorization, and computation outcomes.

## Keywords

numeracy, health literacy, health communication, risk communication, risk perception, data graphics

## 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).

Date received: December 17, 2022; accepted: December 19, 2023



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

Numbers, graphics, and verbal descriptions of probabilities convey critical information for informed decision making in health. As described elsewhere, we undertook a large systematic literature review (Prospero registration No. CRD42018086270) to develop evidence-based guidance on how to communicate numbers in medical and health domains across different data types.<sup>1,2</sup> We organized this literature according to a conceptual model of communication in which a reader performs a cognitive task upon some stimulus that contains data in some data presentation format, prompting a cognitive, affective, perceptual, or behavioral response that is measured with an outcome measure. We restricted our review to stimuli containing quantitative health-related data in 1 or more data presentation formats, including numbers, graphics, and verbal descriptions. We define and enumerate types of tasks, outcomes, data, and data presentation formats in our companion methodology and taxonomy articles.<sup>1,2</sup>

Although the literature review included several types of numbers, the current article focuses only on literature about probability numbers (Table A). We grouped the research literature according to the task performed by the reader or research participant as they examine the stimulus. Point tasks (the focus of the current article) are tasks in which people examine the stimulus for information about single probabilities, such as the chance of disease or the likelihood of side effects. Future articles (Table A) will cover difference tasks, those in which readers seek information specifically about the differences between probabilities (e.g., the effect of a therapy upon probability of recurrence). Other future articles will cover synthesis tasks, in which the reader integrates several probabilities, such as the set of risks of a medication or a list of risks

and benefits. As shown in Table A, a subset of synthesis task research involves interpreting probability information to estimate Bayesian posterior probabilities. Finally, we will present evidence on time-trend tasks, in which readers examine stimuli to evaluate patterns over time.

This article presents the subset of point task evidence pertaining to 5 commonly assessed outcomes: 1) identifying a number in the stimulus (termed “identification”); 2) remembering a number (“recall”); 3) selecting the largest or smallest of a set of probabilities (“contrast”); 4) recognizing which category, such as “elevated” or “within normal range,” a probability falls into, when category boundaries are provided (“categorization”); and 5) performing a computation such as converting a percentage into a number out of 100 or calculating the complement of a probability (“computation”). However, as described in more detail below, the articles in this review frequently did not report enough detail to allow us to distinguish between identification and recall, and thus we group these outcomes in the current article. We include evidence on the effects of all probability presentation formats—numerical, graphical, and verbal—on these outcomes. (Our companion point task results article, titled “Part 2,” presents the evidence on 7 additional important outcomes: probability perceptions, probability feelings, behavioral intentions, behavior, trust in information, preference for a data presentation format, and discrimination.)

## Methods

As described in more detail in our companion methodology article,<sup>2</sup> we searched for experimental (randomized) and quasi-experimental (nonrandomized) research comparing 2 or more formats for presenting quantitative health-related data to patients or other lay audiences. The search was performed on MEDLINE, Embase, CINAHL, the Cochrane Library, PsycINFO, ERIC, and ACM Digital Library, and we conducted hand searches of the tables of contents of *Medical Decision Making*, *Patient Education and Counseling*, *Risk Analysis*, and *Journal of Health Communication*.

The literature review identified 316 eligible articles on communicating probabilities. Of these, 99 articles involved point tasks involving the 5 outcomes reported here. These 99 articles produced 215 distinct findings. Methods for the literature search, screening, risk-of-bias evaluation, data extraction, credibility evaluation of findings, and organization into evidence tables are reported in detail in the methods article. In brief, a broad search was performed to find research comparing 2 or more ways of presenting quantitative health-related data

---

Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA (JSA); Columbia University School of Nursing (NCB), New York, NY, USA; Department of Population Health Sciences, Weill Cornell Medical College, New York, NY, USA (MMS); Department of Population Health, New York University Langone Health, New York, NY, USA (SBJ); Samuel J. Wood Medical Library, Weill Cornell Medical College, New York, NY, USA (MD, DD); Department of 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). 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.

**Table A** Current Article's Scope within the Making Numbers Meaningful Systematic Review<sup>a</sup>

Outcome	Section Number <sup>b</sup>	Probability Task				
		Point	Difference	Trend	Synthesis	Synthesis Bayesian
Identification or recall	1	This article				
Contrast	2	This article				
Categorization	3	This article				
Computation	4	This article				
Probability perceptions or feelings	5					
Effectiveness perceptions or feelings	6					
Behavior or behavioral intention	7					
Trust	8					
Preference	9					
Discrimination	10					

<sup>a</sup>Gray cells represent combinations that are not possible according to the definitions presented in Ancker et al.<sup>1</sup>

<sup>b</sup>This standardized numbering system has been used for results subheadings in this article and across all Making Numbers Meaningful results articles to ensure that readers can find comparable information in all articles.

to patients or other lay audiences. Within each included study, we extracted information about task, stimulus (data and data presentation format), and outcome. Overall, we selected outcomes to track informed by behavioral and risk communication theory (behavior or behavioral intention, probability perceptions or feelings, recall) or empirically on the basis of what was frequently measured by the research included in our review (trust, preference for a format), particularly measures used to measure comprehension (identification, contrast, computation, categorization, discrimination).

Data presentation formats included a) numbers, b) graphics, and c) verbal descriptions of probabilities. a) Numeric formats used for single probabilities (covered in the current article) included percentages, frequencies, proportions, and individual numbers (such as numbers of individuals affected). Among frequencies, we distinguished between 2 types of rates: those formatted as 1 in X (examples include “1 in 5” and “1 of every 25”) and those formatted as a rate per 10<sup>n</sup> (such as “12 in 100” or “2.5 per 1,000”). Unlike some other authors, we reserved the term “natural frequencies” for presentations of a series of probabilities and joint probabilities computed from the same pool of patients in the context of Bayes’ theorem.<sup>3,4</sup> (This definition is congruent with the original formulation of the term,<sup>4</sup> and using the term only for this purpose helps disambiguate otherwise contradictory evidence about the impacts of types of frequencies on comprehension outcomes.<sup>3</sup>) b) Graphical formats for single probabilities included icon arrays, number lines and risk ladders, bar charts, pie charts, and flow charts as well as novel creations such as animated icon arrays and slide shows. c) Verbal formats for single probability

information included probability terms such as “rare” or “likely.” In addition, we collected information on common manipulations of or variations on the formats (gain-loss framing, addition of or variation of contextual information including order effects, addition of anecdotes, and manipulation of the denominators of frequencies).

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. A single study could produce many findings, defined as combinations of task, format comparison, and outcome. For example, a single study could produce one finding on the effects of graphics versus numbers on perceived probability when the reader assessed a single probability and another finding on the effects of graphics versus numbers on recall of that probability. Each finding was rated for credibility by 2 expert reviewers (primarily J.S.A. and B.J.Z.-F., with N.C.B. substituting in cases of conflict of interest), who weighed sample size, statistical methods, face validity and comparability of the stimuli being compared, and the face validity or criterion validity of the outcome measures and covariates at the finding level, together with the S-ROB for the study from which the finding came. Credibility was assigned holistically on a scale from 1 to 10 on the basis of the expert team’s evaluation of these factors, rather than according to a quantitative rubric. The credibility of different findings from the same study often varied. For example, a study might produce a high-credibility finding for its primary outcome but a lower-credibility finding for a secondary outcome not subjected to hypothesis testing.

Findings were grouped by task and outcome and synthesized into guidance statements. We then applied a standard rubric to grade the strength of evidence for each guidance statement according to finding risk of bias, finding credibility, and consistency of findings.

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

Consistency was considered high if all findings were significant in the same direction or if a large majority were significant in one direction with a few lacking in significance, moderate if findings showed a small majority of significant effects in one direction with the remainder lacking significance, and low if the findings showed significant effects in different directions. Findings with high credibility (7 or higher on a scale of 1 to 10) and moderate credibility (4.5–6.5) are discussed below. Findings with lower credibility (4 or lower) are mentioned below, counted in Table B, and listed in our Findings tables, but they do not contribute to the evidence summaries or the statements in the evidence tables.

The freely available Making Numbers Meaningful project at OSF (<https://osf.io/rvxf2/>) contains methodology files (search strategy, S-ROB instrument, and data extraction instrument), a mapping file listing each included study and which Making Numbers Meaningful results article will cover it, and a Probability Findings folder containing each finding listed in this article.

## Results

Each results section 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 elements added for context, effect of gain-loss framing, effect of representations of

uncertainty, effect of manipulations of denominators, effect of animation or interactivity, and manipulations of time period. Table B shows the section headings and numbers of findings in each.

Within each subsection, evidence is arranged from strongest to weakest. Each subsection concludes with a table of the evidence-based guidance, arranged in the same order.

The full spreadsheet of point task findings cited in the current article is available at the free Making Numbers Meaningful Project at OSF (<https://osf.io/rvxf2/>) in the Probability Findings folder.

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

Many researchers administered questions about specific numbers appearing in the stimulus. If the stimulus was present, we classified the outcome as “identification,” but if the stimulus had been removed, the outcome was considered “recall.” As we have previously reported,<sup>2</sup> some articles were unclear about whether participants could refer to the stimulus when asked to “report” information or answer “knowledge questions.” We therefore combined the identification and recall outcomes in this article. When individual findings did clarify the outcome, we reflect that in the summaries below.

*Comparisons between numerical formats in effects on identification/recall of probabilities (subsection 1A).* **ADDITIONAL DIGITS:** A high-credibility finding<sup>5</sup> from a large study by Witteman et al. showed that adding more digits to the right of a decimal reduced recall of the percentage probability of cancer; recall was highest with integers and decreased with each additional digit to the right of the decimal up to 3 digits.

**HEART AGE:** A few studies used risk calculators that express the chance of cardiovascular disease as estimated “heart age” in years rather than as percentage chance of an event, such that an individual with a high risk of cardiovascular disease would show a “heart age” older than their chronological age. A high-credibility finding by Bonner et al.<sup>6</sup> showed better 2-wk recall of “heart age” plus the difference in years between heart age and actual age than of paired percentages plus arithmetic difference between them.

**PERCENTAGES, 1 IN X, RATE PER 10<sup>n</sup>, PROBABILITY BANDS:** In a high-credibility finding, Woloshin and Schwartz<sup>7</sup> demonstrated higher ability to identify percentages in a drug facts box table than a rate

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

Subsection		Section ( <i>n</i> Findings)				Total Findings per Data Presentation Format Comparison
		Identification or Recall	Contrast	Categorization	Computation	
Data Presentation Format Comparison	Section Number/ Subsection Letter <sup>a</sup>	1	2	3	4	
Comparisons between numerical formats	A	1A ( <i>n</i> = 16)	2A ( <i>n</i> = 12)	3A ( <i>n</i> = 1)	4A ( <i>n</i> = 12)	41
Comparisons between graphical formats	B	1B ( <i>n</i> = 23)	2B ( <i>n</i> = 19)	3B ( <i>n</i> = 4)	4B ( <i>n</i> = 11)	57
Comparisons between numerical and graphical formats	C	1C ( <i>n</i> = 23)	2C ( <i>n</i> = 14)	3C ( <i>n</i> = 3)	4C ( <i>n</i> = 14)	54
Comparisons between numerical and verbal probabilities	D	1D ( <i>n</i> = 3)	2D ( <i>n</i> = 1)	3D ( <i>n</i> = 1)	4D ( <i>n</i> = 1)	6
Comparisons of elements added for context	E	1E ( <i>n</i> = 8)	2E ( <i>n</i> = 2)	3E ( <i>n</i> = 8)	4E ( <i>n</i> = 2)	20
Comparisons of frames (gain, loss, and combination)	F	1F ( <i>n</i> = 5)	2F ( <i>n</i> = 7)	3F ( <i>n</i> = 0)	4F ( <i>n</i> = 1)	13
Comparisons of methods for representing uncertainty	G	1G ( <i>n</i> = 3)	2G ( <i>n</i> = 0)	3G ( <i>n</i> = 1)	4G ( <i>n</i> = 1)	5
Comparisons of larger or smaller denominators	H	1H ( <i>n</i> = 2)	2H ( <i>n</i> = 1)	3H ( <i>n</i> = 0)	4H ( <i>n</i> = 5)	8
Comparisons of animation or interactivity	I	1I ( <i>n</i> = 7)	2I ( <i>n</i> = 3)	3I ( <i>n</i> = 1)	4I ( <i>n</i> = 2)	13
Comparisons of shorter v. longer time periods	J	1J ( <i>n</i> = 1)	2J ( <i>n</i> = 0)	3J ( <i>n</i> = 0)	4J ( <i>n</i> = 0)	1
Total findings per outcome		91	59	19	49	218

<sup>a</sup>This standardized numbering system, reflected here and in Table A, has been used for results subheadings across all Making Numbers Meaningful (MNM) results articles. The standard numbers ensure that, for example, studies comparing graphical formats for their effects on categorization are always placed in a subhead labeled subsection 3B (whether or not there is a subsection 3A in that particular article). Our goal is to ensure that readers can use this subhead system to more easily locate similar sections across articles. The full list of section headers is available in the Methodology Files folder at the MNN project at <https://osf.io/rvxf2/>.

per 10<sup>n</sup> (e.g., number per 1,000) or combinations of percentages and rate per 10<sup>n</sup>. However, Henneman et al.,<sup>8</sup> in a moderate-credibility finding, demonstrated no difference in recall between different formats (percentages v. rates). Similarly, in a moderate-credibility finding, Garcia-Retamero and Galesic<sup>9</sup> demonstrated no difference in recall between pairs of rates per 10<sup>n</sup> and pairs of 1-in-X. Two moderate-credibility findings had opposite findings: Ruiz et al.<sup>10</sup> demonstrated that verbatim recall was slightly better with a pair of rates per 10<sup>n</sup> than with a pair of percentages, but Sinayev et al.<sup>11</sup> demonstrated the reverse pattern of improved identification or recall with percentages than with rate per 10<sup>n</sup>. Furthermore, a third

moderate-credibility finding<sup>12</sup> demonstrated that the ability to identify information was better with rates than with a 1-sided probability band (e.g., “up to a 1 in 10 chance”).

