

Running Header: Communicating Risk with Icon Arrays

Witt, J. K., & Dhami, M. K. (2023). Visual organization of icon arrays affects bayesian reasoning and risk judgments. *Journal of Applied Research in Memory and Cognition*, 12(2), 241–254. <https://doi.org/10.1037/mac0000044>

Visual Organization of Icon Arrays Affects Bayesian Reasoning and Risk Judgments

Jessica K. Witt¹ & Mandeep K. Dhami^{2*}

¹Colorado State University

²Middlesex University London

*Corresponding Author (m.dhami@mdx.ac.uk)

Author Note

Jessica K. Witt, Department of Psychology, Colorado State University. Mandeep K. Dhami, Department of Psychology, Middlesex University, London.

Support for this research was provided by the National Science Foundation (grant numbers BCS-1632222 and SES-2030059 to JKW). Stimuli, data, and analysis scripts can be found at <https://osf.io/s569p/>. Please address correspondence to Mandeep K. Dhami, m.dhami@mdx.ac.uk Department of Psychology, Middlesex University, The Burroughs, Hendon, London, NW4 4BT, UK

Abstract

Icon arrays can help patients to understand statistical information such as their risk of having a disease given a diagnostic test result. Bayesian problems such as these are computationally challenging but can be solved via the simple ratio of $\Sigma(\text{disease}|\text{positive})/\Sigma(\text{positive cases})$. We tested whether organizing an icon array to place the two sides of this ratio nearer each other (proximal condition) enhanced risk judgments compared with an organization separating them (distal condition). Organizing icons proximally improved Bayesian reasoning by 31%-43% across two experiments. Participants were more likely to use the simple ratio in the proximal condition, but they used a different, less accurate ratio to judge risk in the distal condition. Current guidelines for risk communication endorse icon arrays but do not provide detailed design instructions. The present research demonstrates the importance of array organization and, in particular, icon proximity for communicating risk and facilitating Bayesian reasoning.

Keywords: Data visualization; Icon arrays; Bayesian Reasoning; Risk Judgment; Medical decision-making

General Audience Summary

Medical diagnostic and screening tests aim to detect specific conditions such as the presence of a disease or genetic anomaly. Such tests can be imperfect and so a person must evaluate the risk given the test result and the types of errors that the test is likely to make. This requires Bayesian reasoning (Bayes theorem) and is cognitively computationally challenging. In order to help people in this reasoning task, one technique is to present the numeric information in a data visualization such as an icon array. An icon array, also known as a pictograph, is a set of icons that are organized through spacing, color, or other visual characteristics to indicate the proportion of people who, for example, have the disease or who tested positive on a diagnostic test. In the present research, we tested whether the organization of an icon array can promote the calculation of risk via a simple ratio that is mathematically equivalent to Bayes theorem. We found that people were more accurate at estimating risk, and more likely to use the computationally easy ratio, when the icons were grouped by test result and colored by disease than when they were grouped by disease and colored by test result. Thus, a simple re-arrangement of the icons can lead to vast improvements in risk comprehension.

Visual Organization of Icon Arrays Affects Bayesian Reasoning and Risk Judgments

Diagnostic tests typically communicate a binary result i.e., positive or negative. These tests are not always 100% accurate and reliable, and so there may be false positives or false negatives. Thus, decisions should not be based on the test result alone but rather on the risk of the disease given the result and diagnostic characteristics of the test. For example, a COVID-19 test may return a negative result if the test is taken too early after disease onset, so the infected person should continue to take precautions rather than trust the negative result (Jetelina, 2021). As another example, prenatal testing may return a false positive for which a genetic abnormality is detected by the test, but the baby does not have Down syndrome. False positives for breast cancer detected with mammography occur at such a high rate that it is debatable when, or even if, annual mammograms should commence (Salisbury, 2019). Therefore, decisions about, for example, treatment, should be based on the positive predictive value (PPV) and negative predictive value (NPV) of the test rather than on the test result per se. But calculating PPV and NPV requires Bayesian reasoning.

Bayesian reasoning is cognitively computationally complex. Although the math is straight forward (see Bayes theorem; Equations 1-2), people often err. In some cases, mistakes are made because the base rate is ignored (Bar-Hillel, 1980). For example, people conclude that an offending car is green rather than blue based on an eyewitness report even when the base rates for green cars are much lower than for blue cars (Tversky & Kahneman, 1982). Similarly, people may worry that a positive mammogram signifies breast cancer even when the PPV is low due to low base rates. These base rate neglect errors are also made by experts such as physicians (Gigerenzer et al., 2008).

$$PPV = P(Disease | Positive) = \frac{P(Positive | Disease) * P(Disease)}{P(Positive)} \quad [1]$$

$$NPV = P(Disease | Negative) = \frac{P(Negative | Disease) * P(Disease)}{P(Negative)} \quad [2]$$

The idea that people are poor Bayesian reasoners has prompted researchers to explore ways to improve their reasoning abilities such as through presenting the numerical information in different formats (e.g., frequencies; Gigerenzer & Hoffrage, 1995) and by provision of equivalent non-numerical information such as visual decision aids (e.g., Galesic et al., 2009; Garcia-Retamero & Dhimi, 2011, 2013). The information structure of Bayesian reasoning tasks can be altered to facilitate the extraction and processing of elements of the Bayesian equation. Past research has shown that changing the way numeric information is presented (e.g., moving from normalized formats such as percentages to non-normalized formats such as frequencies) restructures the task (e.g., into nested sets) and improves performance (for a review see Johnson & Tubau, 2015).

Visualizations can further reduce the complexity of the task, making the problem structure more transparent. These have been found to improve performance for both categories of numeric format (see Johnson & Tubau, 2015). The present research builds on the body of visualization work that involves visual icons. The recommendation in the literature is to communicate risk visually using icon arrays (also called pictographs) rather than numerically because icon arrays can increase understanding of risk amongst those with low numeracy and do not suffer from the problems of other graphical displays due to low graph literacy (e.g., see Garcia-Retamero & Cokely, 2013; 2017; Waters, Fagerlin, & Zikmund-Fisher, 2016; Zipkin et al., 2014).

In an icon array, proportions are shown by altering the appearance or placement of some icons relative to others. For example, an icon array can use colors such as red and gray to signify the relative number of people who have a disease and those who do not. Icon arrays can also use spatial distance to classify icons into two categories such as those who had a positive test result versus a negative one. A combination of color and spatial distance can be used if there are multiple categories of disease occurrence and test result (see Figure 1).

Studies have explored the effects of different aspects of icon array design such as whether arrays should contain icons versus symbols (e.g., Gaissmaier et al., 2012; Stone et al., 1997; Zikmund-Fisher, Witteman et al., 2014; Wangeamsemsuk & Jiamsanguanwong, 2018), the number of icons (denominator) to be used (Garcia-Retamero et al., 2010; Zipkin et al., 2014), and whether the displays should be static or dynamic (e.g., Okan, Garcia-Retamero, Cokely, & Maldonado, 2015; Zikmund-Fisher et al., 2012) or even interactive (e.g., Ancker, Weber, & Kufafka, 2011; Zikmund-Fisher, Dickson, & Witteman, 2011; Zikmund-Fisher et al., 2012). Another potentially important aspect of icon array design, namely its organization or layout, has been relatively underexplored. Although researchers have examined icons that are scattered versus those that are grouped (e.g., Ancker et al., 2011; Feldman-Stewart, Brundage, & Zotov, 2007), they have not examined different groupings, as we do.

Icon array organization is important because it can help or hinder information processing. One study found that people with high and low numeracy processed the information in icon arrays (where icons were grouped) differently (Kreuzmair et al., 2016). Specifically, high numerates counted icons while low numerates used large-area processing. The more effective (counting) strategy, however, may be adversely affected by whether icons are grouped or

organized by test result or by actual disease outcome. The present research aims to contribute to the understanding of effective icon array organization.

One way that an intervention such as an icon array can help people solve a Bayesian problem is to direct them towards solving an easier equation. Indeed, Bayesian problems can be solved using simpler calculations (Equations 3-4) which do not require knowing the base rate or probabilities of various events. Instead, Equations 3-4 require calculating the ratio between two counts. When a test result is positive, the relevant components for the ratio in Equation 3 are the number of people with the disease who get a positive result and the number of positive results. When a test result is negative, the relevant components are the number of people with the disease who get a negative result and the number of negative results (Equation 4).

$$PPV = P(Disease | Positive) = \frac{n(Disease + Positive)}{n(Positive)} \quad [3]$$

$$NPV = P(Disease | Negative) = \frac{n(Disease + Negative)}{n(Negative)} \quad [4]$$

We argue that a visual decision aid could help a person solve Equations 3-4 by positioning the relevant components in close spatial proximity to each other (see Figure 1a). The visual system is highly tuned to proximity, and uses this Gestalt principle to group together elements (Wagemans et al., 2012). Proximity in the icon array could help the visual system spontaneously group and contrast the relevant components required to solve the Bayesian problem. When the test result is positive, the PPV can be found by computing the ratio of the number of red icons to the number of red-and-gray icons on the left side of the icon array shown in Figure 1a (and ignoring information in the right side of this icon array). The red and gray

icons are located near each other, so proximity could help the visual system to group these icons together, and this grouping could help guide people to use the ratio of red icons to red-and-gray icons as an estimate of risk. Given that the gray icons start at the location where the red icons end, grouping may also be boosted by other Gestalt principles such as good continuation. When the test result is negative, the NPV can be calculated as the ratio of the number of red icons to the number of red-and-gray icons on the right side of the icon array.

