Original Article

MDM Medical Decision Making

Medical Decision Making
2019, Vol. 39(3) 196–207

DOI: 10.1177/0272989X19832882
journals.sagepub.com/home/mdm
SAGE

Aiding Risk Information learning through Simulated Experience (ARISE): A Comparison of the Communication of Screening Test Information in Explicit and Simulated Experience Formats

Pete Wegier, Bonnie A. Armstrong, and Victoria A. Shaffer

Objective. To determine whether the use of Aiding Risk Information learning through Simulated Experience (ARISE) to communicate conditional probabilities about maternal serum screening results for Down syndrome promotes more accurate positive predictive value (PPV) estimates and conceptual understanding of screening, compared with explicitly providing individuals with this information via numerical summary or icon array. ***Method.*** In experiment 1, 582 participants completed an online study in which they were asked to estimate the PPV and rate their attitudes toward a screening test when information was presented in either a description (required calculation of the PPV), explicit (PPV was provided and had to be identified), or an ARISE format (PPV was inferred through experience-based learning). In experiment 2, 316 participants estimated the PPV and rated their attitudes toward screening based on information presented in either an icon array (identify the icons that represent the PPV) or ARISE format. ***Results.*** In experiment 1, ARISE elicited the most accurate PPV estimates compared with the description and explicit formats, and both the explicit and ARISE formats led to more unfavorable attitudes toward screening. In experiment 2, both the icon array and ARISE resulted in similar PPV estimates; however, ARISE led to more negative attitudes toward screening. ***Conclusions.*** These findings suggest that ARISE may be superior to other formats in the communication of PPV information for screening tests. However, differences in the complexity of the formats vary and require further investigation.

Keywords
attitudes toward screening, conditional probabilities, positive predictive value, screening tests, simulated experience

Date received: May 25, 2018; accepted: January 21, 2019

Understanding conditional probabilities—the likelihood that one event will occur given that another event has already occurred—is vital to informed medical decision making, specifically in relation to decisions about screening tests. One such test is maternal serum screening for Down syndrome, which is routinely offered to pregnant women in North America and Europe. Down syndrome is a genetic condition characterized by intellectual disability and physical growth delays, due to the presence of an additional copy of chromosome 21. Maternal serum screening for Down syndrome exists as an option

Temmy Latner Centre for Palliative Care, Sinai Health System, Toronto, ON, Canada (PW); Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada (PW); Department of Family & Community Medicine, University of Toronto, Toronto, ON, Canada (PW); Department of Psychology, Ryerson University, Toronto, ON, Canada (BAA); and Department of Psychological Sciences, University of Missouri, Columbia, MO, USA (VAS). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors received no financial support for the research, authorship, and/or publication of this article.

Corresponding Author
Pete Wegier, Temmy Latner Centre for Palliative Care, Mount Sinai Hospital, Joseph & Wolf Lebovic Health Complex, 60 Murray Street, 4th Floor Box 13, Toronto, ON M5T 3L9, Canada.
(pete.wegier@sinaihealthsystem.ca).

for early detection of the condition and is routinely performed in developed nations around the world.[1] If a pregnant woman opts for screening, she would primarily be interested in the probability that her child will have Down syndrome given a positive test result, known as the *positive predictive value* (PPV; i.e., the conditional probability of having Down syndrome). The PPV of the screening can vary due to factors such as maternal age at conception; for the purpose of this article, we have used statistics for the integrated material serum screening test, averaged across a range of maternal ages.[2,3] Given a prevalence of Down syndrome of 0.12% and a typical sensitivity and specificity of 95% for the screening test, the PPV would be approximately 2.2%.

Conditional probabilities represent a class of statistical information that is often difficult for patients to understand, regardless of education level or numeric ability.[4–9] Prior research shows that practicing clinicians and residents,[5,10] patients,[11] and nonmedical samples of adults[12] have difficulty understanding PPVs. A common logical fallacy is to confuse the PPV of a screening test with the sensitivity of the test—the probability that the test returns a positive result given actually having a disease—a logical fallacy known as *confusion of the inverse*.[13,14] In the case of prenatal Down syndrome screening, confusion of the inverse can result in confusing a probability of 2.2% (the true PPV) with a probability of 95% (the sensitivity of the test). The failure to accurately apply test-accuracy evidence to pretest odds of disease has been associated with systematic errors such as overestimating the PPV,[10] which may have negative effects on patients, such as stress or overtreatment.[15,16]

Formats of Risk Communication

Difficulty understanding risk information may not stem from cognitive ability but rather how information is presented. A growing literature suggests that risk might be effectively communicated through experience rather than descriptive summaries.[4,5,10,12,17] Everyday natural frequency information is accumulated through experience. For example, a clinician may gradually learn the prevalence of disease in a patient population through experiencing individual patient cases.[18] The ecological hypothesis posits that humans have evolved to accumulate naturally occurring frequency information through experience to learn about the environment.[19,20] It is argued that natural frequencies foster insight into Bayesian reasoning (i.e., estimating conditional probabilities) because humans have adapted to process these representations.[21]

Recent research shows that decisions based on descriptive formats are made as if the likelihood of rare events are overweighed, whereas the scarcity of infrequent events becomes more salient when information is experienced.[22] In particular, descriptive formats are associated with overestimating the PPV, while experience formats are associated with more accurate PPV estimates and less systematic errors.[4,5,12,18] In addition to accuracy, experience formats have been shown to change subjective risk perception and intentions to screen. Specifically, learning from experience decreases worry about Down syndrome following a positive test result as well as interest in screening.[12,23]

Other formats of risk communication include static icon arrays, which have been shown to improve understanding of risk,[24–28] likely because they represent risks as frequencies rather than probabilities, while conveying both the numerator and denominator.[29] Previous research examined whether interactive pictographs or icon arrays representing risk information that promote active information processing improve risk comprehension over static icon arrays that promote more passive information processing,[30] with several of these studies reporting negative results. For example, Ancker et al.[31] compared static and interactive icon arrays to communicate risk and found no differences in the accuracy of risk estimates or attitudes about risk. In addition, Zikmund-Fisher et al.[32] found that those in an interactive graphic condition made less accurate judgments than those in a static graphic condition—potentially because of high cognitive demands or distraction from relevant information.