**TABLE VERSUS TEXT:** In a moderate-credibility finding, Tait et al.<sup>13</sup> demonstrated better ability to answer questions about numbers of people and differences in rates with rate per 10<sup>n</sup> in a table format rather than the same numbers in text, but another finding by the same author group<sup>14</sup> demonstrated no difference in a similar comparison. In a moderate-credibility finding, Brick et al.<sup>15</sup> found ability to identify specific risks or benefits was higher with a drug fact box table format than with percentages in text.

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

Comparison	Evidence Strength ( <i>n</i> Findings)	Applied Example of Evidence- Based Communication	General Guidance
Additional digits	Weak ( <i>n</i> = 1)	People may find 10% easier to remember than 9.966%.	Percentages presented with fewer decimal points may be easier to remember.
Heart age	Weak ( <i>n</i> = 1)	People may find a “heart age of 45” easier to remember than a “5% chance of heart attack.”	Chance of cardiovascular disease presented as effective age or “heart age” may be easier to remember than a percentage chance of disease.
Tables v. text	Weak ( <i>n</i> = 3)	People may find it easier to identify that the risk of headaches with drug A is 3% when probabilities are shown in a table format instead of in text.	Table formats may make it easier for people to identify particular probabilities versus when numbers are embedded in text.
Percentage, 1 in X, rate per 10 <sup>n</sup> , probability bands	Insufficient— inconsistent findings ( <i>n</i> = 6)	It is not clear whether different numerical formats for probability (e.g., phrasing a probability as a chance of 10%, 0.10, 10 in 100, or “up to 1 in 10”) makes a difference to how easy or hard it is to identify or remember numbers.	
Table v. text and table design	Insufficient—too few findings ( <i>n</i> = 2)	It is not clear whether stating the absolute difference between 2 probabilities in a table affects identification or recall of the probabilities themselves.	
Verbalized numbers	Insufficient— inconsistent findings ( <i>n</i> = 2)	It is not clear whether verbalizing probability statements (e.g., “one out of every 25”) makes probabilities easier or harder to remember.	

**TABLE DESIGN:** In a moderate-credibility finding, Mühlbauer et al.<sup>16</sup> demonstrated that ability to identify numbers in a drug fact box was lower when “how to read this table” instructions were added or the numbers were provided in sentences. However, a similar high-credibility finding<sup>17</sup> demonstrated no difference in identification/recall when arithmetic difference was added to numbers, although the presence of verbal labels in addition to the absolute difference complicates interpretation of this negative finding.

**VERBALIZED NUMBERS:** A moderate-credibility finding<sup>18</sup> suggests that recall was lower when the rate per 10<sup>n</sup> was verbalized as “N out of every 1,000 people,” with no other recall differences across formats. However, a high-credibility finding<sup>19</sup> demonstrated higher recall when 1-in-X pairs were phrased as “one out of every X” than as 1-in-X icon arrays or 1:X numbers.

An additional relevant finding was not summarized due to small sample size and ceiling effects.<sup>20</sup> A recall finding was not summarized here because floor effects (i.e., low overall recall) reduced the credibility of a finding of no difference between formats.<sup>15</sup>

*Comparisons between graphical formats in effects on identification/recall of probabilities (subsection 1B).* Although a number of findings have compared graphics, our ability to draw conclusions is limited by the fact that there are few findings for each specific type of graphic.

**ICON SHAPE:** Two consistent high-credibility findings from Kreuzmair et al.,<sup>21</sup> and Zikmund-Fisher et al.<sup>22</sup> suggest that with icon arrays, recall is higher with stick figure icons than with abstract icons.

**PART-WHOLE GRAPHICS:** A high-credibility finding from Okan et al.<sup>23</sup> showed higher recall with a part-to-whole icon array than with a numerator-only icon array.

**TYPES OR COMBINATION OF GRAPHICS:** In a high-credibility finding, van Weert et al.<sup>24</sup> demonstrated slightly better identification/recall of information for bar charts versus other graphical formats (“clock” pie charts, icon arrays, number lines, unlabeled pie charts). However, another high-credibility finding from Zikmund-Fisher et al.<sup>25</sup> showed better identification/recall with icon arrays than horizontal bar charts.

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Icon shape	Strong ( $n = 2$ )	An icon array of human figures is easier to remember than an icon array of blocks.	Recall is better with an icon array of anthropomorphic shapes than with an icon array of abstract shapes.
Part-whole graphics	Weak ( $n = 1$ )	A chance shown as an icon array with 10 affected people and 90 unaffected ones may be easier to remember than one shown as icons of 10 affected people alone, with no unaffected ones.	Recall may be better with part-to-whole graphics that depict both numerator and denominator of a probability than with numerator-only graphics that depict only the numerator.
Types or combinations of graphics	Insufficient (inconsistent findings; $n = 8$ )	It is not clear whether any specific graphic (bar chart, icon array, etc.) is better than others for improving identification/recall of information.	
Icon array with fewer v. more outcomes	Insufficient (inconsistent findings; $n = 3$ )	It is not clear whether identification/recall of probabilities in an icon array is better when the graphic illustrates fewer outcomes (e.g., survival and mortality) or more outcomes (e.g., cancer mortality, other mortality, and survival).	
Icon array with grouped v. random arrangement	Insufficient (too few findings; $n = 1$ )	It is not clear whether identification/recall of the probability shown in icon arrays is affected by whether icons are grouped or randomly distributed.	
Number line with social comparisons	Insufficient (too few findings; $n = 1$ )	It is not clear whether supplementing a risk ladder of absolute probabilities with social comparison labels (e.g., “higher than average”) affects recall of the absolute probability.	

Similarly, another high-credibility-finding from Fraenkel et al.<sup>26</sup> demonstrated better identification/recall with an icon array (labeled with rate per  $10^n$ ) than with an interactive spinner graphic or an animated slide show. A moderate-credibility finding from Masson et al.<sup>27</sup> also showed somewhat better recall with icon arrays versus bar charts and more so versus number lines, but confidence in this finding is limited due to differences across the graphics. Furthermore, a moderate-credibility finding with a stimulus that showed probabilities changing over time by Hamstra et al.<sup>28</sup> demonstrated no difference in recall between multiple graphical formats (line graph, sets of pie charts, set of bar charts, sets of icon arrays). In moderate-credibility findings from Ghosh et al.,<sup>29</sup> adding an icon array to a bar chart did not improve recall, and one finding from Martin et al.<sup>30</sup> showed recall was similar between icon arrays and a speedometer graphic. Similarly, a moderate-credibility finding by Hawley et al.<sup>31</sup> demonstrated no major differences between several graphic types (bar charts, icon arrays, pie charts).

**ICON ARRAY WITH FEWER VERSUS MORE OUTCOMES:** A moderate-credibility finding from Zikmund-Fisher et al.<sup>32</sup> showed better identification/recall of information with icon arrays that visually

highlighted survival outcomes rather than arrays that highlighted both survival and mortality outcomes, but in a follow-up finding in the same publication, the authors did not find a significant difference in recall. Similarly, 2 moderate-credibility findings published in the same article by McDowell et al.<sup>33</sup> showed no differences in identification/recall between separate icon arrays and integrated multioutcome icon arrays.

**ICON ARRAY WITH GROUPED VERSUS RANDOM ARRANGEMENT:** In a moderate-credibility finding, Ancker et al.<sup>34</sup> showed estimates of the probability were more accurate when icons were grouped versus randomly distributed in an icon array.

**NUMBER LINE WITH SOCIAL COMPARISONS:** In a moderate-credibility finding by Emmons et al.,<sup>35</sup> recall of personal probability of colorectal cancer as an approximate rate per  $10^n$  was not significantly different when a risk ladder (static or interactive) with a scale of rates per  $10^n$  was supplemented by a second risk ladder that showed social comparison labels (e.g., “higher than average”).

Several findings are not summarized due to a lack of hypothesis testing or other limitations (specifically,<sup>36–40</sup> a finding from Feldman-Stewart et al. substudy 4,<sup>40</sup> a finding from Rakow et al. substudy 2<sup>41</sup>).

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Adding numerator-only icon arrays to numbers	Weak ( $n = 2$ )	A message containing a row of 9 icons and a rate of 9 in 100 may be harder to answer questions about or remember than simply saying the probability is 9 in 100.	Adding a foreground-only icon array to a rate per 10 <sup>n</sup> may reduce identification or recall of probabilities.
Adding part-to-whole icon arrays to numbers	Insufficient (inconsistent findings; $n = 5$ )	It is not clear whether adding a part-to-whole icon array (one that shows both numerator and denominator of a probability) to a number will affect identification or recall.	
Numbers v. graphics	Insufficient (inconsistent findings; $n = 11$ )	It is not clear whether numbers alone or graphics alone are generally better for identification or recall of probabilities.	

*Comparisons between numerical and graphical formats, and effect of combinations of numerical and graphical formats, on identification/recall of probabilities (subsection 1C).* **ADDING NUMERATOR-ONLY ICON ARRAYS TO NUMBERS:** Two moderate-credibility findings in the same article by Stone et al.<sup>42</sup> showed that adding numerator-only icon arrays to rates reduced the ability to answer questions about the probability.

**ADDING PART-TO-WHOLE ICON ARRAYS TO NUMBERS:** In a high-credibility finding, Fraenkel et al.<sup>26</sup> showed that the ability to answer questions about probabilities was improved when a part-to-whole icon array was added to a rate, and Reder and Thygesen<sup>43</sup> also demonstrated improvement in the ability to answer questions about probability information with the combination. However, a high-credibility finding by Fagerlin et al.<sup>44</sup> and a moderate-credibility finding by Henneman et al.<sup>8</sup> each suggested that adding part-to-whole icon arrays to rates made no difference. Finally, a moderate-credibility finding<sup>10</sup> demonstrated lower recall when part-to-whole icon arrays were added to rates than when rates or percentages were shown alone.

**NUMBERS VERSUS GRAPHICS:** Eleven findings had different comparisons and different findings. A high-credibility finding<sup>17</sup> demonstrated better ability to answer questions about probabilities for verbalized “1 out of every X” numbers than for 1-in-X icon arrays. Gibson et al.<sup>45</sup> (moderate) showed that ability to answer questions was higher for bar charts than for rates, and Martin et al.<sup>30</sup> (moderate) demonstrated better recall with icon arrays or a speedometer graphic than with percentage alone. Lipkus et al.<sup>46</sup> (moderate) found no differences

between pie charts and percentages in terms of ability to answer questions about probabilities, but Schonlau and Peters<sup>47</sup> (moderate) demonstrated worse ability to identify information with pie chart formats than with percentages in a table. In 2 findings published in 2 substudies in the same article, McDowell et al.<sup>33</sup> found neither identification nor recall differences between tables of rates and either integrated or separated icon arrays. Tait et al.<sup>14</sup> (high-credibility) found a better ability to answer questions about probabilities with part-to-whole icon arrays than with rates either in table or text format, but a moderate-credibility finding from a similarly designed study by the same group<sup>13</sup> demonstrated better performance with rate per 10<sup>n</sup> in tables than with icon arrays. Hawley et al.<sup>31</sup> found a higher ability to answer questions about number of people affected with rates in tables than with icon arrays, pie charts, or bar charts, but this format had lower performance for identifying differences between numbers. A high-credibility finding from a similarly designed study<sup>24</sup> by van Weert also demonstrated the highest ability to answer questions with a table of rate per 10<sup>n</sup> compared with bar charts (next best), icon arrays, pie charts, and number lines.