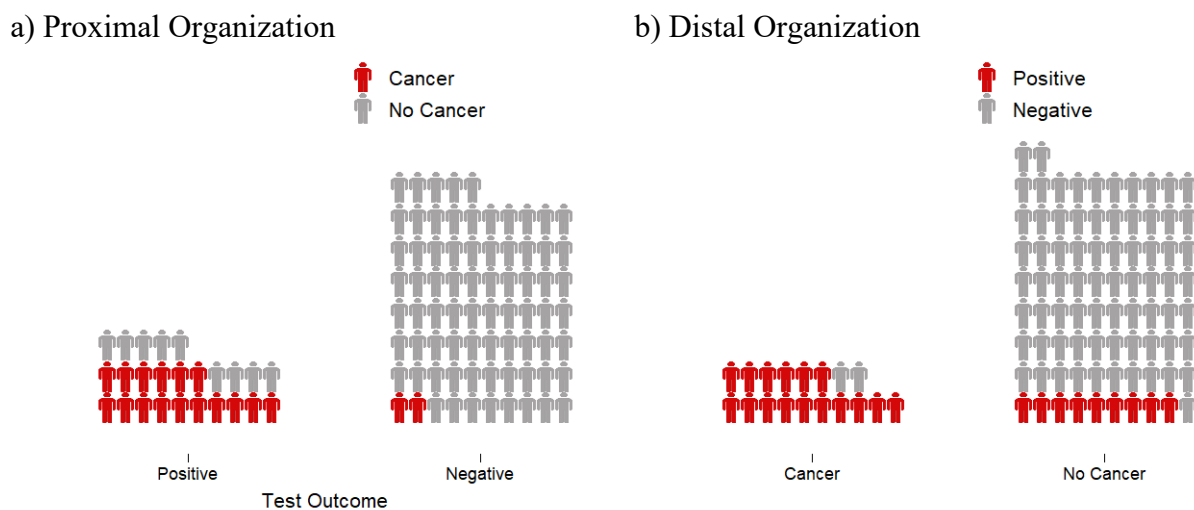


Figure 1. Example showing the proximal (a) and distal (b) organization of the icon array.

Conversely, a visual decision aid could hinder the use of Equations 3-4 by positioning the relevant components at a farther distance from one another (see Figure 1b). In this case, using Equation 3 to calculate the PPV requires a visual comparison of the red icons on the left to the sum of all red icons in the case of a positive result. Using Equation 4 to calculate the NPV requires a visual comparison of the gray icons on the left to the sum of all gray icons. However, two clusters of red icons are so far apart that they do not, to our eyes, seem to be perceptually grouped. Instead, this organization might lead a person to also compare the number of red icons to the number of red-and-gray icons on each side of the icon array. These ratios would involve a

comparison of irrelevant components because their resulting values are unrelated to PPV and NPV.

Hypotheses

H1. People will be more accurate at estimating risk in the proximal condition than in the distal condition.

H2. The higher accuracy in the proximal condition will be due to better precision at estimating risk in the proximal condition than in the distal condition.

H3. The better precision in the proximal condition will be due to participants calculating the ratio of relevant components as specified in Equations 3-4 when in the proximal condition compared to the distal condition.

Accuracy is a commonly used metric to assess performance in Bayesian reasoning tasks. However, accuracy is limited as a metric because there are multiple reasons for inaccurate responses. Reduced accuracy could arise because people are biased in their estimate of the risk or because people are less precise in their judgments of the risk. This distinction between precision and bias has been argued as a better way to characterize performance on Bayesian reasoning tasks (Witt, 2020). Measurement of precision and bias, rather than simply accuracy, requires multiple testing (see Figure 2). Precision is quantified as the slope from a regression model mapping estimated risk onto the actual risk given a test result. Steeper slopes indicate better precision while shallower slopes indicate attenuated precision.

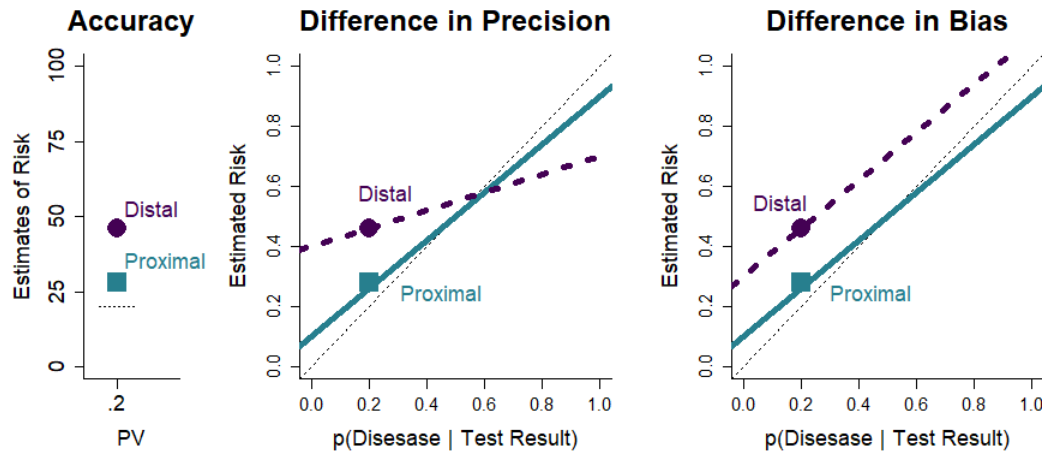


Figure 2. Hypothetical data illustrating estimated risk as a function of the predictive value of the test when only a single scenario is tested (left panel) versus when multiple scenarios are tested (middle and right panels). When only one scenario is used, only accuracy can be assessed, in which case, the proximal condition was better. When multiple scenarios are tested, we can calculate differences in precision (middle panel) or bias (left panel) to further understand what makes the proximal condition better.

We predicted people would be more accurate in the proximal condition (H1) and that they would be more accurate because they would have greater precision in the proximal condition than for the distal condition (H2). To assess the reason for increased precision, we predicted people would use the ratio of proximally-located icons. To assess this prediction, we fit statistical models using one predictor that corresponds to the ratio of red to red-and-gray icons in the proximal condition, and another predictor that corresponds to the ratio of red to red-and-gray icons in the distal condition. Model fit was assessed using the Bayesian information criterion (BIC). We hypothesized that responses in each condition would be best predicted by the ratio of red to red-and-gray icons, and this is why precision would be better when this ratio aligns with the depicted risk, as in the case of the proximal condition, because the proximal condition aligns with Equations 3-4 (H3). We test these hypotheses in two experiments.

Experiment 1

In Experiment 1, we tested whether the organization of an icon array influences performance on a Bayesian reasoning task. We hypothesized that the organization that is more compatible with Equations 3-4 (the proximal condition) would lead to better performance. Specifically, we hypothesized that the proximal organization would lead to greater accuracy (H1) because it would lead to greater precision (H2) in judging one's risk for having a disease given a result on a diagnostic test. We also tested whether the ratios based on visual groupings within each icon array were responsible for greater precision (H3).

Method

Design. In a within-subjects design, we manipulated the main the factor of interest i.e., icon array organization. In the proximal condition, icons were arranged by diagnostic test result (positive v. negative) whereas in the distal condition they were arranged by actual disease outcome (cancer present v. cancer absent). The two organizations are shown in Figure 1.

We also systematically varied the disease base rate (0.18 v. 0.35), diagnostic test sensitivity (low, medium, high), diagnostic test bias (negative, positive, no bias), and diagnostic test result (positive v. negative). These systematic variations produced a set of scenarios with a wide range of PPVs and NPVs (see Appendix). This range enabled us to assess how perceived risk varied as a function of actual risk based on the information within the scenario, thereby enabling measurement of precision.

Participants. The participants were 28 students who volunteered in exchange for course credit in an introductory psychology course. This sample size leads to 80% power to find a medium ($d = .55$) effect size for a two-tailed paired-sample t-test. Participants gave informed consent, and the research was approved by the Colorado State University Institutional Review Board (CSU IRB).