Although interactive icon arrays may not benefit the decision maker, to our knowledge only 1 study has compared risk comprehension when information is presented in either a static icon array or experience format. Specifically, Fraenkel et al.[11] showed that an icon array accompanied by numbers facilitated more accurate knowledge about a lung cancer screening compared with an experience format accompanied by numbers. However, the authors noted that the results may be due to the complexity and amount of information presented at rapid exposure in the experience format compared with the icon array. At present, it is unknown whether an experience format is superior to other formats that present risk information when making Bayesian inferences, such as estimating the PPV of a screening test.

The Present Research

We have recently introduced a new paradigm for the rapid, visual communication of probabilistic information,[12] known as Aiding Risking Information learning through Simulated Experience (ARISE). ARISE combines experience-based learning with the simplicity of

visual aids such as pictographs.[33,34] As it is not possible to have a patient truly experience all the possible outcomes of a medical choice by repeatedly making a choice, ARISE allows for the communication of simulated patient experiences. In addition, using the visual grid approach of pictographs allows for the rapid communication of numerous representative cases. ARISE was designed so that realistic, low-probability events (e.g., 1 in 3000) could be experienced in a quick and simple format.

We previously established that the use of ARISE to communicate information about screening tests was more effective than descriptively providing participants with the relevant probabilities.[12] However, we compared ARISE to a descriptive format that required participants to calculate the PPV using Bayes's theorem. As a clinician would not normally require their patients to calculate a PPV themselves but would more likely provide that statistic to the patient, we decided to investigate how ARISE would compare to explicitly presenting the PPV, in addition to a descriptive format. In experiment 1, participants were presented with the relevant information used to estimate the PPV, in one of three formats: 1) description, a numerical summary explicitly stating the prevalence, sensitivity, and specificity of the screening test in which Bayes' rule could be used to calculate the PPV; 2) explicit, a numerical summary explicitly stating the prevalence, sensitivity, and specificity of the screening test, as well as the PPV; or 3) ARISE, a simulated experience of patient cases in which participants could learn about the PPV over time. In experiment 2, participants were asked to estimate the PPV when information was presented in 1 of 2 conditions: 1) an icon array format or 2) ARISE. Regardless of format, the probabilities presented to participants were the same.

Experiment 1

The goal of the present research was to investigate which format would promote the most accurate PPV estimates. The primary outcome measures were PPV estimate accuracy and change in attitudes toward the screening test across varying format types. In light of previous research, it was hypothesized that those in the explicit and ARISE formats would estimate the PPV more accurately and show unfavorable attitudes toward screening than in the description format[4,12] in experiment 1. Given that no study has examined the accuracy of PPV estimates when the PPV is explicitly provided, no directional hypotheses were set regarding the comparison of the explicit and ARISE formats.

Method

Participants. Two participant samples were recruited. First, a sample of undergraduate students ($N = 285$) completed the study, followed by a replication of the results in an online sample of adults ($N = 297$) to examine the generalizability of the results to an older sample. Undergraduate students were recruited from undergraduate psychology courses at the University of Missouri. The university's research ethics board approved the study. Students provided consent to participate and received course credit for their participation. Online participants were recruited via Amazon's Mechanical Turk service, in which individuals can complete short tasks for small monetary compensation. Online participants provided their consent and were paid $1.50 USD for their participation. Sample characteristics are presented in Table 1.

Design and Procedure. Relevant information about the screening test for Down syndrome was provided in each of the description, explicit, and ARISE formats to investigate which format promotes the greatest understanding of screening test results measured through estimation accuracy of the PPV and attitudes toward screening. Format type was manipulated between subjects. Participants were first presented with introductory information about what Down syndrome is and were asked 3 questions regarding their present attitudes toward screening tests for Down syndrome: 1) "Given what you currently know, how likely would you be to undergo prenatal Down syndrome screening if you (or your partner) were pregnant?" 2) "How concerned would you be if you underwent prenatal Down syndrome screening and were given a POSITIVE test result?" and 3) "How likely are you to recommend prenatal Down syndrome screening to a pregnant friend or loved one?" Responses were given on 6-point Likert-type scales ranging from *extremely unlikely* to *extremely likely* for items 1 and 3 and *extremely unconcerned* to *extremely concerned* for item 2.