Several findings<sup>37,38,48–50</sup> are not summarized because of a lack of hypothesis testing or other limitations.

*Comparisons between numerical and verbal probabilities in effects on identification/recall of probabilities (subsection 1D).* As described in our methods article,<sup>2</sup> we included research on verbal descriptions only when they were verbal probabilities (such as “likely,” “rare,” and “common”). In this section, we summarize research that

**Table 1D** Evidence-Based Guidance for Contrasts between Numerical and Verbal Probability Formats on Identification/Recall of Probabilities

Comparison	Evidence Strength	General Guidance
Adding verbal probability to numbers	Insufficient (too few findings; $n = 1$ )	It is not clear whether adding verbal probabilities to numbers affects identification or recall of probabilities.
Verbal probability compared with numbers	Insufficient (too few findings; $n = 1$ )	It is not clear whether numbers alone or verbal probabilities alone are better for recall of information.

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Adding the population average	Moderate ( $n = 2$ )	In a message that “your chance of disease is 7%,” adding that the average probability is 5% may not affect how easy it is to answer questions about the probability.	Adding a population average to a message about personal chance of an event does not affect ability to answer questions about the probability.
Adding difference labels	Weak ( $n = 1$ )	In a message that drug A has a 5% chance of side effects and drug B has a 7% chance, stating that drug B has a higher chance may not affect ability to answer questions about the numbers.	Adding labels describing which option has higher rates of each outcome may not affect people’s ability to answer questions about probabilities.
Adding anecdotes	Insufficient (inconsistent findings; $n = 4$ )	It is not clear whether adding anecdotes about people who have experienced an event affects people’s ability to answer questions about the probability of the event.	
Adding lifetime chance	Insufficient (too few findings; $n = 1$ )	It is not clear whether changing the time interval of a probability (such as replacing a lifetime chance with a 10-y chance) affects ability to answer questions about the probability.	

**Table 1F** Evidence-Based Guidance for Effect of Gain-Loss Framing on Identification/Recall of Probabilities

Comparison	Evidence Strength	General Guidance
Survival v. mortality curves	Insufficient (inconsistent findings; $n = 2$ )	It is not clear whether ability to answer questions about a probability at one point in time is better with survival curves or mortality curves.
Survival focused v. survival + mortality icon arrays	Insufficient (inconsistent findings; $n = 2$ )	It is not clear whether ability to answer questions about probabilities is better with icon array graphics showing only positive outcomes (survival) or both positive (survival) and negative (mortality) outcomes.

directly contrasted verbal descriptions with numerical probabilities. (A previous article included a larger set of verbal probability articles, including both the ones that

contrasted verbal probability with another format and those that assessed the effect of verbal probability alone with no comparison stimulus.<sup>51</sup>)

**Table 1G** Evidence-Based Guidance for Representing Uncertainty on Identification/Recall of Probabilities

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Point estimates v. ranges	Weak ( $n = 3$ )	It may be somewhat easier to answer questions about a probability described as “10%” than one described as “8 to 12%.”	Ability to answer questions about probabilities may be better when the information is presented as point estimates without uncertainty than when shown as ranges or other displays showing uncertainty.

**Table 1H** Evidence-Based Guidance for Effect of Manipulating Denominators on Identification/Recall of Probabilities

Comparison	Evidence Strength	General Guidance
Same v. different denominators	Insufficient (too few findings; $n = 1$ )	It is not clear whether presenting several rates with the same denominator versus different denominators affects recall.
Denominator size	Insufficient (too few findings; $n = 1$ )	It is not clear whether presenting rates with larger or smaller denominators affects recall.

**ADDING VERBAL PROBABILITY TO NUMBERS:** One moderate-credibility finding (by Sinayev et al.<sup>11</sup>) demonstrated that the ability to answer questions about a probability number or recall of it was improved when a verbal probability was added to percentages or rates.

**VERBAL PROBABILITY COMPARED TO NUMBERS:** A moderate-credibility finding<sup>52</sup> demonstrated no difference in 4-mo recall by whether the chance had been presented with 1 in X or verbal probability.

One finding is not summarized due to small sample size and other limitations.<sup>20</sup>

*Comparisons of elements added for context on identification/recall of probabilities (subsection 1E).* We included research on 4 types of information frequently added to provide context to quantitative probability information.

**ADDING THE POPULATION AVERAGE:** Two moderate-credibility findings<sup>45,53</sup> examined the effect of adding information about the population average to a message about personal chance of an outcome, with neither finding an effect on abilities to answer questions about probabilities.

**ADDING DIFFERENCE LABELS:** A high-credibility finding<sup>19</sup> demonstrated no difference in ability to answer questions about differences between probabilities when a drug facts box-like table including absolute rates and/or absolute probability differences did or did

not include labels describing which drug had the higher frequency of each outcome.

**ADDING ANECDOTES:** Three moderate-credibility findings and 1 high-credibility finding have examined effects of anecdotes, that is, short narratives about people (such as people who experienced a side effect or a benefit from a therapy). These did not have consistent results. Betsch et al. published 2 substudies in 1 paper,<sup>54</sup> with one finding that a larger proportion of adverse event anecdotes improved recall of the adverse event probability and the other finding showing no such effect. Gibson et al.<sup>55</sup> demonstrated that when the proportion of adverse effect anecdotes was not consistent with the actual proportion of adverse effects, it impaired ability to answer questions about the probability, but Fagerlin et al.<sup>44</sup> demonstrated no effect of anecdotes (either statistically proportionate or nonproportionate) on ability to answer questions about probabilities.

**ADDING LIFETIME CHANCE:** One moderate-credibility finding<sup>8</sup> compared lifetime probability alone with lifetime probability plus 10-y probability, finding no effect on recall.

*Comparisons of frames (gain, loss, combination) on identification/recall of probabilities (subsection 1F).* A high-credibility finding<sup>56</sup> demonstrated greater ability to answer questions about individual time points with mortality curves than with survival curves. However, another

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Type of animation	Weak ( $n = 1$ )	Whether an icon array personalizes an icon or moves the icons may not make a difference to recall.	Recall of probabilities may not be affected by different types of animation in animated icon array graphics.
Animation v. static	Insufficient (inconsistent findings; $n = 2$ )	It is not clear whether static or animated graphics are better for recall.	
Interactivity	Insufficient (inconsistent findings; $n = 3$ )	It is not clear whether adding interactivity to graphics affects recall of probabilities.	

**Table 1J** Evidence-Based Guidance for Effect of Varying the Time Period on Identification/Recall of Probabilities

Comparison	Evidence strength	General Guidance
Time period	Insufficient (too few findings; $n = 2$ )	It is not clear whether adding a 10-y chance of disease to a lifetime chance affects probability recall.

moderate-credibility finding<sup>57</sup> demonstrated the reverse pattern.

In a moderate-credibility finding, Zikmund-Fisher et al.<sup>32</sup> demonstrated better recall or ability to identify information when outcomes of breast chemotherapy or hormonal therapy were presented in positively framed icon arrays showing only survival information (with other outcomes unlabeled) than in combination-framed icon arrays showing both survival and mortality statistics. However, a second moderate-credibility replication finding in the same article<sup>32</sup> did not find a significant difference in recall.

A final relevant finding<sup>58</sup> was not synthesized due to limitations.

*Comparisons of methods for representing uncertainty on identification/recall of probabilities (subsection 1G).* We focused on numeric uncertainty representations such as the range or the confidence interval.

In a high-credibility finding,<sup>59</sup> the ability to answer questions about the probabilities of benefit or harm was significantly better when people received a percentage point estimate than when people received a range and were asked to report the maximum value of the range. Similarly, in one finding,<sup>45</sup> ability to answer questions was better with a combined graphic that did not show uncertainty than with several that did show uncertainty, but a number of differences between the graphics makes it somewhat difficult to attribute the difference to the

uncertainty representation. A finding from a smaller study<sup>46</sup> demonstrated no effect of uncertainty (represented as pie charts) on recall.

*Comparisons of larger or smaller denominators on identification/recall of probabilities (subsection 1H).* A moderate-credibility finding<sup>42</sup> suggests that recall of several rates is improved when they all have the same denominator. In a moderate-to-lower-credibility finding, Garcia-Retamero and Galesic<sup>9</sup> demonstrated higher recall of both rates and 1-in-X numbers when denominators were small (100 or 1,000) rather than large (5,000 or 10,000). However, this effect was mostly due to lack of recall of the denominator, and the small sample size limits confidence in this finding.

*Comparisons of animation or interactivity on identification/recall of probabilities (subsection 1I).* Animation was defined as the use of movement in graphics or illustrations, including moving slide shows, cartoons, or graphics such as icon arrays. Interactivity was defined as any function that allowed the user to manipulate or input information.

**TYPE OF ANIMATION:** A high-credibility finding<sup>60</sup> demonstrated no differences in recall with several different types of animation and interactivity applied to icon arrays. Animation features included whether a personal avatar was displayed and whether the avatar moved;

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Rates per 10 <sup>n</sup> v. 1-in-X	Weak ( <i>n</i> = 4)	It may be somewhat easier to identify the larger or smaller of a set presented as “10 in 1,000 and 4 in 1,000” than as “1 in 100 and 1 in 250.”	People may be better able to identify the larger or smaller of a series of probabilities when they are presented as rates per hundred or thousand than when they are presented as 1 in X.
Consistent formats	Weak ( <i>n</i> = 1)	It may be somewhat easier to identify the larger or smaller of a set presented as “1%, 5%, and 10%” than as “1 in 100, 5%, and 0.10.”	People may be better able to identify the larger or smaller of a series of probabilities when all numbers are presented in the same format, e.g., all as the same type of rate per 10 <sup>n</sup> or all as percentages.
Risk-benefit order	Weak ( <i>n</i> = 1)	It may be somewhat easier to identify the larger or smaller harm in this message: “The chance of improvement from this medication is 70%, and the chance of a side effect is 3%” than in this one: “The chance of a side effect from this medication is 3%, and the chance of improvement is 70%.”	In risk-benefit communications, people may be better able to identify the larger or smaller of a series of harm probabilities when benefits are presented first, followed by harms.
Numbers in tables ( <i>n</i> = 1)	Insufficient (inconsistent findings; <i>n</i> = 4)	It is not clear whether presenting probabilities in a table format or embedded in text affects people’s ability to identify the larger or smaller of a set of probabilities.	
Numbers as a list or as a flow chart	Insufficient (too few findings; <i>n</i> = 1)	It is not clear whether presenting number-based flow charts (versus numbers) affects people’s ability to identify the larger or smaller of a set of probabilities.	
Ratio, case count, rate per 10 <sup>n</sup>	Insufficient (too few findings; <i>n</i> = 1)	It is not clear whether different numerical formats for presenting probability of overdiagnosis (a harm) affects ability to identify whether a screening test increases chance of diagnosis.	

interactivity included whether the user could personalize the avatar.

**ANIMATION VERSUS STATIC:** In a high-credibility finding, Fraenkel et al.<sup>26</sup> demonstrated recall was better with a rate per 10<sup>n</sup> combined with a static icon array than with an animated (but not interactive) slide show showing a long series of people affected and unaffected. In a moderate-credibility finding, Houston et al.<sup>61</sup> demonstrated no differences in “verbatim” or “gist” knowledge scores between static icon arrays and 2 types of animated icon arrays emphasizing either subgroups or randomness, but several factors limit confidence in this finding.