Stimuli. The stimuli consisted of icon arrays that depicted information about a diagnostic test for cancer. Each array showed the number out of 100 people who had a true positive test result, a true false positive test result, a true negative test result, and a true false negative test result (see shaded columns in Table A1 in Appendix). The sum of the true positives and true negatives divided by the total sample provides base rate information. Test sensitivity can be deduced by the overall number of errors made by the test (i.e., the number of false positives and false negatives). Finally, test bias is indicated by the relative frequency of different sorts of errors (i.e., false positives versus false negatives). We factorially combined disease base rate, test sensitivity and test bias to create 18 combinations. Each combination was then paired with test result (positive v. negative) for a total of 36 unique scenarios.

The format of the icon arrays was manipulated to create two versions for each scenario thus producing 72 unique displays. For the proximal condition, the icons were arranged by diagnostic test result (positive or negative), and color indicated actual presence of the disease (see Figure 1a). For the distal condition, the icons were arranged by actual disease outcome (presence or absence), and color indicated a positive test result (see Figure 1b). The icon arrays were presented on 19" color monitors.

Procedure. On each trial, participants were shown an icon array and told whether their result was positive or negative (see Figure 3). They then rated their risk for having the disease on a scale from 0 to 10. This response was provided by adjusting a slider. The slider values were in one decimal point, and the numeric value of the slider was presented above the slider. The starting position of the slider was 5. Participants could make multiple adjustments and there was no time limit. When they finished adjusting the slider, participants clicked on a button that said "next" to move to the next trial.

Participants completed two blocks of trials, one with each organization. A block of trials consisted of 36 scenarios, and each scenario was presented twice (i.e., once for each organization) for a total of 72 trials per block. Scenario order within block was randomized. Organization order was randomized across participants.

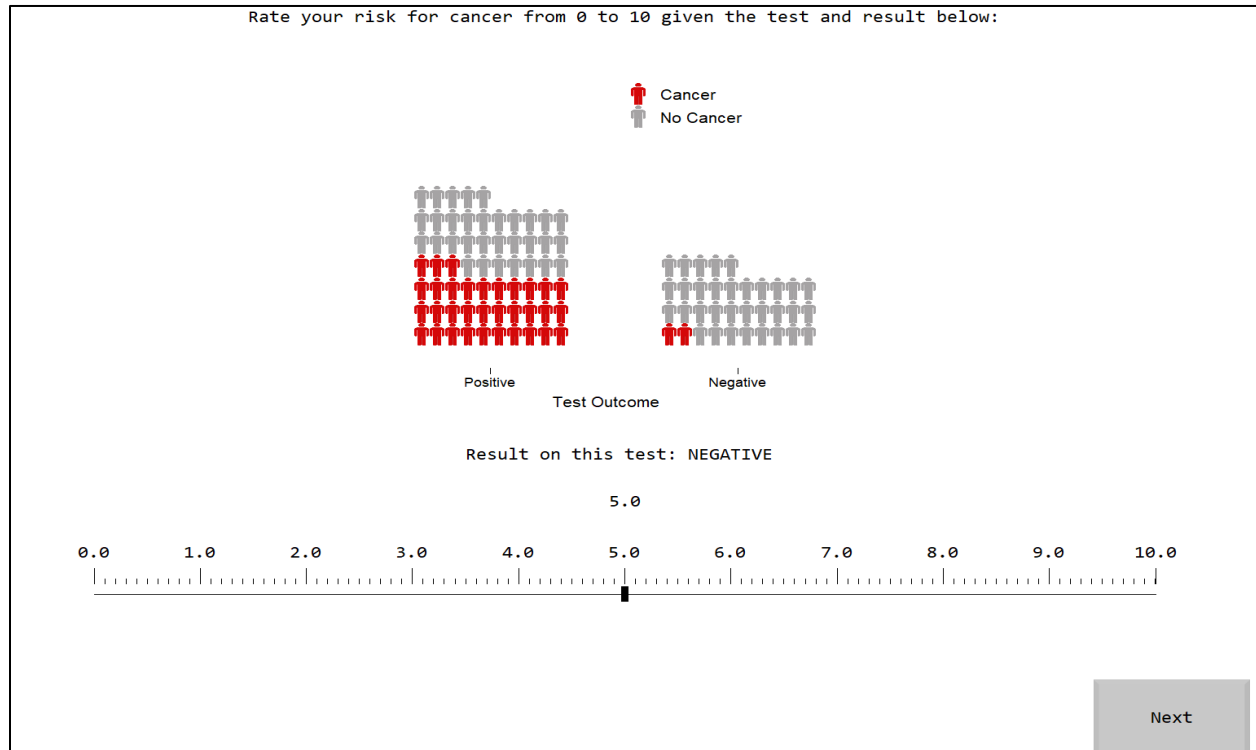


Figure 3. An example trial from Experiment 1 (in the proximal condition).

Analyses and Results

Accuracy. The proportion of correct estimates was calculated for each participant for each icon array organization. Estimates were divided by 10 to put them on a 0-1 scale. An estimate was deemed correct if it was within 5% of the predictive value (PV), which is the probability of having the disease given the result of the test. The PV was equal to the PPV when the test result was positive and the NPV when the test result was negative. We ran a paired-

samples t-test to compare accuracy with the proximal icon arrays versus the distal icon arrays. Accuracy was considerably higher with the proximal icon arrays ($M = 35\%$, $SD = 19\%$) than with the distal icon arrays ($M = 14\%$, $SD = 10\%$), $t = 6.85$, $p < .001$, $df = 27$, $d = 1.29$.

Precision. Accuracy could be lower in the distal condition because people are more biased in estimating risk or because people are less precise at estimating risk than the proximal condition. To differentiate these possibilities, we analyzed the data to assess precision as described above (Witt, 2020). Precision and bias are measured as the slopes and intercepts, respectively, from linear mixed models. The dependent measure was estimated risk. The fixed effects were PV (which was mean-centered by subtracting 0.5), icon array organization and the interaction between icon array organization and PV. Random effects for participant were included in the model. The random effects included intercepts and slopes for each fixed effect.

To test H2, we compared the slopes across the two conditions (see Figure 4). The interaction between disease probability and condition was significant, estimate = -3.93, $SE = 0.70$, $t = -5.61$, $df = 27$, $p < .001$. Consistent with H2, precision in estimating risk was substantially increased in the proximal condition compared with the distal condition. Precision in the proximal condition was 6.77, $SE = 0.65$, $t = 10.34$, $df = 27$, $p < .001$. Perfect precision would be 10, and chance performance (no precision) would be 0. Using this number as the baseline, we calculated the percent increase in precision in the proximal condition, which was a 58% increase ($3.93 / 6.77 * 100\%$).

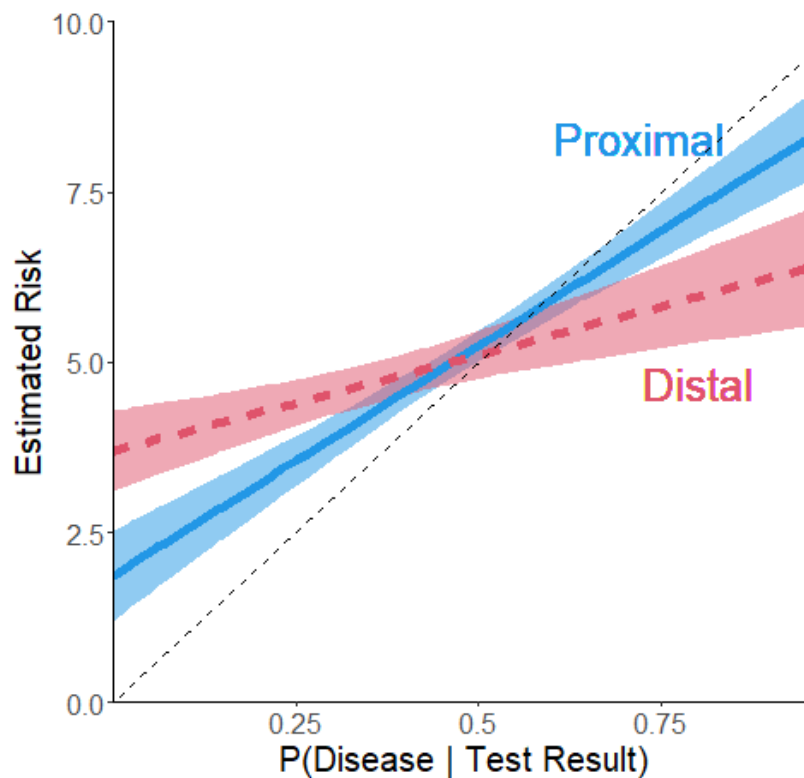


Figure 4. Estimated risk as a function of the probability of having the disease according to the information in the scenario and icon array organization for Experiment 1. Shading represents 95% CIs. The dashed line represents ideal performance.