Participants were randomly assigned to 1 of the 3 format types. Statistics were always presented as frequencies rather than percentages; however, we use percentages here for simplicity. In the description format, participants were provided with a verbal description of the statistics regarding the prevalence of Down syndrome (12 out of 10,000 or 0.12%), and the sensitivity and specificity (both 9500 out of 10,000 or 95%) of the screening test.[2] In the explicit format, participants were provided with the same information as the description format but were also explicitly presented with a verbal description

Table 1 Participant Characteristics

	Experiment 1			Experiment 2	
	Description	**Explicit**	**ARISE**	**Icon Array**	**ARISE**
Undergraduate sample					
n	97	96	92	138	178
Gender, M/F	31/66	31/65	20/72	70/68	87/90/1[a]
Age, y, M (SD)	18.4 (0.6)	18.8 (2.2)	18.7 (1.7)	19.0 (1.0)	19.2 (1.1)
Age range, y	18–21	18–38	18–32	18–24	18–25
Caucasian race, % (*n*)	86 (83)	86 (83)	85 (78)	80 (111)	80 (143)
Online sample					
n	99	100	98		
Gender, M/F	55/44	51/49	50/47/1[a]		
Age, y, M (SD)	34.8 (9.8)	34.5 (10.6)	33.9 (10.0)		
Age range, y	21–61	19–69	19–64		
Caucasian race, % (*n*)	75 (74)	76 (76)	73 (72)		

[a]Participant did not report gender.

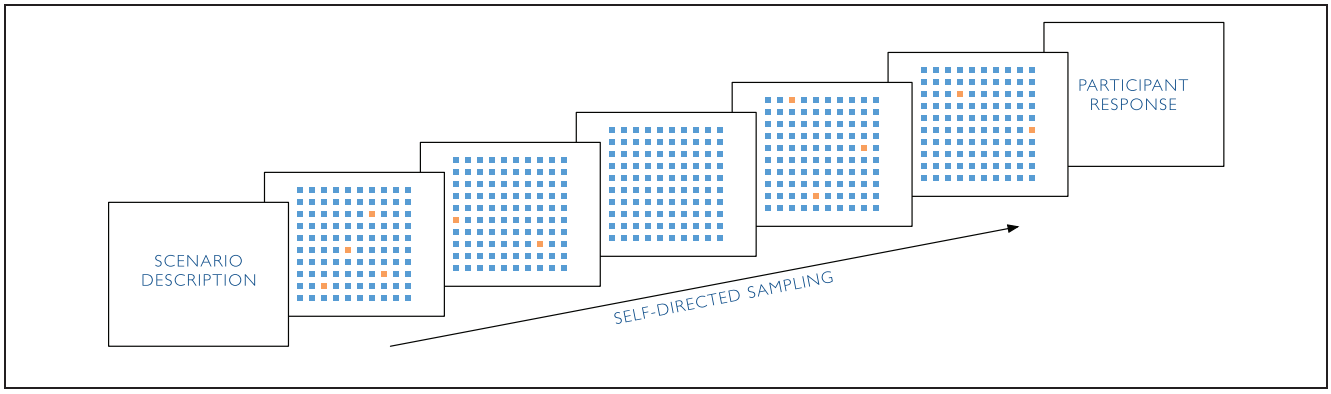

Figure 1 Schematic of the Aiding Risk Information Learning through Simulated Experience (ARISE) format. One hundred (10 × 10) screening test results are presented per grid, with orange squares denoting true positives (positive test result and has Down syndrome) and blue squares denoting false positives (positive test result and does not have Down syndrome).

of the PPV (22 out of 10,000 or 2.2%; i.e., "given a positive test result, there is a 2.2% probability that a fetus will actually have Down syndrome"). In ARISE, participants were shown a series of 10 × 10 grids of colored squares with each square representing a result of a fictional patient who had undergone prenatal screening for Down syndrome and had returned a positive result (Figure 1). Orange squares represented positive test results for those with the disease (true positives), while blue squares represented positive test results for those without the disease (false positives). Participants could view up to 50 grids, representing 5000 simulated test results, with search effort self-directed. The base rates of the orange and blue squares presented in the grids were based on an underlying probability distribution that reflects the true PPV of the screening test. Participants experienced 11 true positives for every 500 simulated test results, representing the true PPV of 2.2%.

Subsequent to the presentation of probability information, participants were asked to estimate the PPV, providing their response as a frequency estimate (i.e., a numerator and denominator) to the question "Given a positive test result, what is the likelihood that a fetus has Down syndrome?" In the description format, participants could derive the PPV using Bayes's theorem (i.e., divide the true positives by the sum of true and false positives); in the explicit format, participants were explicitly provided with the PPV and had to recall this information; and in ARISE, participants could estimate the PPV based on the simulated outcomes they experienced over time. Participants then rated their attitudes toward screening again on the same 3 scales used prior to

Table 2 Estimates Positive Predictive Value (PPV) across Formats

	n[a]	Proportion Correct, % (n)	χ^{2}[b]	P	φ	PPV Estimate Error,[c] M [95% CI]	Mean Difference	t[d]	P	d
Experiment 1 (collapsing undergraduate and online samples)										
Description	174	31% (53)				62.0 [55.6, 68.4]				
Explicit	182	64% (117)				27.6 [21.7, 33.5]				
ARISE	180	78% (141)				10.1 [6.5, 13.7]				
Description v. Explicit			39.4	<0.001	0.34		34.4	7.78	<0.001	0.82
Description v. ARISE			79.9	<0.001	0.48		51.9	14.00	<0.001	1.50
Explicit v. ARISE			8.0	0.004	0.16		17.5	5.00	<0.001	0.52
Experiment 2										
Icon array	131	60% (79)				9.6 [5.0, 14.1]				
ARISE	155	65% (101)				15.3 [10.9, 19.7]				
Icon array v. ARISE			0.52	.469	—		5.74	1.79	.074	—

[a]The number of participants providing PPV estimates does not equal the number of participants recruited, as some participants gave no answer and were excluded from the analysis.
[b]Pearson's chi-square test on proportion of correct PPV estimates.
[c]True PPV was 2.2% and 2% for experiments 1 and 2, respectively.
[d]Welch 2-sample t tests on PPV estimate error.

information exposure to obtain a postinformation measure of their attitudes toward screening. Stimuli were presented via an interactive web application built using JavaScript.