**INTERACTIVITY:** Three studies used very different types of interactive graphics. In a high-credibility finding, Fraenkel et al.<sup>26</sup> demonstrated recall was better when rates were combined with a static icon array than with an interactive “spinner” graphic that the participant could spin. It is not clear whether the difference was due to the

specific design of the spinner or influenced by the relative unfamiliarity of the spinner. However, a moderate-credibility finding<sup>50</sup> demonstrated no recall difference between information in a static bar chart and an interactive one in which the respondent was asked to adjust the height. Emmons (moderate credibility) found that recall of personal probability of colorectal cancer as an approximate rate per 10<sup>n</sup> was not significantly different between static risk ladders and interactive ones with the ability to toggle risk factors to see their effects on one’s personal probability of cancer.<sup>35</sup>

Another finding<sup>39</sup> from a study comparing animated and static number lines and icon arrays lacked hypothesis testing and is not summarized.

*Comparisons of shorter versus longer time periods on identification/recall of probabilities (subsection 1J).* In one moderate-credibility finding from a small study,<sup>8</sup> there

**Table 2B** Evidence-Based Guidance for Effects of Graphical Formats on Ability to Contrast Probabilities

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Grouped v. random icon arrays	Moderate ( $n = 4$ )	It is easier to see which probability is the highest with icon arrays that have grouped icons than with icon arrays of randomly arranged ones.	Ability to select highest or lowest probabilities or see that 2 probabilities are equal is higher with grouped icon arrays than random icon arrays.
Bar charts	Insufficient (too few findings; $n = 1$ )	It is not clear whether ability to select highest or lowest probabilities in a bar chart is better with a y-axis that starts at 0 or a compressed axis that started at $Y > 0$ .	
Multiple graphics	Insufficient (inconsistent findings; $n = 9$ )	It is not clear whether ability to select highest or lowest probabilities is affected by different graphical displays (bar charts, pie charts, grouped icon displays, random icon displays, integrated icon arrays).	

were no recall differences by whether lifetime chance was given alone or combined with the 10-y chance.

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

Asking participants to select the largest or smallest in a list of numbers or rank the numbers in order of size was classified as a *contrast* outcome.

Within the *contrast* outcome, there were no relevant findings on the effects of stating or illustrating numerical uncertainty (category 2G in Table B) or effects of varying the time period (category 2J).

*Comparisons between numerical formats in effects on ability to contrast probabilities (subsection 2A).* **RATES PER  $10^n$  VERSUS 1 IN X:** Two high-credibility findings<sup>62,63</sup> demonstrated that people's ability to identify which probability was larger was better with pair of rates with the same denominator than with pair of 1-in-X frequencies. However, in another high-credibility finding, Cuite et al.<sup>64</sup> demonstrated no difference in ability to identify which probability was higher between 1-in-X and rates per hundred or thousand, and that 1-in-X was better than percentages for this task. Similarly, in a moderate-to-high-credibility finding, Pighin et al.<sup>65</sup> demonstrated no difference in overall ability to compare probabilities by whether the probabilities were in 1-in-X or rate per  $10^n$  format.

**CONSISTENT FORMATS:** In a moderate-to-high-credibility finding, Nagle et al.<sup>66</sup> demonstrated that the

ability to compare probabilities and choose the higher/lower was best when both were in the same format.

**ORDER OF RISKS AND BENEFITS:** In a moderate-to-high-credibility finding, Ubel et al.<sup>67</sup> demonstrated people were better at identifying the larger or smaller chance when information was presented benefit-then-harm than when presented harm-then-benefit.

**TABLES:** In a moderate-credibility finding, Tait et al.<sup>13</sup> demonstrated ability to identify which group was more affected by chances or benefits or drug was better when the rate per  $10^n$  was embedded in a table versus in text. However, a high-credibility finding by the same author group<sup>14</sup> demonstrated no difference in this ability, and other formatting factors may have contributed to the identified effects. A moderate-credibility finding<sup>10</sup> demonstrated no difference in ability to identify the higher risk by format of the number (rate per thousand or percentage), but the small sample size reduces confidence in the negative finding. Brick et al.<sup>15</sup> showed better performance on a composite comprehension measure with a fact box than with numbers in a paragraph. However, because only 5 of 12 comprehension measure questions involved contrast outcomes, the contrast outcome finding is classified as only moderate credibility.

**NUMBERS AS LIST VERSUS NUMBERS AS FLOW CHART:** In a moderate-credibility finding, Dolan et al.<sup>68</sup> demonstrated no difference in ability to identify larger chances between a list of rates per thousand or the same numbers formatted as a number-based flow chart (or several graphical formats).

**RATIO VERSUS CASE COUNT VERSUS RATE PER  $10^n$ :** In a moderate-credibility finding, Waller et al.<sup>69</sup> demonstrated no difference in ability to answer

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Part-whole icon arrays added to numbers	Weak ( $n = 10$ )	It may be easier to identify the higher of 2 probabilities (e.g., 3 in 100 and 5 in 100) if the numbers are supplemented with icon arrays that show both numerator and denominator.	Adding part-whole icon array graphics to numbers may improve ability to identify highest or lowest probability.
Numerator-only icon arrays	Weak ( $n = 2$ )	It may not be any easier to identify the higher of 2 probabilities (e.g., 3 in 100 and 5 in 100) if the numbers are supplemented with a row of 3 icons (for the 3 in 100 chance) and a row of 5 icons (for the 5 in 100 chance).	Adding numerator-only icon arrays to rate per 10 <sup>n</sup> may not alter ability to identify the highest probability.

whether screening mammograms increase the likelihood of diagnosis when information about overdiagnosis was presented in 3 numerical formats (an odds ratio, a case count, or a rate per 10<sup>n</sup>), but concerns about question wording and stimulus design somewhat limit confidence in this negative finding.

*Comparisons between graphical formats in effects on ability to contrast probabilities (subsection 2B).* **GROUPED VERSUS RANDOM ICON ARRAYS:** A high-credibility finding<sup>70</sup> demonstrated better ability to identify biggest/smallest/equal probabilities with grouped icon arrays (static or animated) versus random icon arrays (static or animated). Similarly, in 2 moderate-credibility findings, Feldman-Stewart et al.<sup>40</sup> substudy 3 and Feldman-Stewart et al.<sup>71</sup> showed that ability to identify biggest/smallest was better with grouped icon arrays versus random icon arrays. In the same finding, Feldman-Stewart et al.<sup>71</sup> also showed that adding axes to graphics improved performance for grouped icon arrays but worsened performance for random icon arrays. However, another moderate-credibility finding<sup>72</sup> demonstrated that ability to identify the largest/smallest probabilities (measured using an outcome that combined single risks and effects) was not associated with whether an icon array was grouped or random.

**BAR CHARTS:** In a moderate-credibility finding, Okan et al.<sup>73</sup> (substudy 16) showed that ability to identify largest/smallest in a bar chart was facilitated by a standard y-axis starting at zero and impaired by a y-axis that started at 400.

**MULTIPLE GRAPHICS:** A variety of findings have compared ability to identify larger or smaller probabilities across graphics types, with varying results. Aside

from the evidence suggesting that grouped icon arrays are superior to randomly arranged ones (discussed above), there is insufficient similarity and coherence among the available studies to determine whether icon arrays, bar charts, number lines, or other types of graphics are better for improving people's ability to contrast quantities.

In 2 moderate-credibility findings, Feldman-Stewart et al.<sup>40</sup> and Feldman-Stewart et al.<sup>71</sup> showed that ability to identify biggest/smallest was best with vertical bar charts and grouped icon arrays and worst with pie charts and random icon arrays. In this latter study, Feldman-Stewart et al.<sup>71</sup> also showed that adding axes to graphics improved performance for grouped icon arrays and both horizontal and vertical bar charts but worsened performance for random icon arrays and pie charts. In a somewhat similar finding, Downen et al.<sup>74</sup> (moderate-credibility) demonstrated that ability to identify the best/worst option was better with vertical or horizontal bar charts (grouped to show proportions dead/alive at several time points) than with pairs of pie charts or survival curves. In moderate-credibility findings from 2 substudies published in the same article, McDowell et al.<sup>33</sup> demonstrated improved ability to identify options with higher probabilities with sets of individual icon arrays versus an integrated icon array. However, a variety of findings have demonstrated no differences between graphical formats in identifying larger or smaller probabilities: neither a moderate-to-high-credibility finding<sup>31</sup> nor a very similar moderate-credibility finding<sup>24</sup> demonstrated any differences in this ability between rate per 10<sup>n</sup> in tables, bar charts (horizontal or vertical), pie charts, or icon arrays (vertical or horizontal). Dolan et al.<sup>68</sup> also demonstrated no differences between a

horizontal number line with log-scale axis, a vertical bar chart, and a  $25 \times 20$  icon array. Similarly, a moderate-credibility finding<sup>75</sup> demonstrated no difference between icon arrays and sequential experience format in people's ability to compare cancer detections versus false alarms.

Two lower-credibility findings (Timmermans et al.,<sup>76</sup> Okan et al.<sup>73</sup> substudy 7) were not synthesized due to small sample size. Two low-credibility findings published in the same article<sup>41</sup> were also not synthesized due to confounding in the manipulation, while another low-credibility finding<sup>37</sup> had both a small sample size and an aggregate measure that combined multiple types of outcomes, decreasing confidence in a contrast effect.

*Comparisons between numerical and graphical formats, and combinations of numerical and graphical format, in effects on ability to contrast probabilities (subsection 2C).* **NUMERATOR-ONLY ICON ARRAYS ADDED TO NUMBERS:** In 2 moderate-credibility findings from substudies published in the same article, Stone et al.<sup>42</sup> showed that adding a foreground-only icon array graphic to rate per  $10^n$  did not affect people's ability to select the most common disease.

**PART-TO-WHOLE ICON ARRAYS:** In a high-credibility finding, Tait et al.<sup>14</sup> demonstrated ability to identify more affected groups was better with icon arrays than with rate per  $10^n$  in table or text format, while a moderate-credibility finding of similar design<sup>13</sup> demonstrated icon arrays or the rate per  $10^n$  in table format led to improvements over rate per  $10^n$  in text. Similarly, a moderate-to-high-credibility finding<sup>75</sup> demonstrated that adding icon arrays (or an animated graphic, discussed elsewhere) to rates improved ability to identify more common events. In moderate-credibility findings from substudies published in the same article, McDowell et al.<sup>33</sup> demonstrated improved ability to identify options with higher probabilities with sets of individual icon arrays versus an integrated icon array or rate per  $10^n$  in a tabular facts-box format. However, both a moderate-to-high-credibility finding<sup>31</sup> and a moderate-credibility finding<sup>24</sup> demonstrated no differences in this ability between rate per  $10^n$  in tables, bar charts (horizontal or vertical), pie charts, or icon arrays (vertical or horizontal). Similarly, in a moderate-credibility finding, Dolan et al.<sup>68</sup> demonstrated no differences in ability to identify higher disease probabilities between a list of rate per  $10^n$ , a number-based flow chart, a number line, a vertical bar chart, or a  $20 \times 25$  icon array. A low-to-moderate-credibility finding<sup>37</sup> also demonstrated no differences between percentages combined with rate per  $10^n$ , vertical stacked bar charts, pairs of icon arrays, or sets of pie

charts. A moderate-credibility finding<sup>10</sup> demonstrated no difference in ability to identify the higher probability by format of the number (icon arrays with rate per  $10^n$  versus the rates alone or percentages alone), but the small sample size reduces the confidence in the negative finding.

Two lower-credibility findings<sup>47,76</sup> were not synthesized due to limitations for this outcome.

*Comparisons between numerical and verbal probabilities in effects to contrast probabilities (subsection 2D).* One lower-credibility finding had a small sample size and confounded manipulations.<sup>77</sup>

*Comparisons of elements added for context on ability to contrast probabilities (subsection 2E).* **CHANCE OF COMPARISON EVENTS:** In a moderate-to-high-credibility finding, Ubel et al.<sup>67</sup> demonstrated that adding information about chance of other cancers eliminated order effects, so ability to identify the larger/smaller harm was not affected by whether information was presented benefit-then-harm or harm-then-benefit.

**INTERPRETIVE LABELS:** One moderate-credibility finding<sup>78</sup> suggests that classifying items into verbally described categories helps readers more than providing numbers does. However, this finding was limited by the fact that the numbers provided were complex, reducing confidence in this finding.

*Comparisons of frames (gain, loss, or combination) on ability to contrast probabilities (subsection 2F).* **SURVIVAL AND MORTALITY CURVES:** Both a high-credibility finding<sup>56</sup> and a moderate-credibility finding<sup>57</sup> demonstrated that ability to select a medication with better survival at specific time points was better with survival curves (gain framing) versus mortality curves (loss framing).