Relevant and Irrelevant Ratios. The results show that the organization of the icon array is critical for its effectiveness at communicating information in a Bayesian reasoning task. According to H3, the proximal organization is the superior format because it puts the relevant parts of the ratio in close proximity. When a person receives a positive result, the relevant ratio is the number of true positives to the number of positive results. When a person receives a negative result, the relevant ratio is the number of false negatives to the number of negative results. In the proximal condition, the icons are arranged such that each side of the ratio are closer to each other. By contrast, in the distal condition, the relevant sides of the ratios are further apart. This distance could lead to worse precision in judging risk.

Alternatively, the distal condition may lead to perceptual grouping that emphasizes an irrelevant ratio: If participants use the ratio of red to red-and-gray icons on one side of the icon array to judge risk, this ratio is only partially correlated with the probability of having the disease ($r = .57$, linear regression slope coefficient = 0.50), so risk judgments would be far less precise if this irrelevant ratio were used to estimate risk. The visual grouping likely to occur due to proximity and other Gestalt grouping principles such as good continuation in the proximal condition was hypothesized to both facilitate identification of the relevant visual comparison and make the comparison easier to perform (H3).

We tested H3 by modeling whether responses were a function of the ratios of red to red-and-gray icons presented in the icon arrays for each organization. Predictor 1 corresponded to the ratio of red icons to red-and-gray icons on each side of the icon arrays in the proximal condition. Predictor 1 (hereafter also called the relevant ratio) was equal to the number of true positives divided by the number of positive results when the diagnostic test result was positive (64% in Figure 1a) and was equal to the number of false negatives divided by the number of negative results when the diagnostic test result was negative (3% in Figure 1a). Predictor 2 (also called the irrelevant ratio) corresponded to the ratio of red icons to red-and-gray icons on each side of the icon arrays in the distal condition. Predictor 2 was equal to the number of true positives divided by the number of people who had the disease when the diagnostic test result was positive (89% in Figure 1b) and was equal to the number of false positives divided by the number of people who did not have the disease when the diagnostic test result was negative (11% in Figure 1b).

For each participant in each array organization, responses were analyzed with two models, one with Predictor 1 (relevant ratio) and one with Predictor 2 (irrelevant ratio) as the independent measure. We extracted the Bayesian information criterion (BIC) for each model.

Lower BIC indicates better model fit and BIC scores within 2 of each other suggest equivalent models. Three participants had BIC scores within 2 of each other and thus could not be classified as being better fit by either predictor. Excluding these participants, the proportion of participants whose data were better fit by Predictor 1 (relevant ratio) than Predictor 2 (irrelevant ratio) was substantially greater for the proximal condition than the distal condition (92% vs 27%, respectively, $\chi^2 = 20.45$, $df = 1$, $p < .001$).

For illustration, we conducted a linear mixed model with estimated risk as the dependent variable. The fixed effects were Predictor 1, Predictor 2, condition, and interactions with condition and each predictor. Random effects for participant included intercepts and slopes for each fixed effect. Predictor 1 was a better predictor of participants' responses in the proximal condition, and Predictor 2 was a better predictor of participants' responses in the distal condition (see Figure 5). Rather than processing the correct ratio for both conditions but simply being more precise when each side of the ratio was closer together in the icon array, the results suggest that participants were processing different ratios based on the icon array organization. This is consistent with the idea that participants calculated the proportion of red icons to red-and-gray icons, and the organization of icons by test result encouraged the appropriate calculation.

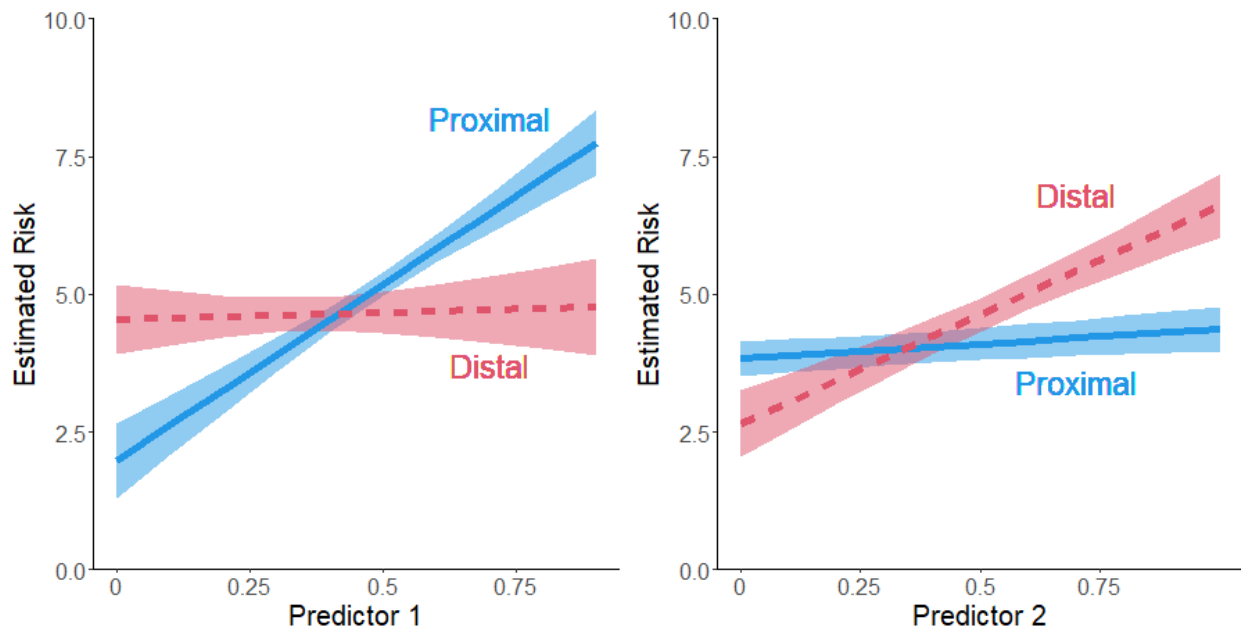


Figure 5. Estimated risk as a function of condition and Predictor 1 (relevant ratio; left) or Predictor 2 (irrelevant ratio; right) for Experiment 1. See text for details on calculation of Predictors 1 and 2. Shading represents 95% CIs calculated from the model.

Order Effects. We also tested whether condition order affected precision given that experience with the proximal organization could later improve precision with the distal organization. A linear mixed model with estimated risk as the dependent measure was computed. The fixed effects were the probability of having the disease given the information in the scenario, the icon array organization, order of presentation, and all two- and three-way interactions. The random effects for participant included intercepts and slopes for each fixed effect that was within-subjects. None of the interactions with order were statistically significant, $ps > .30$. Precision was higher in the proximal condition than in the distal condition for participants who started with the proximal organization and for participants who started with the distal organization, $ps < .004$.

Discussion

The results showed that icon array organization had a large effect on accuracy of risk estimates. Furthermore, the results showed that people were more accurate with the proximal organization because they had greater precision than with the distal organization. To account for the higher precision, we found people were more likely to use the relevant ratio (as specified in Equations 3-4) when viewing the proximal icon arrays and to use an irrelevant ratio when viewing the distal icon arrays.

Experiment 2

The primary purpose of Experiment 2 was to replicate Experiment 1 using a larger sample of participants, and with participants judging the risk for an unknown other person rather than themselves, as well as participants providing responses on a 0-100 (rather than 0-10) scale. In addition, we explored individual differences in the effectiveness of the two different icon array organizations. Specifically, we focused on individual differences in numeracy. It has been suggested that icon arrays are particularly useful for individuals with low numeracy (Galesic et al., 2009). We hypothesized that the proximal organization of icon arrays would be particularly beneficial for people lower in numeracy compared with the distal organization.

Method

Participants. The participants were 62 students who volunteered in exchange for course credit. This sample size leads to 80% power to find a small effect size ($d = .36$). Participants provided informed consent. The experiment was approved by the CSU IRB.

Design and Stimuli. We used the same scenarios (icon arrays) as in Experiment 1. In addition, we included the 4-item multiple-choice Berlin numeracy scale (Cokely et al., 2012).

Procedure. The experiment was administered as a Qualtrics survey which participants completed on their own device. The survey started with three specific Bayesian problems, which

were part of a separate line of research inquiry and so will not be discussed further. Then, participants completed the numeracy scale.

Participants then completed the main trials as per Experiment 1. On each trial, an icon array was presented along with a test result (positive or negative). Participants estimated how many people out of a 100 with that result on that test would have cancer (see Figure 6). They positioned a slider which ranged from 0–100 (no decimal places), then clicked to continue onto the next trial. Participants each completed 36 trials (one for each scenario) under one icon organization condition before completing 36 trials under the other condition (72 trials in total). Order of scenarios was randomized within participant. Starting condition was randomized across participants.