Analyses. We used Pearson's chi-square test with Yates's continuity correction to compare the proportions of participants providing a correct PPV estimate when information was presented via the description, explicit, and ARISE formats. Correct estimates were defined as any estimate greater than zero but less than or equal to 7.2% (i.e., any estimate within ± 5 percentage points of the true PPV of 2.2%). In a second step, we quantified PPV error by mean absolute error (i.e., the absolute difference between the PPV estimate and true PPV) and compared the estimate error between formats using independent t tests (Table 2).

As we found no differences between the undergraduate and online samples in PPV estimate accuracy, we present analyses with the 2 samples collapsed. However, independent samples t tests showed differences between the 2 samples in attitudes toward screening; thus, the results of attitudes toward screening for the 2 samples are presented separately. Using paired-samples t tests, attitudes toward screening were also compared before and after participants had been presented with the PPV information. All analyses were conducted using R.[35]

Given the sample size of the undergraduate sample ($N = 285$) and the online sample ($N = 297$) and the between-subjects design, the statistical power to detect medium-sized effects[36] with an alpha of 0.05 for both samples was 0.97 for the format factor.[37]

Results

Participants shown ARISE sampled an average of 22.6 grids (95% confidence interval [CI] 20.2, 24.9). We found significant differences in PPV estimate accuracy between the 3 formats, with 31% of participants in the description format providing correct PPV estimates compared with 64% in the explicit format and 78% with ARISE (Table 2). Critically, participants deviated from the true PPV to a much smaller degree in ARISE. That is, participants who provided incorrect responses (PPV = 0, or >7.2 and ≠95) still provided low estimates of the PPV, relative to the other 2 formats. Specifically, the average PPV estimate erred by 10% when using ARISE, which is closer to the correct PPV of 2.2% than the average error in the explicit format (28%, $P < 0.001$) and in the description format (62%, $P < 0.001$).

In addition to examining PPV estimate accuracy and the magnitude of error, we were interested in the quality of the estimates themselves. We classified all participant estimates into 6 categories. Correct estimates were any estimate that was greater than 0 and less than or equal to 7.2%—the same accuracy criterion as our previous analysis.[12] Of these, any estimate that was equal to 2.2% was classified as exactly correct.[1] As the prevalence of

[1]Only estimates from the explicit format could be classified as "exactly correct," as we found participants gave estimates to 1 decimal place only when they were repeating information that was explicitly provided in the text (i.e., the PPV presented in the explicit format or the prevalence, sensitivity, or specificity in the description and explicit formats); otherwise, participants only gave estimates as whole numbers.

Table 3 Positive Predictive Value (PPV) Estimate Quality across Conditions (Collapsing Undergraduate and Online Samples)

	Experiment 1			Experiment 2	
	Description	**Explicit**	**ARISE**	**Icon Array**	**ARISE**
n[a]	174	184	180	131	155
Correct responses, % (n)	31 (53)	64 (117)	78 (141)	60 (79)	65 (101)
Exactly correct[b]	0 (0)	47 (86)	0 (0)	60 (79)	12 (18)
Correct (PPV >0 and $\leq$7.2%)[c]	18 (31)	13 (24)	78 (141)	—	53 (83)
Confusion of the prevalence (PPV = 0.12%)[d]	13 (22)	4 (7)	—	—	—
Incorrect responses, % (n)	69 (121)	36 (65)	22 (39)	40 (52)	35 (54)
Incorrect (PPV = 0 or >7.2 and $\neq$95)[e]	17 (30)	16 (29)	22 (39)	40 (52)	35 (54)
Confusion of the sensitivity/specificity (PPV = 95%)[f]	52 (91)	20 (36)	—	—	—

[a]The number of participants providing PPV estimates does not equal the number of participants recruited because some participants gave no answer and were excluded from the analysis.
[b]"Exactly correct" was defined as a PPV estimate of 2.2% and 2% for experiments 1 and 2, respectively.
[c]"Correct" was defined as a PPV estimate that was >0 and $\leq$7.2%. We had previously defined this as an appropriate range of acceptable values. However, in the icon array format of experiment 2, this range was not used, and only estimates that were exactly correct were counted as correct responses. Because of the nature of the icon array format presenting only a single possible value (2%) to participants, we felt that any other responses participants provided were incorrect responses.
[d]"Confusion of the prevalence" was defined as a PPV estimate of 0.12%.
e "Incorrect" was defined as any PPV estimate that did not fall into any of the other categories.
f "Confusion of the sensitivity/specificity" was defined as a PPV estimate of 95%.

the disease was 0.12% and fell in our range of a correct response (i.e., 0%–7.2%), rather than classifying estimates of 0.12% as "confusion of the prevalence," we chose to classify these estimates as "correct" responses, as we could not determine if they were confusing the prevalence with certainty. Incorrect estimates were any estimates equal to 0 or greater than 7.2%. Any estimates equal to 95%, which represented both the sensitivity and specificity of the screening test, were also incorrect and classified as "confusion of the sensitivity/specificity." Finally, no response or nonsense responses were classified as "no answer." The results of this classification can be found in Table 3.