**POSITIVE VERSUS NEGATIVE FRAMING:** Four high- and moderate-credibility findings examined effect of framing on risky choice problems, which is a preference-based choice more than a selection of a normatively bigger or smaller option. Peters and Levin<sup>79</sup> and Damnjanovic and Gvozdenovic<sup>80</sup> demonstrated a tendency to pick risky options (v. sure-thing options) in a health domain involving negative outcomes when options were presented in negative frame (e.g., 10 out of 100 die) versus positive frame (e.g., 90 out of 100 saved). Two low-to-moderate-credibility findings published in the same article<sup>81</sup> also examined framing in a more complicated design, with inconsistent findings.

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Chance of comparison events	Weak ( $n = 1$ )	In a communication about the benefits and side effects of a cancer drug, people may be less swayed by the order of presentation if they are also told the probability of a different cancer.	In risk-benefit communication, adding the probabilities of other harms for comparison may reduce the effects of benefit v. harm ordering.
Interpretive labels	Insufficient (too few findings; $n = 1$ )	It is not clear whether adding interpretive labels to percentages (versus numbers alone) helps people identify the highest or lowest probability.	

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

Comparison	Evidence Strength	Applied Example of Evidence-Based Communication	General Guidance
Survival v. mortality curves	Strong ( $n = 2$ )	It is easier for people to identify the group with the higher survival at a specific time point if shown a survival curve rather than a mortality curve.	Showing probabilities over time as survival curves (versus mortality curves) improves ability to select groups with higher survival at specific time points.
Positive v. negative framing	Moderate ( $n = 4$ )	This formulation may influence people to choose the sure option A: "With A, 200 out of 600 people will survive. With B, 1/3 probability that all 600 people will survive and a 2/3 probability that 0 will survive." This formulation may influence people to choose the risky option B. "With A, 400 out of 600 people will die. With B, 1/3 probability that 0 people will die and a 2/3 probability that all 600 will die."	In choices about negative health outcomes, people are more likely to select a risky option (rather than a sure thing) when options are in negative frames (number with negative event) than when they are in positive frames (number without negative event).

**Table 2H** Evidence-Based Guidance for Effect of Manipulating Denominators of Probabilities on Ability to Contrast Probabilities

Comparison	Evidence Strength	General Guidance
Consistent denominator	Insufficient (too few findings; $n = 1$ )	It is not clear whether ability to rank events from least to most common is highest when all were shown as rates with the same denominator.

An additional finding is not summarized due to confounding of the data formats and the question format.<sup>73</sup>

*Comparisons of larger or smaller denominators on ability to contrast probabilities (subsection 2H).* CONSISTENT DENOMINATORS: Stone et al.<sup>42</sup> in a moderate-credibility finding demonstrated that ability to rank

events from least to most common was highest when all were shown as rates with the same denominator. When denominators were different, ability to rank them was poor (note that adding an icon array did not help).

*Comparisons of animation or interactivity on ability to contrast probabilities (subsection 2I).* INTERACTIVITY: A

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

Comparison	Strength of Evidence	Applied Example of Evidence-Based Communication	General Guidance
Interactivity	Weak ( $n = 1$ )	People may have an easier time identifying the smallest or largest probability if they use traditional static graphics and numbers rather than interactive graphics.	Ability to choose the largest or smallest probability may be better with static graphics (combined with rates per $10^n$ ) than with a particular type of interactivity in which participants drew a graph of the probability.
Animation	Insufficient (too few findings; $n = 2$ )	It is not clear whether animated or static graphics are better for helping people identify the larger or smaller of a series of probabilities.	

high-credibility finding<sup>82</sup> demonstrated that ability to choose the therapy with the lowest probability was better with static icon arrays combined with rate per  $10^n$  than with an interactive graphic that invited respondents to illustrate the probability expressed in the rate.

**ANIMATION:** A high-credibility finding<sup>70</sup> demonstrated no improvements (and some decreases) in ability to identify larger or equal probabilities with a variety of types of animated icon arrays versus static icon arrays. Similarly, a moderate-credibility finding<sup>61</sup> demonstrated no difference in ability to identify higher or lower values (combined with higher/lower effects of screening) between static icon arrays and 2 different types of animated icon arrays; the small sample reduces confidence.

### *Effects of Different Formats on Ability to Classify a Number into Categories (Categorization Outcome): Section 3*

Research in risk communication sometimes involves questions about whether the participant can recognize which category they fall into. For example, patients receiving a personalized chance of cancer recurrence may need to understand whether cancer experts consider this recurrence probability to be high, moderate, or low. Similarly, a patient may be able to make more informed decisions about lifestyle if they recognize that their chance of developing diabetes is above average. A question asking participants to assess which category a probability belongs to, when categories are provided, is called a categorization outcome.

Within the categorization outcome, there were no relevant findings for gain-loss framing (category 3F), denominator manipulation (3H), animation and interactivity (3I), or time period variation (3J).

*Comparisons between numerical formats in effects on categorization of probabilities (subsection 3A).* In a study comparing table versus text formats with a 12-item composite comprehension measure, 1 item pertained to contrast outcomes. However, no hypothesis testing was performed specific to this outcome, so this low-credibility finding is not synthesized.<sup>15</sup>

*Comparisons between graphical formats in effects on categorization of probabilities (subsection 3B).* **GRAPHIC TYPE:** In a moderate-credibility finding, Masson et al.<sup>27</sup> demonstrated no difference in people's ability to place their personal probability into provided categories between percentages combined with icon arrays, percentages combined with a bar chart, or a number line with verbal probability.

Three lower-credibility findings<sup>35,83,84</sup> are not synthesized due to limitations.

*Comparisons between numerical and graphical formats, and combinations of numerical and graphical formats, in effects on ability to classify a number into a category (subsection 3C).* **NUMBERS VERSUS GRAPHICS:** Two moderate-credibility findings contrasted numbers versus graphics for helping people to correctly classify a probability. Timmermans and Oudhoff<sup>85</sup> demonstrated that a verbal comparative probability (higher/lower than average) was as effective as numbers combined with an icon array showing a person's chance of disease as higher or lower than average. Brewer et al.<sup>86</sup> demonstrated that women were more accurate at classifying their probability of a health event into categories when it was displayed on a horizontal number line that superimposed the risk onto a labeled category than with various numbers or

**Table 3B** Evidence-Based Guidance for Effect of Graphical Format on Ability to Classify a Number into a Category

Comparison	Evidence Strength	General Guidance
Graphic type	Insufficient (too few findings; $n = 1$ )	It is not clear whether ability to assess which category a probability falls into is affected by the type of graphic used to convey the information.

graphics that included the category information only in a text label. A low-credibility finding<sup>84</sup> is not synthesized.

*Comparisons between numerical and verbal probabilities in effects on categorization of probabilities (subsection 3D).* **1-IN-X FORMAT VERSUS VERBAL PROBABILITY:** One moderate-credibility finding<sup>52</sup> suggests that women were more likely to correctly recognize they were in the “low-risk” category rather than the “no-risk” category with a 1-in-X rather than a verbal probability alone.

*Comparisons of elements added for context on categorization of probabilities (subsection 3E).* Three types of context features were added to probability information: interpretive labels about the categories, a population or average probability, and the chances of other events (“comparison risks”).

**INTERPRETIVE LABELS:** Three findings examined the effect of interpretive labels on people’s ability to correctly classify their probability of disease, generally finding that the labels helped. In a high-credibility finding, Marteau et al.<sup>87</sup> demonstrated that when a test result was labeled “abnormal,” there was no added effect of providing absolute or relative probability numbers. In a high-credibility finding, Johnson<sup>88</sup> demonstrated that, with air quality indicators, there was no difference between a label of “unhealthful” and “unhealthful for sensitive groups.” In a moderate-credibility finding of the Oncotype DX report, Brewer et al.<sup>86</sup> demonstrated that several graphics that showed cancer recurrence probability in graphics with interpretive labels outperformed the original Oncotype report, but multiple differences between the different graphics make it difficult to attribute the performance to the interpretive labels.

**ADDING A POPULATION VALUE:** Two findings examined the effect of adding a population value to a risk message. A moderate-credibility finding from Lipkus et al.<sup>89</sup> showed that women were better able to classify their own probability as higher or lower than others’ when they were shown the typical probability. In a moderate-credibility finding, Timmermans and Oudhoff<sup>85</sup> showed that displaying the population

probability improved accuracy of perception of personal probability (by increasing perceived risk); the format of the comparison to the average did not matter.

**COMPARISON RISKS:** In a moderate-credibility finding, Lipkus et al.<sup>90</sup> showed that displaying probabilities of different cancers for comparison did not affect ability to classify their own probability as higher or lower than others’.

Two lower-credibility findings<sup>83,84</sup> are not synthesized.

*Comparisons of methods of representing uncertainty on categorization of probabilities (subsection 3G).* In 1 moderate-credibility finding<sup>86</sup> comparing several formats as alternatives to the Oncotype Dx report, the presence/absence of confidence intervals in various formats did not appear to affect ability to classify probability.

*Comparisons of animation or interactivity on categorization of probabilities (subsection 3I).* A relevant finding was considered lower credibility because of the small sample size and lack of hypothesis testing specific to this comparison.<sup>35</sup>

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

It is widely accepted that it is poor communication practice to require patients to perform computations and that communicators should instead do calculations for their readers.<sup>91</sup> Nevertheless, there are situations in which the appropriate calculation cannot be performed for every reader who might need it, or in which readers might want to compute additional metrics that are not provided. For these reasons, researchers sometimes assess participants’ ability to perform computations on information provided in the stimulus. Performance on computations is likely to depend in part on numeracy and in part on the difficulty of the computation. Computation questions in the research we reviewed ranged from relatively simple (subtracting a percentage from another percentage to compute an absolute risk reduction) to somewhat more

**Table 3C** Evidence-Based Guidance for Contrasts between Numerical and Graphical Formats, and Combinations of Numerical and Graphical Formats, on Categorization of Probabilities

Comparison	Evidence Strength	General Guidance
Numbers v. graphics	Insufficient (too few findings; $n = 2$ )	It is not clear whether there is any difference between numbers or graphics in general for helping people understand how their chance of disease falls into categories.

**Table 3D** Evidence-Based Guidance for Contrasts between Numerical and Verbal Probabilities on Categorization of Probabilities

Comparison	Evidence Strength	General Guidance
1-in-X v. verbal probability	Insufficient (too few findings; $n = 1$ )	It is not clear whether 1 in X or verbal probability alone is more effective at helping people categorize a probability.

**Table 3E** Evidence-Based Guidance for Effect of Adding Context on Categorization of Probabilities

Comparison	Strength of Evidence	Applied Example of Evidence-Based Communication	General Guidance
Interpretive labels	Weak ( $n = 3$ )	People may have an easier time placing their test result into a category if it is labeled “normal” or “abnormal.”	Adding interpretive labels to numerical information may help people classify their probability better than showing numbers without labels.
Adding a population value	Weak ( $n = 2$ )	People may have an easier time placing their probability into a category such as “above average” if they can also see the population average probability.	Displaying population values in addition to an individual probability may help people classify their probability compared to that of others better than showing the individual probability alone.
Comparison risks	Insufficient (too few findings; $n = 1$ )	It is not clear whether adding 1 or more comparison probabilities (e.g., supplementing the chance of breast cancer with chances of different cancers) affects ability to classify probabilities accurately.	

difficult and multistep (converting a percentage into a rate per 1,000, thus expecting participants to understand they should first convert the percentage into a proportion and then multiply it by 1,000) to moderately complex (computing a relative increase or decrease). These issues of numeracy and heterogeneity of types of computations assessed make it impossible to attribute performance entirely to the stimulus. Nevertheless, performance on computations does provide some information about the clarity and ease of use of the data presentation format.

In the computation outcome, there were no relevant findings about time period variation (category 4J in Table B).