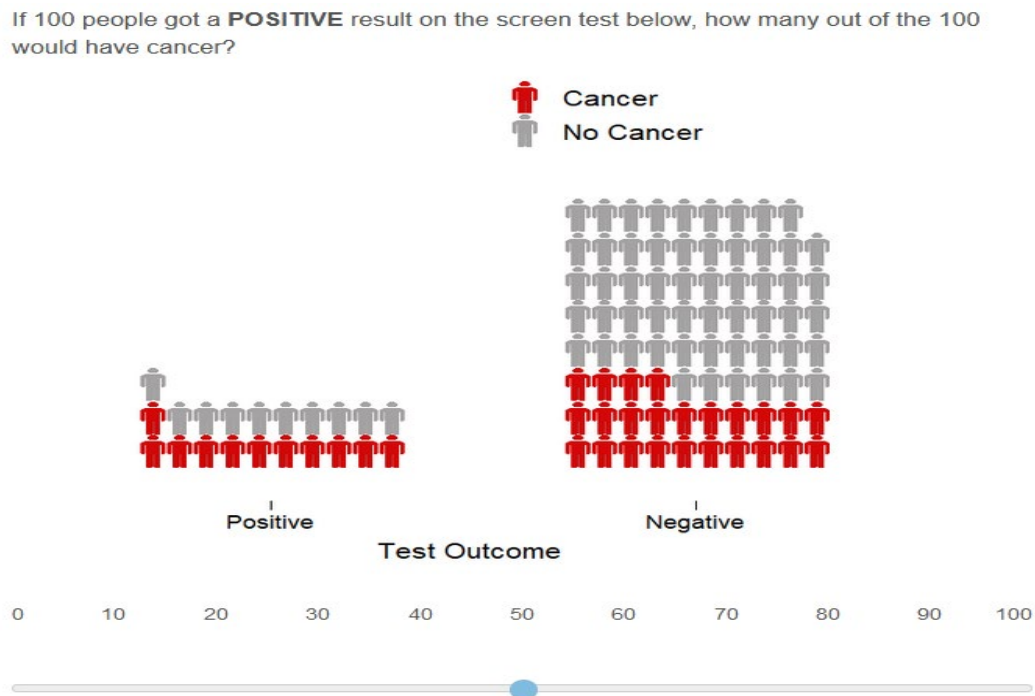


Figure 6. A sample trial display from the proximal condition in Experiment 2.

Results

The data analysis plan followed that for Experiment 1.

Accuracy. The proportion of correct responses were calculated for each participant for each condition. Participants estimated risk correctly more often in the proximal condition than in the distal condition, $t = 3.56$, $df = 60$, $p < .001$, $d = 0.46$ (Proximal: $M = 44\%$, $SD = 32\%$; Distal: $M = 31\%$, $SD = 30\%$).

Precision. To test whether the better accuracy in the proximal condition was due to better precision, we compared the slopes across the two conditions (see Figure 7). The interaction between disease probability and icon array organization was significant, estimate = -17.01, $SE = 3.98$, $t = -4.28$, $df = 60.28$, $p < .001$. Precision in estimating risk increased by 35% in the proximal condition compared with the distal condition. This result is consistent with H2. We used the random effect coefficients to assess differences in the precision across the two organizations for each participant and found that 77% (48 out of 62 participants) had greater precision with the proximal than the distal organization.

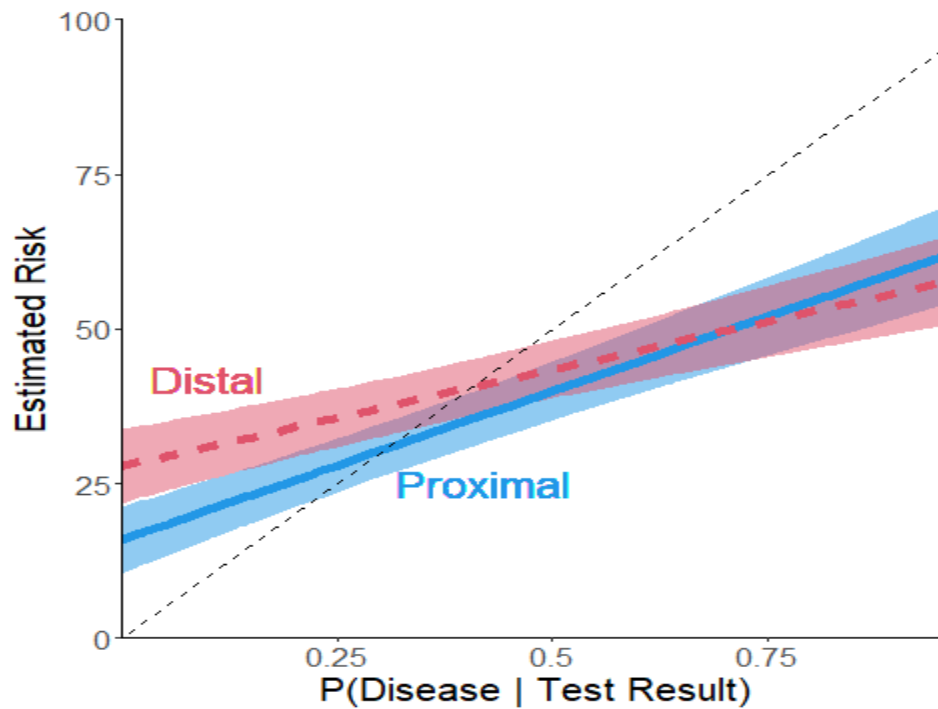


Figure 7. Estimated risk as a function of the probability of having the disease according to the information in the scenario and icon array organization for Experiment 2. Shading represents 95% CIs.

Relevant and Irrelevant Ratios. To test H3, we examined whether the ratios based on proximity in the two conditions best explained participants' responses. Predictors 1 (relevant ratio) and 2 (irrelevant ratio) were calculated as in Experiment 1 to align with the ratio of red to red-and-gray icons on each side of each icon array. Participants were more likely to respond based on the relevant ratio than the irrelevant ratio in the proximal condition compared with the distal condition (78% versus 56%, respectively, $\chi^2 = 7.17$, $df = 1$, $p = .007$). The result is consistent with H3. This pattern also matched the outcomes from a linear mixed model with response as the dependent variable, fixed effects being condition, Predictors 1 and 2, and their interaction with condition, and random effects for participant as well as intercepts and slopes for each fixed effect (see Figure 8).

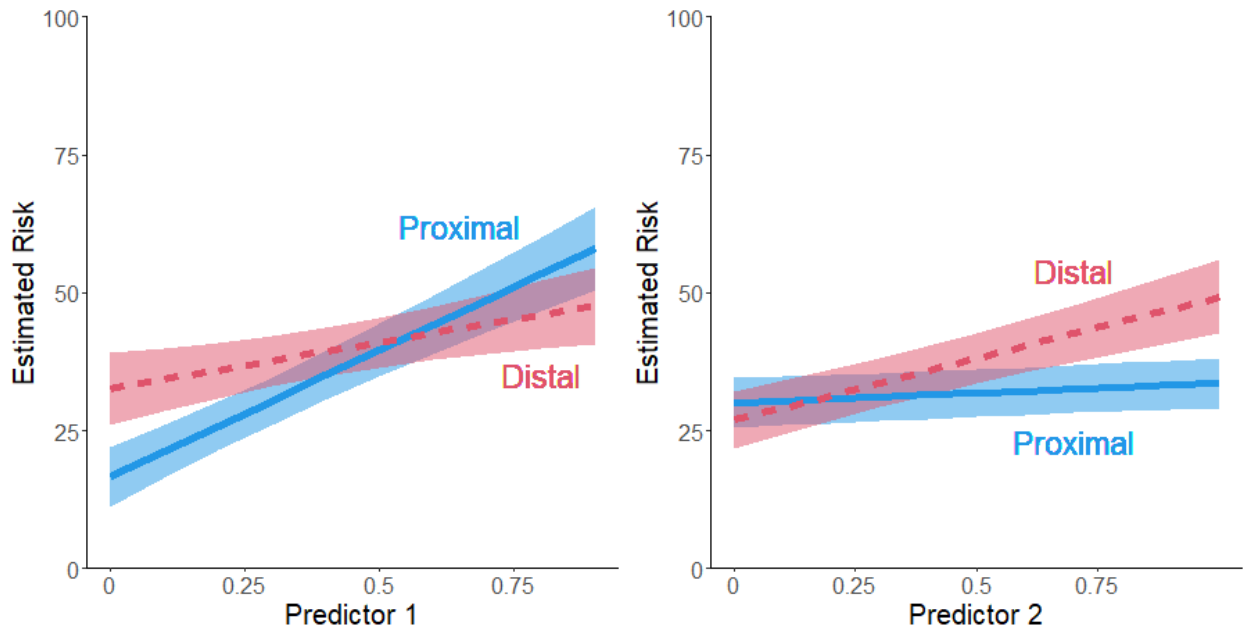


Figure 8. Estimated risk as a function of condition and Predictor 1 (relevant ratio; left) or Predictor 2 (relevant ratio; right) for Experiment 2. See text for details on calculation of Predictors 1 and 2. Shading represents 95% CIs calculated from the model.