Only 31% of participants in the description format provided a correct estimate, with 13% of the 31% being the result of the participants confusing the PPV with the provided prevalence of Down syndrome (i.e., 0.12%). No participant provided the exact PPV or confused the sensitivity and specificity with the PPV in the description format. Despite being provided with the PPV explicitly, participants in the explicit format provided an exactly correct estimate only 47% of the time. Participants gave an estimate that was equal to other statistical information they had been provided 17% of the time in the explicit format, suggesting it was difficult for them to discern which statistic represented the PPV. No participants provided an exact estimate of 2.2% or made the confusion of the prevalence or confusion of the sensitivity/specificity errors in ARISE, as this information was not numerically provided to participants but had to be inferred over time.

Attitudes toward screening did not change for participants in the description format. However, attitudes were more negative in the explicit and ARISE formats toward screening, with more negative ratings of screening in ARISE relative to the explicit format (and with more negative postratings emerging in the online sample compared with the undergraduate sample; Table 4).

Discussion

We investigated whether presenting probability information regarding maternal serum screening test results for Down syndrome using ARISE[12] would result in more accurate PPV estimates and unfavorable attitudes toward screening compared with a description and explicit format. In line with our hypothesis, the use of ARISE to communicate conditional probability information resulted in more accurate estimates of the PPV, compared with providing participants with a numerical description of the information to calculate the PPV (description format) or explicitly providing the PPV (explicit format).

Overall, the results suggest that those who were explicitly provided the PPV and did not correctly report it may have confused the PPV with other information provided in the description. In the case of the screening test used, confusing the prevalence (0.12%) as the PPV would result in a PPV estimate that is relatively close to the true PPV. However, confusing the PPV with the sensitivity or specificity of the test (95%) may lead to different and potential harmful decisions. The use of simulated

Table 4 Experiment 1: Pre/Post Differences in Attitudes toward Screening[a]

	Undergraduate Sample						Online Sample					
	Pre, M [95% CI]	Post, M [95% CI]	Mean Difference[b]	*t*	*P*	*d*	Pre, M [95% CI]	Post, M [95% CI]	Mean Difference[b]	*t*	*P*	*d*
Likelihood of undergoing screening												
Description	4.1 [3.8, 4.4]	4.1 [3.8, 4.4]	−0.02	0.24	0.810	—	4.6 [4.3, 4.9]	4.5 [4.2, 4.8]	−0.11	0.98	0.329	—
Explicit	4.2 [3.9, 4.4]	3.9 [3.7, 4.2]	−0.21	1.82	0.071	—	4.4 [4.0, 4.7]	4.1 [3.8, 4.4]	−0.25	2.34	0.021	0.24
ARISE	4.1 [3.8, 4.4]	3.9 [3.5, 4.2]	−0.20	1.51	0.134	—	4.4 [4.0, 4.7]	3.9 [3.6, 4.3]	−0.46	3.77	<0.001	0.39
Concern regarding a positive test result												
Description	4.4 [4.1, 4.6]	4.3 [4.1, 4.6]	−0.04	0.50	0.614	—	5.1 [4.9, 5.4]	5.0 [4.8, 5.2]	−0.14	1.84	0.068	—
Explicit	4.5 [4.3, 4.8]	4.2 [3.9, 4.4]	−0.33	3.42	<0.001	0.36	4.9 [4.7, 5.2]	4.6 [4.3, 4.8]	−0.35	2.86	0.005	0.29
ARISE	4.4 [4.1, 4.6]	3.9 [3.6, 4.2]	−0.50	5.12	<0.001	0.54	5.1 [4.9, 5.3]	4.6 [4.4, 4.8]	−0.54	4.22	<0.001	0.44
Likelihood of recommending screening												
Description	3.8 [3.6, 4.1]	4.0 [3.7, 4.2]	0.14	−1.63	0.107	—	4.4 [4.1, 4.6]	4.2 [3.9, 4.5]	−0.16	2.18	0.031	0.22
Explicit	4.0 [3.7, 4.3]	3.7 [3.5, 4.0]	−0.26	2.52	0.013	0.26	4.1 [3.8, 4.4]	3.8 [3.5, 4.2]	−0.24	2.83	0.005	0.28
ARISE	4.0 [3.8, 4.3]	3.7 [3.4, 4.0]	−0.35	2.93	0.004	0.31	4.2 [3.9, 4.5]	3.8 [3.5, 4.2]	−0.37	3.36	0.001	0.35

[a]Ratings were made on a 6-point Likert-type scales, with 1 = *extremely unlikely* and 6 = *extremely likely* for items 1 and 3, and 1 = *extremely unconcerned* and 6 = *extremely concerned* for item 2.
[b]Mean differences were calculated as the postestimate positive predictive value (PPV) rating minus the preestimate PPV rating.

experiences for the communication of this information not only promoted more accurate PPV estimates but also shifted participants' attitudes further away from screening. The shift in attitudes observed may be evidence of greater comprehension of the PPV when learning about risk information via the ARISE format, given that the PPV used in the current study was low.

Experiment 2

Prior research has shown that learning probability information through simulated experience or graphics such as icon arrays promotes more accurate PPV estimates compared with descriptive formats of numerical summaries.[4,5,12,33,38] Further, experience-based formats[12] and icon arrays[39] have previously been shown to shift attitudes away from screening in comparison with descriptive formats. Only 1 study to our knowledge has compared an experience format with an icon array format, with the icon array leading to more accurate knowledge about risk than simulated experiences.[11] However, this study did not examine estimates based on conditional probabilities.