*Comparisons between numerical formats in effects on ability to perform computations on probabilities (subsection*

*4A).* **NUMERICAL FORMATS:** Two high-credibility findings (Cuite et al.<sup>64</sup> substudy 3 and Garcia-Retamero and Galesic,<sup>92</sup> respectively) and a moderate-credibility finding (Cuite et al.,<sup>64</sup> substudy 5) demonstrated that computations with probabilities were more accurate with percentages than with rate per 10<sup>n</sup> or 1 in X. However, a high-credibility finding<sup>65</sup> and moderate-credibility findings by Knapp et al.<sup>93</sup> and Ruiz et al.<sup>10</sup> demonstrated no effect of number format (percentage, 1 in X, rate per 10<sup>n</sup>, or combinations of these) on ability to perform computations with probability numbers. Two moderate-credibility findings<sup>12,94</sup> demonstrated greater accuracy with rate per 10<sup>n</sup> than with the rate-based probability band established by the European Commission (EC; e.g., “up to 1 in 10”). By contrast, a moderate-credibility finding (Hill and Brase<sup>95</sup> substudy 3) demonstrated better

**Table 3G** Evidence-Based Guidance for Effect of Stating or Illustrating Numerical Uncertainty on Categorization of Probabilities

Comparison	Evidence Strength	General Guidance
Confidence intervals	Insufficient (too few findings; $n = 1$ )	It is not clear whether ability to assess which category a probability falls into is affected by showing uncertainty in the form of a confidence interval.

calculations of cumulative probability with common-denominator rates than with probabilities. Two additional moderate-credibility findings (Hill and Brase<sup>95</sup> substudies 1 and 2) demonstrated no effect of number format (percentage, 1 in X, rate per 10<sup>n</sup>, or combinations) on ability to perform computations with probability numbers.

A lower-credibility finding<sup>96</sup> is not summarized.

*Comparisons between graphical formats in effects on ability to perform computations on probabilities (subsection 4B).* **ICON ARRAYS:** A moderate-credibility finding demonstrated an advantage when icon arrays were presented in combination with numbers (data labels) compared with icon arrays alone.<sup>97</sup> However, a second moderate-credibility finding suggested that supplementing numbers with a bar chart or data table was preferable to supplementing them with an icon array.<sup>98</sup>

A high-credibility finding by Okan et al.<sup>23</sup> did not find a difference in ability to perform computations between part-to-whole and numerator-only icon arrays.

**SURVIVAL CURVES:** Two low-to-moderate-credibility findings examining types of survival curves published in 1 article by Rakow et al.<sup>41</sup> demonstrated no differences by what sort of survival curve was used to display the probability.

**MULTIPLE GRAPHICS:** Three other moderate-credibility findings that contrasted multiple graphics (icon arrays both random and grouped, bar charts both horizontal and vertical, pie charts, flowcharts, number lines) either demonstrated no differences<sup>99</sup> or produced different rankings.<sup>40,68</sup> Another low-to-moderate-credibility finding in the Feldman-Stewart et al.<sup>40</sup> article was underpowered to determine differences between graph types.

Two lower-credibility findings were not summarized.<sup>47,100</sup>

*Comparisons between numerical and graphical formats, and combinations of numerical and graphical formats, in effects on ability to perform computations on probabilities (subsection 4C).* **COMBINING NUMBERS AND**

**GRAPHICS:** Five moderate- to high-credibility findings suggest that the combination of graphics and numbers is generally superior to either alone, in contrast to 3 moderate-to-lower-credibility findings that did not show this effect. Specifically, a finding by Garcia-Retamero and Galesic<sup>101</sup> and 2 findings from substudies in the same article by Garcia-Retamero et al.<sup>102</sup> demonstrated that computations using rates per 100 or 1,000 numbers were more accurate when sizes of the treated and untreated groups were the same or when the numbers were supplemented with icon arrays. Another finding by Garcia-Retamero and Dhimi<sup>103</sup> also showed that the combination of numbers and icon arrays was better than numbers alone. Similarly, a moderate-credibility finding<sup>104</sup> demonstrated that supplementing percentages with a bar chart helped facilitate computations. However, both a moderate-credibility finding<sup>10</sup> and 2 additional moderate findings (in the same Dragicevic and Jansen<sup>104</sup> article) demonstrated no evidence that supplementing numbers with graphics helps computations.

**NUMBERS VERSUS GRAPHICS:** In a moderate-credibility finding, Feldman-Stewart et al.<sup>40</sup> demonstrated that vertical bar charts were superior to numbers in terms of facilitating computations, but grouped icon arrays were about equal to numbers, and horizontal bar charts, random icon arrays, and pie charts were all worse. A low-to-moderate-credibility finding from a different substudy by the same author<sup>40</sup> had similar results, and another low-to-moderate-credibility finding<sup>45</sup> demonstrated computational ability was best when numbers were supplemented by bar charts than when they were shown alone or supplemented with a histogram.

Lower-credibility findings from 3 substudies published in 2 articles were not summarized.<sup>47,100</sup>

*Comparisons between numerical and verbal probabilities in effects on ability to perform computations on probabilities (subsection 4D).* **ADDING INTERPRETIVE LABELS:** In communicating chance of a series of side effects, adding a verbal label to a rate per 10<sup>n</sup> did not improve ability to compute the chance of having any side effect.<sup>12</sup>

**Table 4A** Evidence-Based Guidance for Effects of Numerical Format on Ability to Perform Computations on Probabilities

Comparison	Evidence Strength	General Guidance
Numerical formats (percentages, 1 in X, rate per 10 <sup>n</sup> , band)	Insufficient (inconsistent findings; $n = 11$ )	It is not clear whether different number formats consistently affect people's ability to do computations with probabilities across different sorts of computations.

*Comparisons of elements added for context on ability to perform computations on probabilities (subsection 4E).* ANECDOTES: In a moderate-credibility finding, Gibson et al.<sup>55</sup> demonstrated that ability to convert from proportion to percentages was better when accompanying anecdotes reinforced the proportions stated in the scenario.

A low-credibility finding from a relevant study is not summarized.<sup>45</sup>

*Comparisons of frames (gain, loss, combination) on ability to perform computations on probabilities (subsection 4F).* GAIN- VS LOSS-FRAMING: A single moderate-credibility finding<sup>105</sup> studied whether gain versus loss framing affects computations, finding no effect.

*Comparisons of methods of representing uncertainty on ability to perform computations on probabilities (subsection 4G).* One finding is not synthesized here due to limitations in this comparison.<sup>45</sup>

*Comparisons of larger or smaller denominators on ability to perform computations on probabilities (subsection 4H).* CONSISTENT VERSUS DIFFERENT DENOMINATORS: Presenting people with frequency information with different denominators impairs people's ability to compute relationships between probabilities, according to 4 moderate-to-high-credibility findings that showed that computations were more accurate when sizes of treated and untreated groups were the same (Garcia-Retamero and Galesic,<sup>101</sup> Garcia-Retamero et al. findings from studies 1<sup>102</sup> and 2<sup>102</sup>, Okan et al.<sup>105</sup>).

A lower-credibility finding<sup>103</sup> is not summarized.

*Comparisons of animation or interactivity on ability to perform computations on probabilities (subsection 4I).* Two moderate-credibility findings were inconsistent. Okan et al.<sup>97</sup> demonstrated that accuracy in computation was higher when icon arrays were animated and had reflective questions added. However, Houston et al.<sup>61</sup>

demonstrated no difference in ability to perform computations between static and animated icon arrays.

## Summary of Evidence

We categorized the majority of evidence from this synthesis as insufficient, with only small numbers of strong, moderate, and weak guidance statements.

There is **moderate or strong evidence** for the following:

- icon arrays with human figures are more memorable than icon arrays with blocks (subsection 1B: recall outcome, graphics v. graphics comparison);
- it is easier to identify points of highest survival with a survival curve than with a mortality curve (subsection 1B: identification outcome, graphics v. graphics comparison);
- adding an average population-level probability to a message about individual probability may not affect the memorability of the probability (subsection 1E: recall outcome, context comparison);
- computations are easier when bar charts are combined with data labels than with either numbers or graphics alone (subsection 4C: computation outcome, numbers v. graphics comparison);
- computations with frequencies are more accurate when denominators match (subsection 4A: computation outcome, numbers v. numbers comparison); and
- gain-loss framing has substantial effects on risky choices, with selection of a risky outcome being more common when options are presented in negative frames (subsection 2F: contrast outcome, framing comparison).

**Weak evidence** was more common and included the following:

- reducing significant digits may improve people's ability to identify numbers (subsection 1A: identification/recall outcome, numbers v. numbers comparison);

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

Comparison	Evidence Strength	General Guidance
Graphical formats	Insufficient (inconsistent findings; $n = 9$ )	It is not clear whether different graphics consistently affect people's ability to do computations with probabilities across different sorts of computations.

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

Comparison	Strength of Evidence	Applied Example of Evidence-Based Communication	General Guidance
Combining numbers and graphics	Moderate ( $n = 8$ )	It may be easier for people to do computations with a bar chart labeled with percentages than with percentages alone or a bar chart alone.	Combining graphics and numbers, as compared with providing either graphics or numbers alone, improves peoples' ability to perform calculations on probabilities. However, the effect probably depends on what computation they are asked to perform. (It is preferable not to ask people to perform computations.)
Numbers v. graphics	Insufficient (too few findings; $n = 3$ )	It is not clear whether any specific graphic formats are superior to any specific number formats in facilitating ability to perform computations on probabilities.	

- interpretive labels for numbers may improve people's ability to categorize (subsection 3E: categorization outcome, context comparison);
  - "heart age" communications may improve identification/recall (subsection 1A: identification/recall outcome, numbers v. numbers comparison);
  - table formats may improve ability to identify numbers (subsection 1A: identification/recall outcome, numbers v. numbers comparison);
  - numerator-denominator icon arrays may improve both recall and contrast ability over numerator-only ones (subsections 1B/1C and 2C: recall and contrast outcomes, graphics v. graphics and numbers v. graphics comparisons);
  - adding labels may improve ability to identify the option with a higher rate (subsection 2E: contrast outcome, context comparison);
  - rates per 100 may improve ability to identify or remember numbers as compared with numerator-only icon arrays (subsection 1C: identification/recall outcome, numbers v. graphics comparison);
  - using point estimates rather than ranges may improve identification/recall (subsection 1G: identification/recall outcome, uncertainty comparison);
  - matching formats and denominators of probabilities and presenting benefit probabilities before harm probabilities may each improve ability to find highest or lowest values (subsection 2A: contrast outcome, numbers v. numbers comparison);
  - adding probabilities of other harms for context may also improve ability to find high/low values (subsection 2E: contrast outcome, context comparison); and
  - showing the population probability may help people understand what category their probability falls into (subsection 3E: categorization outcome, context comparison).
- Weak evidence also suggested a lack of benefit of different types of animation (subsections 2I: contrast outcome, animation comparison). However, notably, despite the ubiquity of online and computer-mediated health communication, animation and interactivity have not been well-studied to date.

## Discussion

This article synthesizes the evidence pertaining to the impact of data presentation formats on 5 outcomes

**Table 4D** Evidence-Based Guidance for Contrasts between Numerical and Verbal Probabilities, or Combinations of Numerical and Verbal Probabilities, on Ability to Perform Computations on Probabilities

Comparison	Evidence Strength	General Guidance
Adding interpretive labels	Insufficient (too few findings; $n = 1$ )	It is not clear whether adding verbal labels to numbers affects people's ability to do computations with probabilities across different sorts of computations.

**Table 4E** Evidence-Based Guidance for Effect of Adding Context on Ability to Perform Computations on Probabilities

Comparison	Evidence Strength	General Guidance
Anecdotes	Insufficient (too few findings; $n = 1$ )	It is not clear whether providing contextual features such as anecdotes affects people's ability to do computations with probabilities across different sorts of computations.

**Table 4F** Evidence-Based Guidance for Effect of Gain-Loss Framing on Ability to Perform Computations on Probabilities

Comparison	Evidence Strength	General Guidance
Gain v. loss framing	Insufficient (too few findings; $n = 1$ )	It is not clear whether the framing of probability information affects ability to compute relationships between probabilities across different sorts of computations. The optimal framing probably depends on what computation the reader is asked to perform.

**Table 4H** Evidence-Based Guidance for Effect of Manipulating Denominators of Probabilities on Ability to Perform Computations on Probabilities

Comparison	Strength of Evidence	Applied Example of Evidence-Based Communication	General Guidance
Consistent v. different denominators	Strong ( $n = 4$ )	It is easier for people to perform computations on probabilities shown as 25 in 1,000 and 50 in 1,000 than on probabilities shown as 1 in 40 and 1 in 20.	Presenting rates or frequencies with matching denominators, as compared with different denominators, improves peoples' ability to compute relationships between the probabilities. However, the effect probably depends on what computation they are asked to perform, and it is preferable not to ask people to perform computations.