Order Effects. We explored the effect of order of presentation on precision by including order and all interactions in the linear mixed model. The three-way interaction with PV, condition, and order was non-significant, $p = .65$. The difference in precision across conditions was similar for participants who started with the proximal organization, estimate = 15.20, $SE = 5.62$, $z = 2.70$, $p = .007$, and for participants who started with the distal organization, estimate = 18.80, $SE = 5.70$, $z = 3.30$, $p = .001$.

Numeracy. In order to measure individual differences in the effectiveness of the two icon array organizations, we calculated the proportion of correct responses on the numeracy task ($M = .32$, Median = .25), and then re-ran the linear mixed model and included numeracy and all two-way and three-way interactions as fixed effects. Condition was entered as a factor with the distal

organization as the reference category. The three-way interaction between PV, condition, and numeracy was significant, estimate = -34.72, $SE = 16.11$, $t = -2.16$, $p = .035$ (see Figures 9-10).

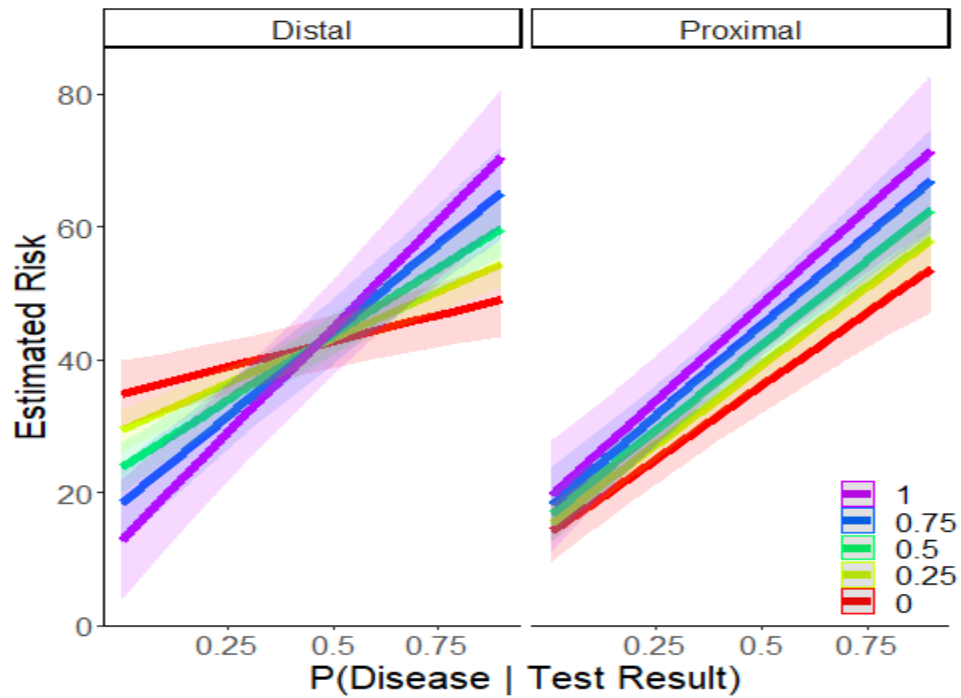


Figure 9. Estimated risk as a function of the probability of having the disease according to the information in the scenario, icon array organization, and proportion of correct responses on the numeracy task for Experiment 2. Shading represents 1 SEM (standard error of the mean).

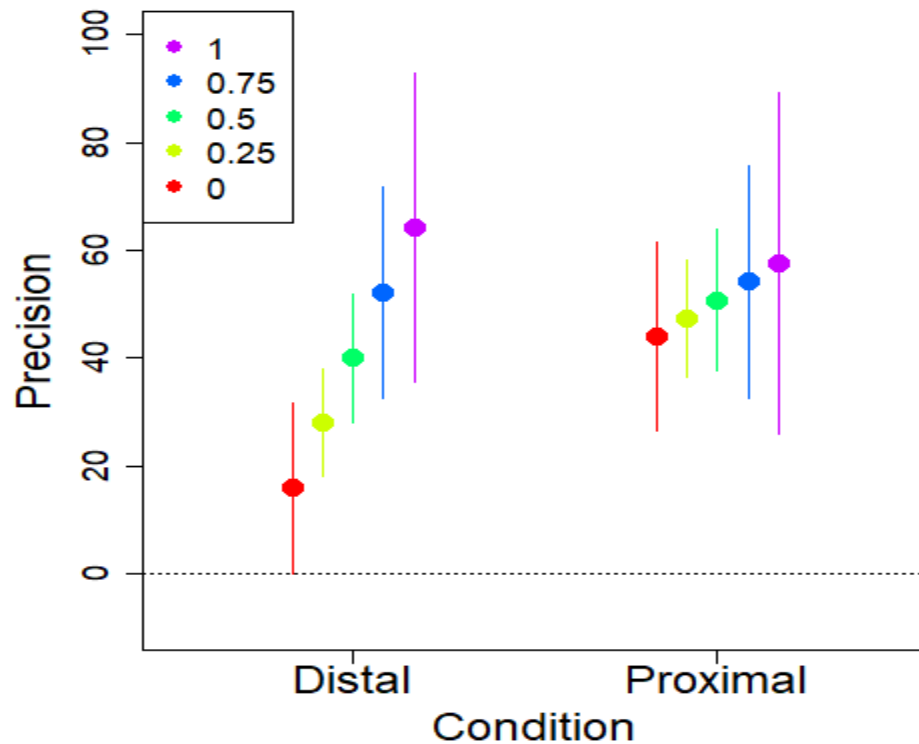


Figure 10. Mean precision scores in Experiment 2 as a function of condition and numeracy (calculated as proportion of correct responses from 0 to 1; higher scores indicate higher numeracy). Error bars represent 95% CIs as calculated by the model.

For participants higher in numeracy (i.e., those who answered 3 or 4 out of 4 questions correctly; blue and purple in Figures 9-10), the icon array organization did not significantly affect their precision, as shown by a lack of significant interaction between disease probability and condition, estimate = 2.30, $SE = 4.12$, $z = 0.61$, $p = .54$. In contrast, for participants lower in numeracy (i.e., those who answered 0 or 1 question correctly; red and yellow in Figures 8-9), the proximal organization increased their precision over 100% compared with the distal organization, estimate = 23.70, $SE = 4.98$, $z = 4.77$, $p < .001$. Indeed, the proximal condition increased precision for people with low numeracy so much that the statistical significance of the interaction between disease probability and numeracy for the proximal condition was non-significant, estimate = 13.60, $SE = 22.40$, $z = 0.61$, $p = .54$. Although non-significant, the most

numerate individuals were 31% more precise than the least numerate individuals, and although the proximal condition hugely boosted the precision of those with lower numeracy, precision was still not as high as achieved by those highest in numeracy.

Exploring Bias Across Both Experiments

The PV of a diagnostic test is a combination of several factors including the prevalence of the disease, the quality of the diagnostic test, and biases within the diagnostic test. To accurately estimate risk, one must consider all these factors and consider them in combination with one another. For example, a positive result is a better indicator that a person has the disease when the test quality is high or when the test is biased to give negative results whereas a positive result is less indicative of disease when the test quality is low or the test is biased to give positive results.

We explored whether the icon array organizations led participants to differ in the weighting of these factors and their combinations. The data across the experiments was first combined (risk estimates from Experiment 1 were multiple by 10 so they were on a scale from 0-100 as per Experiment 2). We quantified weighting as the coefficients from linear mixed models. The dependent measure was estimated risk. The fixed effects were test result, diagnostic test quality, diagnostic test bias, base rate of the disease, and all two-way interactions. For interpretability, we only included two-way interactions. Test result (negative, positive) was coded as a factor. Diagnostic test quality (low, medium, high) was coded as 0, 1, 2. Diagnostic test bias (bias to give positive results, no bias, and bias to give negative results) was coded as -.75, 0, .75. Base rates (18%, 35%) was coded as a factor. We ran one model for estimates made with the distal icon arrays and another model for estimates made with the proximal icon arrays. To calculate ideal weightings, we ran the model with PV as the dependent measure and extracted

the coefficients. The results are shown in Figure 11. The lengths of the lines indicate the deviation from the ideal coefficient, and differences in line lengths across the conditions signify differences in the coefficients between the distal and proximal conditions.