Mirroring experiment 1, the primary outcome measures were PPV estimate accuracy and change in attitudes toward the screening test across varying format types—in this case ARISE compared with an icon array. We hypothesize that ARISE will promote more accurate PPV estimates and reduce favorable attitudes toward screening after information exposure compared with an icon array in experiment 2. Importantly, no numbers were presented in either format, the stimuli presented was simple, and the presentation of each grid was self-paced for ARISE.

Method

Participants. Undergraduate students ($N = 316$) completed the study. Participants were recruited from undergraduate psychology courses at the University of Missouri and could not have participated in experiment 1. The university's research ethics board approved the study. Students provided consent to participate and received course credit for their participation. Sample characteristics are presented in Table 1.

Design and Procedure. The design of experiment 2 was very similar to experiment 1, except we compared ARISE to an icon array format. The icon array was nearly identical in appearance to a single grid from ARISE: a static 10 × 10 grid of 100 squares that represented true-positive and false-positive test results. Orange squares represented true positives, and blue squares represented false positives. A description of what each color represented remained on the screen while the participant made his or her estimate (i.e., orange: received a positive test and HAS Down syndrome; blue: received a positive test and DOES NOT have Down syndrome). All participants assigned to the icon array format were presented the same icon array: 2 orange squares placed in the top left

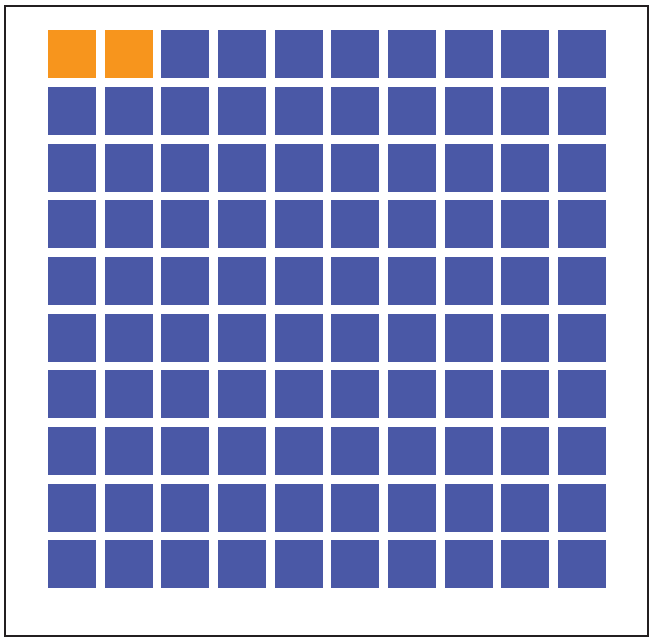

Figure 2. Experiment 2 icon array. One hundred (10 × 10) screening test results are presented, with orange squares denoting true positives (positive test result and has Down syndrome) and blue squares denoting false positives (positive test result and does not have Down syndrome).

corner of the array, with the remaining 98 squares colored blue (Figure 2). That is, to correctly estimate the PPV, participants had to identify that the 2 orange squares reflected the PPV. This slightly shifted the underlying PPV for experiment 2 to be 2%, rather than 2.2%.

Analyses. Similar to experiment 1, we used Pearson's chi-square test with Yates's continuity correction to compare the proportions of participants providing a correct PPV estimate in the icon array and ARISE formats. The classifications for correct and incorrect responses varied slightly from experiment 1. "Exactly correct" was defined as being 2%. Only estimates that were exactly correct were counted as correct responses for the icon array format. Because of the nature of the icon array format presenting only a single possible value (2%) to participants, we felt that any other responses participants provided were incorrect responses. For the ARISE format, PPV estimates were classified as "correct" if estimates were between 0 and $\leq 7.2\%$, as we had in experiment 1, and "exactly correct" as 2%. All other responses were defined as "incorrect." As participants were not provided with explicit numerical probabilities, confusion of the prevalence or sensitivity/specificity with the PPV was not expected. In line with experiment 1, we compared estimate error across formats using independent *t* tests (Table 2). Using paired-samples *t* tests, attitudes toward screening were also compared before and after participants had been presented with the PPV information.

Given the sample size of the undergraduate sample ($N = 316$) and the between-subjects design, the statistical power to detect medium-sized effects,[36] with an alpha of 0.05, was 0.99 for the format factor.[37]

Results

Participants shown ARISE sampled an average of 18.3 grids (95% CI 16.2, 20.4). The results showed no significant effect of format, with 60% of participants in the icon array format correctly estimating the PPV and 65% of those in ARISE correctly estimating the PPV. The average PPV estimate erred approximately 10% in the icon array format and 15% in ARISE relative to the true PPV. Table 3 presents the details of the estimate classification.

Similar to experiment 1, participants rated how likely they were to undergo a prenatal Down syndrome screening, how concerned they would be if they received a positive test result, and how likely they would be to recommend the screening to others before and after exposure to screening information. Interestingly, the willingness to undergo screening significantly decreased for those in ARISE but not for those in the icon array format. Table 5 shows participants' pre/post ratings of their attitudes towards screening in both formats.