**Table 4I** Evidence-Based Guidance for Effect of Animation or Interactivity on Ability to Perform Computations

Comparison	Evidence Strength	General Guidance
Animation v. static	Insufficient (inconsistent findings; $n = 2$ )	It is not clear whether animation affects ability to perform computations.

(identification, recall, contrast, categorization, computation) involving performing point tasks on stimuli containing probabilities.

The granular nature of this evidence showcases nuances in the conclusions that can be drawn. For example, evidence is strong that survival curves are better than mortality curves for helping people identify the point at which survival is the highest (the contrast outcome), and yet there is insufficient evidence to determine whether they are also better for helping people identify the exact probability at 1 point of time (the identification outcome).

The outcomes discussed here are sometimes lumped together in both the research and practice literatures as comprehension or sometimes broken down as verbatim versus gist comprehension. Indeed, improving what is alternately described as audience comprehension, understanding, or knowledge of probability information is perhaps the most commonly stated objective for probability communications. Yet, as this review shows, it is precisely the imprecision of such terms that has led to confusion about what is and is not known about the effectiveness of different probability data presentation formats. Indeed, it matters whether the people receiving probability information need to simply recognize what is in front of them (identification), remember it later (recall), or perform numerical operations on it (computation). It matters whether they need to compare a number to other data points (contrast) or to a set of thresholds that define classes or categories (categorization). Part of the reason that we were able to draw so few strong conclusions from our review is that those studies that do exist often compared similar formats using distinctly different measures.

As a result, as yet unanswered with the current evidence are broad questions such as, “Is there a specific number format or graphic that is superior to other formats for all of these outcomes?” “In general, does adding a graphic to a number improve multiple outcomes?” and “Is recall of the probability of an event better with numbers alone or verbal probabilities alone?” While resolution of these questions must await further research, the heterogeneity of findings across the outcomes studied in this article suggest that there may be very few formats that improve all outcomes. Instead, our evidence is consistent with the idea that the right question will usually be “which format or graphic is superior for the specific outcome that is most important in this particular context?”

Limitations include the possibility of overlooked studies, the use of a small expert group to evaluate risk of bias and credibility, and the highly granular data

extraction that focused on narrow comparisons rather than global assessment of research. The literature was heterogeneous, ranging from large to very small sample sizes, including strong and weak study designs, and different types of participants with some including patients, other members of the general public, and others primarily undergraduates or other educated samples. Overall, the number of highly credible comparable articles within any category was small, limiting the strength of the evidence that could be derived from this literature. We did not analyze articles by participant or population characteristics (such as education, culture, or numeracy) because of the relatively small numbers of comparable papers for each relevant characteristic. Such potential confounders might contribute to the heterogeneity of findings when studies are grouped and may also limit generalizability to different settings and populations.


As illustrated previously in Table A, this review addresses the research evidence only regarding point cognitive tasks, that is, situations in which the audience for probability communications is asked to focus on single data points (presented either singly or in larger sets). It does not touch upon the evidence for communicating probability differences (difference tasks), time trends (trend tasks), or situations in which the audience is asked to consider multiple types of probability information simultaneously (synthesis tasks). It also considers only a narrow slice of 5 specific outcomes, while a companion article presents the set of evidence pertaining to the effect of data presentation formats on 7 additional outcomes involving point tasks (probability perceptions, probability feelings, behavioral intentions, behavior, trust in information, preference for a data presentation format, and discrimination).<sup>106</sup> As such, it represents a fundamentally incomplete picture of the implications of using particular data presentation formats for communicating probabilities. We urge readers to consider this article as but 1 segment of the larger compendium of findings from this project. To the extent that the research evidence presented here provides guidance to the practice of probability communication, the findings shown here must be integrated with and balanced against the findings regarding both the other cognitive tasks that users may need to perform and the larger set of outcomes that such communications create.

### Acknowledgments

We thank the Numeracy Expert Panel for contributions to conceptualizing the MNM project (Cynthia Baur, Sara Cjaza, Angela Fagerlin, Carolyn Petersen, Rima Rudd, Michael Wolf,

and Steven Woloshin). We are grateful to Marianne Sharko, MD, MS, Andrew Z. Liu, MPH, and Lisa Grossman Liu, MD, PhD, for contributions to article screening and risk-of-bias assessment. We also thank Jordan Brutus for assisting with data management.

## ORCID iDs

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

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

## Availability of Research Resources

All research resources are available at the Making Numbers Meaningful Project at OSF (<https://osf.io/rvxf2/>). This project includes a Methodology Files folder (containing the search strategy, the data extraction instrument, and the study risk of bias [S-ROB] rubric), the list of each included article mapped to the Making Numbers Meaningful review article that covers it, and a Probability Findings folder displaying the extracted findings for each of the Making Numbers Meaningful review articles.

## References

1. 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
2. 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
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. Witteman HO, Zikmund-Fisher BJ, Waters EA, Gavaruzzi T, Fagerlin A. Risk estimates from an online risk calculator are more believable and recalled better when expressed as integers. *J Med Internet Res.* 2011;13(3):e54. DOI: 10.2196/jmir.1656
6. 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
7. Woloshin S, Schwartz LM. Communicating data about the benefits and harms of treatment: a randomized trial. *Ann Intern Med.* 2011;155(2):87–96. DOI: 10.7326/0003-4819-155-2-201107190-00004
8. Henneman L, van Asperen CJ, Oosterwijk JC, Menko FH, Claassen L, Timmermans DR. Do preferred risk formats lead to better understanding? A multicenter controlled trial on communicating familial breast cancer risks using different risk formats. *Patient Prefer Adherence.* 2020;14:333–42. DOI: 10.2147/PPA.S232941
9. Garcia-Retamero R, Galesic M. Using plausible group sizes to communicate information about medical risks. *Patient Educ Couns.* 2011;84(2):245–50. DOI: 10.1016/j.pec.2010.07.027
10. 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
11. Sinayev A, Peters E, Tusler M, Fraenkel L. Presenting numeric information with percentages and descriptive risk labels: a randomized trial. *Med Decis Making.* 2015;35(8):937–47. DOI: 10.1177/0272989X15584922
12. Knapp P, Gardner PH, Raynor DK, Woolf E, McMillan B. Perceived risk of tamoxifen side effects: a study of the use of absolute frequencies or frequency bands, with or without verbal descriptors. *Patient Educ Couns.* 2010;79(2):267–71. DOI: 10.1016/j.pec.2009.10.002
13. 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
14. 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
15. Brick C, McDowell M, Freeman ALJ. Risk communication in tables versus text: a registered report randomized trial on ‘fact boxes’. *R Soc Open Sci.* 2020;7(3):190876. DOI: 10.1098/rsos.190876
16. Mühlbauer V, Prinz R, Mühlhauser I, Wegwarth O. Alternative package leaflets improve people’s understanding of drug side effects—a randomized controlled exploratory survey. *PLoS One.* 2018;13(9):e0203800. DOI: 10.1371/journal.pone.0203800
17. Sullivan HW, O’Donoghue AC, Aikin KJ. Communicating benefit and risk information in direct-to-consumer print advertisements: a randomized study. *Ther Innov Regul Sci.* 2015;49(4):493–502. DOI: 10.1177/2168479015572370
18. Callison C, Gibson R, Zillmann D. How to report quantitative information in news stories. *Newsp Res J.* 2009;30(2):43–55. DOI: 10.1177/073953290903000205
19. 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

20. Fausset CB, Rogers WA. Younger and older adults' comprehension of health risk probabilities: understanding the relationship between format and numeracy. *Proc Hum Factors Ergon Soc Annu Meet.* 2012;56(1):120–4. DOI: 10.1177/1071181312561002
21. Kreuzmair C, Siegrist M, Keller C. Does iconicity in pictographs matter? The influence of iconicity and numeracy on information processing, decision making, and liking in an eye-tracking study. *Risk Anal.* 2016;37(3):546–66. DOI: 10.1111/risa.12623
22. Zikmund-Fisher BJ, Witteman HO, Dickson M, et al. Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. *Med Decis Making.* 2014;34(4):443–53. DOI: 10.1177/0272989X13511706
23. 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
24. 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):109–17. DOI: 10.1016/j.pec.2020.06.031
25. 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
26. Fraenkel L, Nowell WB, Stake CE, et al. The impact of information presentation format on preference for total knee replacement surgery. *Arthritis Care Res.* 2019;71(3):379–84. DOI: 10.1002/acr.23605
27. 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
28. Hamstra DA, Johnson SB, Daignault S, et al. The impact of numeracy on verbatim knowledge of the longitudinal risk for prostate cancer recurrence following radiation therapy. *Med Decis Making.* 2015;35(1):27–36. DOI: 10.1177/0272989X14551639
29. Ghosh K, Crawford BJ, Pruthi S, et al. Frequency format diagram and probability chart for breast cancer risk communication: a prospective, randomized trial. *BMC Womens Health.* 2008;8:18. DOI: 10.1186/1472-6874-8-18
30. Martin RW, Brower ME, Gerald A, Gallagher PJ, Tellinhuysen 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
31. Hawley S, Zikmund-Fisher B, Ubel P, Jankovic A, Lucas T, Fagerlin A. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Educ Couns.* 2008;73(3):448–55. DOI: 10.1016/j.pec.2008.07.023
32. Zikmund-Fisher B, Fagerlin A, Ubel P. A demonstration of “less can be more” in risk graphics. *Med Decis Making.* 2010;30(6):661–71. DOI: 10.1177/0272989X10364244
33. McDowell M, Gigerenzer G, Wegwarth O, Rebitschek FG. Effect of tabular and icon fact box formats on comprehension of benefits and harms of prostate cancer screening: a randomized trial. *Med Decis Making.* 2019;39(1):41–56. DOI: 10.1177/0272989X18818166
34. Ancker JS, Weber EU, Kukafka R. Effect of arrangement of stick figures on estimates of proportion in risk graphics. *Med Decis Making.* 2011;31(1):143–50.
35. 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
36. Kasper J, Heesen C, Köpke S, Mühlhauser I, Lenz M. Why not? Communicating stochastic information by use of unsorted frequency pictograms—a randomised controlled trial. *Psychosoc Med.* 2011;8:Doc08. DOI: 10.3205/psm000077
37. Tait AR, Voepel-Lewis T, Brennan-Martinez C, McGonegal M, Levine R. Using animated computer-generated text and graphics to depict the risks and benefits of medical treatment. *Am J Med.* 2012;125(11):1103–10. DOI: 10.1016/j.amjmed.2012.04.040
38. Barnes AJ, Hanoch Y, Miron-Shatz T, Ozanne EM. Tailoring risk communication to improve comprehension: do patient preferences help or hurt? *Health Psychol.* 2016;35(9):1007–16. DOI: 10.1037/hea0000367
39. Kasper J, van de Roemer A, Pöttgen J, et al. A new graphical format to communicate treatment effects to patients—a Web-based randomized controlled trial. *Health Expect.* 2017;20(4):797–804. DOI: 10.1111/hex.12522
40. 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.
41. Rakow T, Wright RJ, Bull C, Spiegelhalter DJ. Simple and multistate survival curves can people learn to use them? *Med Decis Making.* 2012;32(6):792–804. DOI: 10.1177/0272989X12451057
42. Stone ER, Gabard AR, Groves AE, Lipkus IM. Effects of numerical versus foreground-only icon displays on understanding of risk magnitudes. *J Health Commun.* 2015;20(10):1230–41. DOI: 10.1080/10810730.2015.1018594
43. Reder M, Thygesen LC. Crowd-figure-pictograms improve women's knowledge about mammography screening: results from a randomised controlled trial. *BMC Res Notes.* 2018;11(1):332. DOI: 10.1186/s13104-018-3437-z
44. Fagerlin A, Wang C, Ubel PA. Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? *Med Decis Making.* 2005;25(4):398–405.