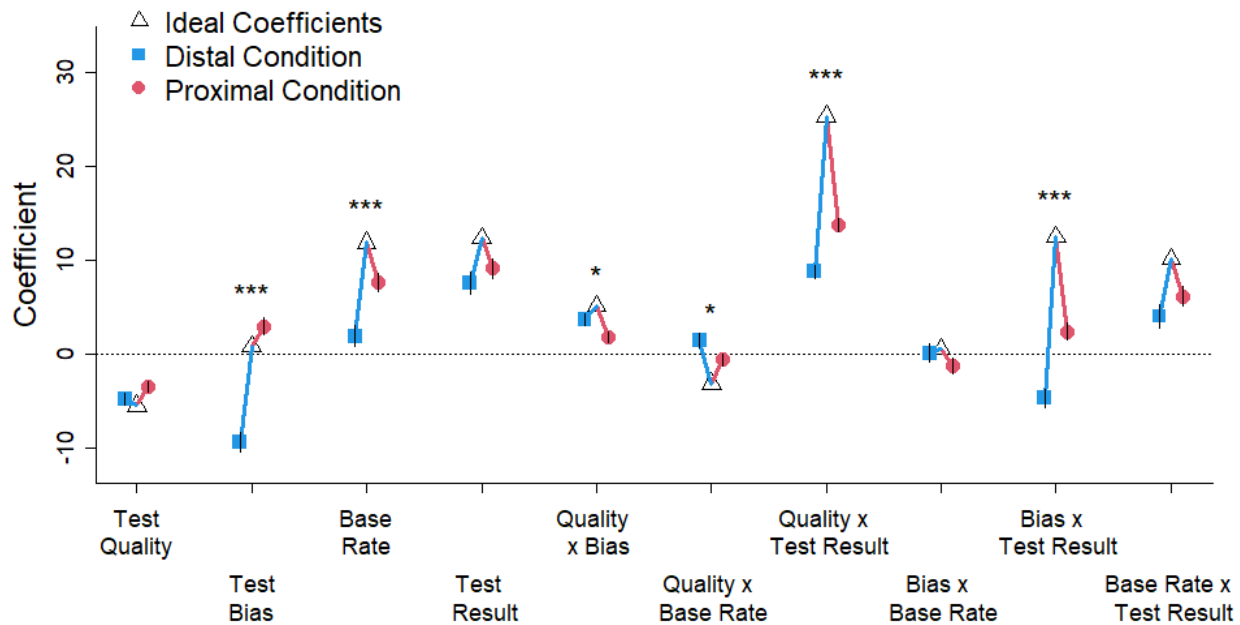


Figure 11. Coefficients from mixed linear regressions are shown as a function of the fixed effect (x-axis). Black open symbols correspond to ideal coefficients. Blue square symbols correspond to coefficients from responses on the distal organization. Red circle symbols correspond to coefficients from responses on the proximal organization. Red and blue lines highlight the deviation of the coefficients from each condition from the ideal coefficient, so longer lines indicate greater bias. Coefficients below the ideal indicate a bias to under-weight the information, and coefficients above the ideal indicate a bias to over-weight the information. Error bars are 1 SEM. When the coefficients from the two conditions are significantly different from each other, asterisks indicate the level of statistical significance (* $p < .05$; *** $p < .001$).

We highlight a few results here. There was a significant difference in the coefficients for base rate. As shown in Figure 11, participants under-weighted information about base rate, and they under-weighted this information more in the distal condition than in the proximal condition (both symbols below the triangle and the blue square is lower than the red circle).

With respect to test bias, the ideal coefficient is close to 0, which makes sense because the bias of the test in and of itself has no implication for whether a person has a disease. Yet participants were biased by this information. In the distal condition, the coefficient was negative, which means that as the test was biased to produce positive results, participants estimated risk as higher. In the proximal condition, the bias was much less than in the distal condition and was in the opposite direction: risk was estimated as higher when the diagnostic test was biased to produce negative results.

Test bias has large implications for PV when combined with information about the test result. The ideal coefficient for Bias x Test Result was nearly 12.5. A positive test is more indicative of disease when the test is biased to produce negative results than when the test is biased to produce positive results. However, participants did not use the integration of this information as much as they should have, and in the distal condition, their coefficient was negative. The negative coefficient means participants were more likely to estimate risk as higher when the test result was positive when the test was biased to produce positive results than when it was biased to produce negative results.

As another example of a failure to integrate different pieces of information, test outcome also has a larger impact on PV when the test quality is high than when it is low (the ideal coefficient for Quality x Test Result is 25). Coefficients were in the right direction but were far less than ideal for both conditions, but the coefficient was higher (better) in the proximal condition than in the distal condition.

With respect to test result, participants underweighted this piece of information in both conditions, and the difference between the conditions was not statistically significant, $p = .32$.

General Discussion

Icon arrays are considered an effective means for communicating medical risks, especially for those with lower levels of numeracy (e.g., see Zipkin et al., 2014). Beyond demonstrating the effectiveness of this form of communication relative to other forms such as numbers, prior research has also focused on identifying the most beneficial aspects of icon array design, including whether icons should be scattered or grouped (Ancker et al., 2006; Feldman-Stewart et al., 2007). The present research tackled another design-related question, namely the most effective grouping or organization of the icons. We tested whether organizing the icons by the diagnostic test result (positive v. negative) would lead to superior performance in terms of estimated risk than organizing the icons by actual disease outcome (presence v. absence).

When a diagnostic test can lead to false positives or negatives, estimating risk can be done through Bayes theorem (Equations 1-2). However, these Bayesian problems can be reduced to a simple ratio (see Equations 3-4). Here, we tested which icon array organization was better for guiding the visual system to spontaneously compute the simple ratio. One icon array organization grouped the relevant components in close proximity, then used color to differentiate the various parts of the ratio (proximal condition, see Figure 1a). The other icon array organization grouped the relevant components using the same color, then used spatial distance to differentiate the various parts (distal condition, see Figure 1b). While both proximity and color can lead to perceptual grouping (Wagemans et al., 2012), we found that the organization in the proximal condition was better for guiding people to solve a Bayesian reasoning task. The proximal icon array organization appeared to encourage the calculation of the relevant ratio, which would increase performance on the Bayesian reasoning task compared to the distal icon array organization for which the relevant icons were positioned further apart from each other and in a way that did not make it easy to see which parts of the array were relevant to compare.

The data from two experiments supported our hypothesis – participants demonstrated significantly greater precision and less bias in estimates of risk when in the proximal condition than when in the distal condition. Furthermore, the proximal condition led participants to the use of the relevant ratio (red to red-and-gray icons) whereas the distal condition led to increased use of the irrelevant ratio (red to red-and-gray icons). This suggests that icon array organization can help participants attend to and process the relevant ratio as specified in Equations 3-4.

Errors in Bayesian reasoning occur partly because people ignore (or underweight) some of the relevant information (e.g., Bar-Hillel, 1980), and the effectiveness of icon arrays is partly due to their ability to make all information explicitly available for processing. Our research suggests that processing is made easier when the relevant information is organized in close proximity to each other. Research in the field of judgment and decision-making has often depicted people as ‘cognitive misers’ in an effort to describe and explain why they may prefer to use intuitive heuristic strategies over more rational thought (see Kahnemann, 2011). The proximal organization of icon arrays leverages the desire that people may have to take mental shortcuts and helps them to successfully perform a Bayesian reasoning task in a less cognitively effortful way.

Beyond restructuring the task and making it more transparent so as to signal the appropriate ratios for comparison, researchers have recently argued that visual displays better match the information being presented with that being requested compared to so-called text-based or word problems (e.g., Tubau, 2021). In increasing this match, it has been shown that performance on Bayesian reasoning tasks can be improved (e.g., Johnson & Tubau, 2017; Tubau, Rodríguez-Ferreiro, Barberia, & Colomé, 2019; see also Talboy & Schneider, 2018). While the present research did not aim to directly contribute to the debate on the alignment of problem and

question, one might argue that there is greater alignment in the proximal than distal condition. Future research could explore this possibility.

The fact that icon arrays appear to be a particularly advantageous way of communicating risks to people with low levels of numeracy (Galesic et al., 2009; Garcia-Retamero & Galesic, 2010) means that any innovations in array design ought to at least retain benefits for this subgroup, if not further enhance the benefits. We found that the proximal organization provided a sizeable boost to the performance of individuals with lower numeracy, although their performance remained lower than that of their numerate counterparts. While others have relied on instructional manipulations to try to improve the performance of low numerates (e.g., by asking them to count icons; Kreuzmair et al., 2016), the approach of reorganizing icon arrays considered in the present research circumvents this need to put the onus on individuals to be both motivated to change their cognitive behavior and to be capable of doing so. Simply reorganizing the information presented to them could result in far reaching improvements in risk comprehension.