Discussion

We investigated whether ARISE would facilitate more accurate PPV estimates and affect attitudes toward screening compared with an icon array format. First, no difference was observed between formats on the accuracy of PPV estimates, with both ARISE and the icon array format promoting similar levels of accuracy. Prior research has shown that icon arrays facilitate accurate estimates of medical risks and can reduce cognitive biases such as denominator neglect,[28,33] with Fraenkel et al.[11] showing that an icon array format accompanied by numbers led to more accurate PPV estimates compared with a simulated experience format. However, as we presented participants with only 2 possible outcomes—true positives and false positives—for both formats in experiment 2, it remains to be seen how icon arrays and ARISE might compare with more complex sets of outcomes (i.e., the addition of true- and false-negative test results).

Table 5 Experiment 2: Pre/Post Differences in Attitudes toward Screening[a]

	Pre, M [95% CI]	Post, M [95% CI]	Mean Difference[b]	*t*	*P*	*d*
Likelihood of undergoing screening						
Icon array	4.2 [4.0, 4.4]	4.0 [3.7, 4.2]	−0.22	1.86	0.066	—
ARISE	3.9 [3.7, 4.1]	3.7 [3.5, 3.9]	−0.20	2.11	0.036	0.16
Concern regarding a positive test result						
Icon array	4.5 [4.2, 4.7]	4.1 [3.8, 4.3]	−0.38	4.06	<0.001	0.36
ARISE	4.5 [4.3, 4.6]	4.0 [3.8, 4.2]	−0.50	6.60	<0.001	0.50
Likelihood of recommending screening						
Icon array	4.1 [3.9, 4.4]	3.8 [3.6, 4.0]	−0.32	3.27	0.001	0.29
ARISE	3.9 [3.7, 4.1]	3.7 [3.4, 3.9]	−0.21	2.38	0.019	0.18

[a]Ratings were made on 6-point Likert-type scales with 1 = *extremely unlikely* and 6 = *extremely likely* for items 1 and 3, and 1 = *extremely unconcerned* and 6 = *extremely concerned* for item 2.
[b]Mean differences were calculated as the postestimate positive predictive value (PPV) rating minus the preestimate PPV rating.

Second, although we hypothesized that ARISE would lead to more accurate PPV estimates, given the relative simplicity of the information presented, we expected a greater proportion of people to correctly estimate the PPV in the icon array format. For those who have less experience with medical screening tests, it may be counterintuitive to come across a screening test with a high frequency of false positives, which may have discouraged participants from estimating a low PPV in the icon array format. Future research is encouraged to further examine how icon arrays representing true and false positives are comprehended by laypeople and whether predictive value estimates based on icon arrays lead to informed decision making in a real-life medical context.

Finally, similar to experiment 1, attitudes toward screening were more negative after information exposure compared with before information exposure, with larger shifts observed for ARISE. Gradually experiencing patient outcomes over time may highlight how rare true positives are, which may have led to less favorable attitudes toward screening. Repeated exposure to rare outcomes may be the key feature of ARISE that further ingrains how rare true positives are compared with the icon array format that presents representative frequencies. Whether the attitudes affected by the icon array and ARISE formats lead to differences in real medical decisions (e.g., to undergo screening or not) remains to be addressed in future work.

General Discussion

The accuracy of estimating conditional probabilities is sensitive to the format of information presentation.[4,12,19,33] There is evidence that descriptive formats in which statistical summaries are presented can lead to inaccurate conditional probability estimates and systematic errors.[40] It has been posited that estimating predictive values such as the PPV is outside of our cognitive capacity when information is described in a numerical summary.[20] Alternative ways to communicate probabilistic information that are more digestible for the decision maker, such as learning via experience or graphical representation, have been proposed and have shown to elicit more accurate predictive value estimates compared with descriptive formats.[4,12,33]

The primary goal of the current study was to examine the effect varying formats have on PPV estimate accuracy and attitudes toward screening tests. Consistent with our prior research,[4,5,12] experiencing patient cases over time elicited more accurate PPV estimates compared with description formats. A clear advantage of ARISE is that estimates are less variable across participants, with most PPV estimates close to the true PPV of the test (Figure 2). The results of experiment 1 replicate prior work showing evidence of systematic error—with the PPV commonly being confused with other properties of the screening test such as the rate of sensitivity in the description formats.[40] This is true even when the PPV is explicitly provided. There are 2 reasons why ARISE may be superior to description and explicit formats. First, ARISE does not present probabilistic information that could be distracting or used as a heuristic that may lead to error in judgment. Second, experiencing information via ARISE highlights how rare or how common an event is. Medical patients vary in working memory as well as in numeric and graphical abilities. ARISE was designed to bypass reliance on high numeracy and constraints of cognitive flexibility and instead present frequency information to be gradually learned to arrive at the gist of probability.

Critically, this study is the first to show that even when the PPV is explicitly provided, systematic errors

(e.g., confusing the PPV with the sensitivity or specificity of the test) and large estimation errors are commonly made. One explanation is that nonmedical experts may not conceptually understand true- and false-positive test results. For example, it may be thought that screening tests always accurately detect the presence of disease. When asked, "Given a positive test result, what is the probability of having a disease?" it may be a common misconception to think this probability will be high. Although speculative, in the description and explicit formats, a high probability (e.g., the sensitivity) may be sought out in the information provided and used as the PPV estimate and low probabilities undervalued because they are incongruent with the schema of screening tests accurately detecting the presence of disease. In addition, the use of Bayes's rule to correctly estimate a predictive value may be unavailable or unintuitive to most people. These results hold practical implications for what and how information is shown to patients by clinicians. In particular, these results suggest that it cannot be assumed that the PPV is well understood by nonmedical experts even when explicitly provided.