45. Gibson JM, Rowe A, Stone ER, Bruine De Bruin W. Communicating quantitative information about unexploded ordnance risks to the public. *Environ Sci Technol*. 2013;47(9):4004–13. DOI: 10.1021/es305254j
46. Lipkus IM, Klein WM, Rimer BK. Communicating breast cancer risks to women using different formats. *Cancer Epidemiol Biomarkers Prev*. 2001;10(8):895–8.
47. Schonlau M, Peters E. Comprehension of graphs and tables depend on the task: empirical evidence from two Web-based studies. *Stat Politics Policy*. 2012;3:1–35.
48. Davis CR, McNair AG, Brigic A, et al. Optimising methods for communicating survival data to patients undergoing cancer surgery. *Eur J Cancer*. 2010;46(18):3192–9. DOI: 10.1016/j.ejca.2010.07.030
49. Henneman L, Oosterwijk JC, Van Asperen CJ, et al. The effectiveness of a graphical presentation in addition to a frequency format in the context of familial breast cancer risk communication: a multicenter controlled trial. *BMC Med Inform Decis Mak*. 2013;13(1):55. DOI: 10.1186/1472-6947-13-55
50. Mason D, Boase S, Marteau T, et al. One-week recall of health risk information and individual differences in attention to bar charts. *Health Risk Soc*. 2014;16(2):136–53. DOI: 10.1080/13698575.2014.884544
51. Andreadis K, Chan E, Park M, et al. Imprecision and preferences in interpretation of verbal probabilities in health: a systematic review. *J Gen Intern Med*. 2021;36(12):3820–9. DOI: 10.1007/s11606-021-07050-7
52. Marteau TM, Saidi G, Goodburn S, Lawton J, Michie S, Bobrow M. Numbers or words? A randomized controlled trial of presenting screen negative results to pregnant women. *Prenat Diagn*. 2000;20(9):714–8.
53. Weinstein ND, Atwood K, Puleo E, Fletcher R, Colditz G, Emmons K. Colon cancer: risk perceptions and risk communication. *J Health Commun*. 2004;9(1):53–65.
54. Betsch C, Ulshöfer C, Renkewitz F, Betsch T. The influence of narrative v. statistical information on perceiving vaccination risks. *Med Decis Making*. 2011;31(5):742–53. DOI: 10.1177/0272989X11400419
55. Gibson R, Callison C, Zillmann D. Quantitative literacy and affective reactivity in processing statistical information and case histories in the news. *Media Psychol*. 2011;14(1):96–120. DOI: 10.1080/15213269.2010.547830
56. Zikmund-Fisher BJ, Fagerlin A, Ubel PA. Mortality versus survival graphs: improving temporal consistency in perceptions of treatment effectiveness. *Patient Educ Couns*. 2007;66(1):100–7.
57. Armstrong K, Schwartz JS, Fitzgerald G, Putt M, Ubel PA. Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. *Med Decis Making*. 2002;22(1):76–83.
58. Huys J, Evers-Kiebooms G, d'Ydewalle G. Framing biases in genetic risk perception. *Adv Psychol*. 1990;68:59–68. DOI: 10.1016/S0166-4115(08)61315-1
59. Sladakov J, Jansen J, Hersch J, Turner R, McCaffery K. The differential effects of presenting uncertainty around benefits and harms on treatment decision making. *Patient Educ Couns*. 2016;99(6):974–80. DOI: 10.1016/j.pec.2016.01.009
60. Witteman HO, Fuhrel-Forbis A, Wijesundera HC, et al. Animated randomness, avatars, movement, and personalization in risk graphics. *J Med Internet Res*. 2014;16(3):e80. DOI: 10.2196/jmir.2895
61. Houston AJ, Kamath GR, Bevers TB, et al. Does animation improve comprehension of risk information in patients with low health literacy? A randomized trial. *Med Decis Making*. 2020;40(1):17–28.
62. Grimes DA, Snively GR. Patients' understanding of medical risks: implications for genetic counseling. *Obstet Gynecol*. 1999;93(6):910–4.
63. Van Vliet HAAM, Grimes DA, Popkin B, Smith U. Lay persons' understanding of the risk of Down's syndrome in genetic counselling. *Br J Obstet Gynaecol*. 2001;108(6):649–50. DOI: 10.1016/S0306-5456(00)00151-0
64. Cuite CL, Weinstein ND, Emmons K, Colditz G. A test of numeric formats for communicating risk probabilities. *Med Decis Making*. 2008;28(3):377–84. DOI: 10.1177/0272989X08315246
65. Pighin S, Savadori L, Barilli E, et al. Communicating Down syndrome risk according to maternal age: “1-in-X” effect on perceived risk. *Prenat Diagn*. 2015;35(8):777–82. DOI: 10.1002/pd.4606
66. 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
67. Ubel PA, Smith DM, Zikmund-Fisher BJ, et al. Testing whether decision aids introduce cognitive biases: results of a randomized trial. *Patient Educ Couns*. 2010;80(2):158–63. DOI: 10.1016/j.pec.2009.10.021
68. 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
69. Waller J, Whitaker KL, Winstanley K, Power E, Wardle J. A survey study of women's responses to information about overdiagnosis in breast cancer screening in Britain. *Br J Cancer*. 2014;111(9):1831–5. DOI: 10.1038/bjc.2014.482
70. Zikmund-Fisher B, Witteman H, Fuhrel-Forbis A, Exe N, Kahn V, Dickson M. Animated graphics for comparing two risks: a cautionary tale. *J Med Internet Res*. 2012;14(4):e106. DOI: 10.2196/jmir.2030
71. Feldman-Stewart D, Brundage MD, Zotov V. Further insight into the perception of quantitative information: judgments of gist in treatment decisions. *Med Decis Making*. 2007;27(1):34–43. DOI: 10.1177/0272989X06297101
72. Wright AJ, Whitwell SCL, Takeichi C, Hankins M, Marteau TM. The impact of numeracy on reactions to different graphic risk presentation formats: an experimental analogue study. *Br J Health Psychol*. 2009;14(pt 1):107–25. DOI: 10.1348/135910708X304432

73. Okan Y, Galesic M, Garcia-Retamero R. How people with low and high graph literacy process health graphs: evidence from eye-tracking. *J Behav Decis Making*. 2016;29(2-3): 271–94. DOI: 10.1002/bdm.1891
74. Downen F, Sidhu K, Broadbent E, Pilmore H. Communicating projected survival with treatments for chronic kidney disease: patient comprehension and perspectives on visual aids. *BMC Med Inform Decis Mak*. 2017;17(1):137. DOI: 10.1186/s12911-017-0536-z
75. Fraenkel L, Peters E, Tyra S, Oelberg D. Shared medical decision making in lung cancer screening: experienced versus descriptive risk formats. *Med Decis Making*. 2015;36(4):518–25. DOI: 10.1177/0272989X15611083
76. Timmermans D, Molewijk B, Stiggelbout A, Kievit J. Different formats for communicating surgical risks to patients and the effect on choice of treatment. *Patient Educ Couns*. 2004;54(3):255–63.
77. Man-Son-Hing M, O'Connor AM, Drake E, Biggs J, Hum V, Laupacis A. The effect of qualitative vs. quantitative presentation of probability estimates on patient decision-making: a randomized trial. *Health Expect*. 2002;5(3): 246–55.
78. Steiner MJ, Dalebout S, Condon S, Dominik R, Trussell J. Understanding risk: a randomized controlled trial of communicating contraceptive effectiveness. *Obstet Gynecol*. 2003;102(4):709–17.
79. Peters E, Levin IP. Dissecting the risky-choice framing effect: numeracy as an individual-difference factor in weighting risky and riskless options. *Judgm Decis Mak*. 2008;3(6):435–48.
80. Damjanovic K, Gvozdenovic V. Influence of the probability level on the framing effect. *Psihologijske Teme*. 2016;25(3):405–29.
81. Kühberger A. The framing of decisions: a new look at old problems. *Organ Behav Hum Decis Process*. 1995;62(2): 230–40. DOI: 10.1006/obhd.1995.1046
82. Zikmund-Fisher BJ, Dickson M, Witteman HO. Cool but counterproductive: interactive, Web-based risk communications can backfire. *J Med Internet Res*. 2011;13(3):e60.
83. Edmonds SW, Cram P, Lu X, et al. Improving bone mineral density reporting to patients with an illustration of personal fracture risk. *BMC Med Inform Decis Mak*. 2014;14(1):101. DOI: 10.1186/s12911-014-0101-y
84. Damman OC, Vonk SI, van den Haak MJ, van Hooijdonk CMJ, Timmermans DRM. The effects of infographics and several quantitative versus qualitative formats for cardiovascular disease risk, including heart age, on people's risk understanding. *Patient Educ Couns*. 2018;101(8):1410–8. DOI: 10.1016/j.pec.2018.03.015
85. Timmermans DR, Oudhoff J. Indicating risks in the Dutch Cancer Society Cancer Risk Test: indicating population risks improves risk perception. *Ned Tijdschr Geneesk*. 2012;156(21):A4961.
86. Brewer NT, Richman AR, Defrank JT, Reyna VF, Carey LA. Improving communication of breast cancer recurrence risk. *Breast Cancer Res Treat*. 2012;133(2):553–61. DOI: 10.1007/s10549-011-1791-9
87. Marteau TM, Senior V, Sasieni P. Women's understanding of a "normal smear test result": experimental questionnaire based study. *BMJ*. 2001;322(7285):526–8.
88. Johnson BB. Communicating air quality information: experimental evaluation of alternative formats. *Risk Anal*. 2003;23(1):91–103. DOI: 10.1111/1539-6924.00292
89. Lipkus IM, Biradavolu M, Fenn K, Keller P, Rimer BK. Informing women about their breast cancer risks: truth and consequences. *Health Commun*. 2001;13(2):205–26. DOI: 10.1207/S15327027HC1302\_5
90. Lipkus IM, Crawford Y, Fenn K, et al. Testing different formats for communicating colorectal cancer risk. *J Health Commun*. 1999;4(4):311–24. DOI: 10.1080/108107399126841
91. Shoemaker SJ, Wolf MS, Brach C. *The Patient Education Materials Assessment Tool (PEMAT) and User's Guide*. AHRQ Publication 14-0002-EF. Rockville (MD): Agency for Healthcare Research and Quality; 2013.
92. Garcia-Retamero R, Galesic M. Who profits from visual aids: overcoming challenges in people's understanding of risks [published erratum appears in *Soc Sci Med*. 2010;70(12):2097]. *Soc Sci Med*. 2010;70(7):1019–25. DOI: 10.1016/j.socscimed.2009.11.031
93. Knapp P, Gardner P, McMillan B, Raynor DK, Woolf E. Evaluating a combined (frequency and percentage) risk expression to communicate information on medicine side effects to patients. *Int J Pharm Pract*. 2013;21(4):226–32. DOI: 10.1111/j.2042-7174.2012.00254.x
94. Knapp P, Raynor DK, Woolf E, Gardner PH, Carrigan N, McMillan B. Communicating the risk of side effects to patients: an evaluation of UK regulatory recommendations. *Drug Saf*. 2009;32(10):837–49. DOI: 10.2165/11316570-000000000-00000
95. Hill WT, Brase GL. When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Q J Exp Psychol*. 2012;65(12):2343–68. DOI: 10.1080/17470218.2012.687004
96. Knapp P, Gardner P, McMillan B, Raynor DK, Woolf L. The effects of numeracy on the perceived risk of medicine side effects. *Int J Pharm Pract*. 2010;2:23–24. DOI: 10.1111/j.2042-7174.2010.tb00509.x
97. 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
98. 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
99. Etnel JRG, de Groot JM, El Jabri M, et al. Do risk visualizations improve the understanding of numerical risks? A randomized, investigator-blinded general population

- survey. *Int J Med Inform.* 2020;135:104005. DOI: 10.1016/j.ijmedinf.2019.104005
100. Leonhardt JM, Robin Keller L. Do pictographs affect probability comprehension and risk perception of multiple-risk communications? *J Consum Aff.* 2018;52(3): 756–69. DOI: 10.1111/joca.12185
  101. Garcia-Retamero R, Galesic M. Communicating treatment risk reduction to people with low numeracy skills: a cross-cultural comparison. *Am J Public Health.* 2009;99(12):2196–202. DOI: 10.2105/AJPH.2009.160234
  102. Garcia-Retamero R, Galesic M, Gigerenzer G. Do icon arrays help reduce denominator neglect? *Med Decis Mak-ing.* 2010;30(6):672–84. DOI: 10.1177/0272989X10369000
  103. 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
  104. 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
  105. Okan Y, Garcia-Retamero R, Cokely ET, Maldonado A. Individual differences in graph literacy: overcoming denominator neglect in risk comprehension. *J Behav Decis Mak.* 2012;25(4):390–401. DOI: 10.1002/bdm.751
  106. Ancker JS, Benda NC, Sharma MM, et al. How point (single-probability) tasks are affected by probability format, part 2: a Making Numbers Meaningful systematic review. *MDM Policy Pract.* 2025;10(1):23814683241255337. DOI: 10.1177/23814683241255337