Indeed, the superiority of the proximal organization over the distal organization was demonstrated both when participants judged risk for themselves (Exp 1) and when they judged risk for other people (Exp 2). This suggests that different user groups (e.g., clinicians advising others and patients making decisions for themselves) can also benefit from icon array reorganization.

Limitations

While the present research supports the use of the proximal organization for icon arrays, several limitations must be noted. Both experiments involved convenience samples of University students, who are likely to be higher educated as well as possess other differences that

may be relevant for Bayesian reasoning ability. In addition, the present research employed a within-subjects design, so participants had multiple opportunities to practice making risk judgments. While order effects did not account for the difference in precision across the two icon array organizations, practice effects could lead to different estimates compared with some real-world situations where only one estimate of risk is made (e.g., by a patient). The fact that professionals such as clinicians may make several estimates over time suggests that the present research is more akin to those situations.

Another limitation is that when precision is attenuated, this leads to different kinds of biases depending on whether the PPV or NPV is high versus low. Attenuated precision even with no overall bias would, by definition, lead to overestimation at lower levels of risk and underestimation of lower levels of risk. While the particulars of the present stimuli could have had an influence on the results, it is also worth noting that risk perception at very low levels of risk (e.g., $< 1\%$) were not specifically examined in the present research. This level of risk is difficult to communicate non-numerically (Dhimi & Mandel, 2022), and it is debatable whether a different kind of visualization such as a risk ladder might be more effective in these cases.

Finally, we did not systematically manipulate the contrast between the colors or the distance between the two sides of the icon arrays. It is possible that shorter distance coupled with greater cues for grouping by similarity (color, size, shape) could have led to the opposite results with better performance in the distal condition. However, closing the gap between the two sides of the icon array could lead to confusion over which side corresponds to which category whereas the present stimuli align more closely with icon arrays seen in the medical literature (e.g., <https://patientdecisionaid.org/>).

Summary

The present research provides an important, useful and easy-to-implement recommendation for designing icon arrays that aim to communicate risk (for other design features see Garcia-Retamero & Cokely, 2017). The icon arrays should be designed such that the icons are split by test result, with icons corresponding to positive test results on one side and icons corresponding to negative test results on the other side. Within each side, the icons should then be marked by disease such as using colors to designate which icons have disease and which do not. This organization helps transform the cognitively complex Bayesian reasoning task into a simple visual calculation of ratios, thereby improving risk comprehension.

References

- Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, 13(6), 608–618. <https://doi.org/10.1197/jamia.M2115>
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233. [https://doi.org/10.1016/0001-6918\(80\)90046-3](https://doi.org/10.1016/0001-6918(80)90046-3)
- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, 7(1), 25–47.
- Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, <https://doi.org/10.1016/j.tics.2022.03.002>
- Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: Ten steps to better risk communication. *Journal of the National Cancer Institute*, 103(19), 1436–1443.
- Feldman-Stewart, D., Brundage, M. D., & Zotov, V. (2007). Further insight into the perception of quantitative information: Judgments of gist in treatment decisions. *Medical Decision Making*, 27(1), 34–43.
- Gaissmaier, W., Wegwarth, O., Skopec, D., Müller, A., Broschinski, S., & Politi, M. C. (2012). Numbers can be worth a thousand pictures: Individual differences in understanding graphical and numerical representations of health-related information. *Health Psychology*, 31(3), 286.
- Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, 28(2), 210–216.
- Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, 22(5), 392–399.

- Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics. *Human Factors*, 59(4), 582–627.
- Garcia-Retamero, R., & Dhami, M. K. (2011). Pictures speak louder than numbers: On communicating medical risks to immigrants with limited non-native language proficiency. *Health Expectations*, 14, 46–57.
- Garcia-Retamero, R., & Dhami, M. K. (2013). On avoiding framing effects in experienced decision makers. *Quarterly Journal of Experimental Psychology*, 66(4), 829–842.
- Garcia-Retamero, R., & Galesic, M. (2010). Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science and Medicine*, 70(7), 1019–1025.
- Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Medical Decision Making*, 30(6), 672–684.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz M., L., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2), 53–96.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704.
- Jetelina, K. (2021, November 21). Using antigen tests for Thanksgiving week. *Your Local Epidemiologist*. <https://yourlocalepidemiologist.substack.com/p/using-antigen-tests-for-thanksgiving>.
- Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, 6:938. doi: 10.3389/fpsyg.2015.00938

- Johnson, E.D., & Tubau, E. (2017). Structural mapping in statistical word problems: A relational reasoning approach to Bayesian inference. *Psychonomic Bulletin and Review*, 24, 964–97. <https://doi.org/10.3758/s13423-016-1159-6>
- Kahnemann, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Kreuzmair, C., Siegrist, M., & Keller, C. (2016). High numerates count icons and low numerates process large areas in pictographs: Results of an eye-tracking study. *Risk Analysis*, 36(8), 1599–1614.
- Salisbury, H. (2019). Helen Salisbury: Should I persuade patients to have mammograms? *Bmj*, 365.
- Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4, 31–60.
- Stone, E. R., Yates, J. F., & Parker, A. M. (1997). Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied*, 3(4), 243.
- Talboy, A. N., & Schneider, S. L. (2018). Focusing on what matters: Restructuring the presentation of Bayesian reasoning problems. *Journal of Experimental Psychology: Applied*, 24(4), 440-458.
- Tubau, E. (2021). Why can it be so hard to solve Bayesian problems? Moving from number comprehension to relational reasoning demands. *Thinking and Reasoning*, DOI: 10.1080/13546783.2021.2015439.
- Tubau, E., Rodríguez-Ferreiro, J., Barberia, I. & Colomé, A. (2019). From reading numbers to seeing ratios: a benefit of icons for risk comprehension. *Psychological Research*. 83, 1808-1816.

- Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 153–160). MIT Press.
- Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6), 1172.
- Witt, J. K. (2020). The precision-bias distinction for evaluating visual decision aids for risk perception. *Medical Decision Making*, 0272989X20943516.
- Zikmund-Fisher, B. J., Witteman, H. O., Dickson, M., Fuhrel-Forbis, A., Kahn, V. C., Exe, N. L., Valerio, M., Holtzman, L. G., Scherer, L. D., & Fagerlin, A. (2014). Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. *Medical Decision Making*, 34(4), 443–453.
- Zipkin, D. A., Umscheid, C. A., Keating, N. L., Allen, E., Aung, K., Beyth, R., Kaatz, S., Mann, D. M., Sussman, J. B., Korenstein, D., & others. (2014). Evidence-based risk communication: A systematic review. *Annals of Internal Medicine*, 161(4), 270–280.

Appendix

Table A1. The 18 distributions combined with two possible diagnostic test results (positive, negative) to create 36 scenarios and corresponding probabilities of having cancer used in Experiments 1 and 2.

Scenario	Base rate	Sensitivity (d')	Bias (c)	#True+	#False+	#False-	#True-	Positive Predictive Value	Negative Predictive Value
1	0.18	0.5	-0.75	15	57	3	25	0.21	0.11
2	0.35	0.5	-0.75	29	45	6	20	0.39	0.23
3	0.18	0.5	0	11	33	7	49	0.25	0.13
4	0.35	0.5	0	21	26	14	39	0.45	0.26
5	0.18	0.5	0.75	6	13	12	69	0.32	0.15
6	0.35	0.5	0.75	11	10	24	55	0.52	0.3
7	0.18	1.5	-0.75	17	41	1	41	0.29	0.02
8	0.35	1.5	-0.75	33	32	2	33	0.51	0.06
9	0.18	1.5	0	14	19	4	63	0.42	0.06
10	0.35	1.5	0	27	15	8	50	0.64	0.14
11	0.18	1.5	0.75	9	5	9	77	0.64	0.1
12	0.35	1.5	0.75	18	4	17	61	0.82	0.22
13	0.18	2.5	-0.75	18	25	0	57	0.42	0
14	0.35	2.5	-0.75	34	20	1	45	0.63	0.02
15	0.18	2.5	0	16	9	2	73	0.64	0.03
16	0.35	2.5	0	31	7	4	58	0.82	0.06
17	0.18	2.5	0.75	12	2	6	80	0.86	0.07
18	0.35	2.5	0.75	24	1	11	64	0.96	0.15

Notes. + indicates a positive result; - indicates a negative result, so #True+ indicates the number of true positives. The measures of d' and c are standard signal detection theory measures and correspond to test quality and test bias, respectively. Lower d' scores indicate worse test quality. For c , 0 corresponds to no bias, negative scores correspond to result in a bias to produce positive test outcomes, and positive scores result in a bias to produce negative test outcomes. Base rate, d' , and c were systematically manipulated to generate the 18 distributions.