In line with previous research,[33,41] the icon array format facilitated accurate PPV estimates, producing similarly accurate estimates to those of ARISE. As briefly noted above, one explanation for these results could be due to the simplicity of the icon array. Whereas the static icons in the icon array format directly infer the PPV, ARISE demonstrates the variability of screening test accuracy across a large data set (i.e., a maximum of 5000 patient cases). Exposure to the variance of event outcomes—disease status and test result—may have influenced more error than the static grid used in the icon array format.

Secondarily, we found shifts in attitudes toward screening in both experiments. Learning via ARISE often resulted in participants reporting they would be less likely to undergo screening in the future, would be less concerned with a positive screening result, and would be less likely to recommend screening to a loved one. Similar shifts were observed in the explicit format of experiment 1 and the icon array format of experiment 2; these effects tended to be smaller when compared with ARISE. These shifts in attitudes toward screening are likely to be the result of an improved understanding of the underlying probabilities of the screening test.

Limitations

First, only 2 outcomes were presented in the ARISE and icon array formats of experiments 1 and 2, namely, true-positive and false-positive test results. Although this is the relevant information for maternal serum screening for Down syndrome, the task of estimating the PPV in these formats may have been simple given the binary nature of the information provided. The description and explicit formats of experiment 1 provided more information to participants (i.e., specificity, sensitivity, prevalence), which could have affected their estimates. Future research should investigate whether icon array and the ARISE formats promote accurate PPV estimates when more complex information is presented (i.e., true and false, positive and negative test results). Second, the samples tested in the current experiment likely comprised many participants who are not currently pregnant or are parents and who were making hypothetical decisions. Pregnant participants, for example, may be more motivated to complete the task because of personal relevance of the study. It is important for future research to determine whether varying the formats of risk information affect patients who make real-life medical decisions. Third, the samples across both experiments were primarily Caucasian participants. To ensure the results are generalizable to the greater population, future research is encouraged to include a distribution of participants varying in demographics such as education level, socioeconomic status, and race/ethnicity.

Conclusion

There exist many different formats for conveying probabilistic information to patients required to make medical decisions, with some formats proving to be more effective than others. Previously, we demonstrated that ARISE was superior to a description format that required calculation of conditional probabilities on the patients' part. Here, we investigated how ARISE compares with more explicit formats of risk information, namely, explicitly providing the PPV to participants or presenting information via a simple icon array. Unlike those in the ARISE format, systematic errors were observed when the PPV was explicitly provided, suggesting that nonmedical experts may not conceptually understand predictive values. Although we found comparable estimate accuracy of the PPV using ARISE compared with the icon array format, ARISE affected attitudes toward screening compared with the icon array format. These findings add to the growing literature that simulated experiences are a simple and viable way to communicate conditional probability information in medical decision scenarios and can be a powerful tool for improving patient understanding in medical choice.

ORCID iD

Pete Wegier https://orcid.org/0000-0003-0191-136X

References

1. Twiss P, Hill M, Daley R, Chitty LS. Non-invasive prenatal testing for Down syndrome. *Semin Fetal Neonat M*. 2014;19(1):9–14. doi: 10.1016/j.siny.2013.10.003.
2. California Department of Public Health. *The California Prenatal Screening Program Provider Handbook*. Sacramento, CA: California Department of Public Health; 2009.
3. MacDonald ML, Wagner RM, Slotnick RN. Sensitivity and specificity of screening for Down syndrome with alpha-fetoprotein, hCG, unconjugated estriol, and maternal age. *Obstet Gynecol*. 1991;77(3):63–8.
4. Armstrong B, Spaniol J. Experienced probabilities increase understanding of diagnostic test results in younger and older adults. *Med Decis Making*. 2017;37(6):670–9. doi: 10.1177/0272989X17691954.
5. Armstrong B, Spaniol J, Persaud N. Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St Michael's Hospital, Toronto, Canada. *BMJ Open*. 2018;8(2):e019241–6. doi: 10.1136/bmjopen-2017-019241.
6. Gigerenzer G, Galesic M. Why do single event probabilities confuse patients? *BMJ*. 2012;344:e245. doi: 10.1136/bmj.e245.
7. Reyna VF, Nelson WL, Han PK, Dieckmann NF. How numeracy influences risk comprehension and medical decision making. *Psychol Bull*. 2009;135(6):943–73. doi: 10.1037/a0017327.
8. Nelson W, Reyna VF, Fagerlin A, Lipkus I, Peters E. Clinical implications of numeracy: theory and practice. *Ann Behav Med*. 2008;35(3):261–74. doi: 10.1007/s12160-008-9037-8.
9. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. *Med Decis Making*. 2001;21(1):37–44. doi: 10.1177/0272989X0102100105.
10. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. *Psychol Sci Publ Interest*. 2007;8(2):53–96. doi: 10.1111/j.1539-6053.2008.00033.x.
11. Fraenkel L, Peters E, Tyra S, Oelberg D. Shared medical decision making in lung cancer screening: experienced versus descriptive risk formats. *Med Decis Making*. 2016;36(4):518–525. doi: 10.1177/0272989X15611083.
12. Wegier P, Shaffer VA. Aiding Risk Information learning through Simulated Experience (ARISE): using simulated outcomes to improve understanding of conditional probabilities in prenatal Down syndrome screening. *Patient Educ Couns*. 2017;100(10):1882–9. doi: 10.1016/j.pec.2017.04.016.
13. Eddy DM. Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. *Judgement under Uncertainty: Heuristics and Biases*. Cambridge (UK): Cambridge University Press; 1982. p 249–67.
14. Steurer J, Fischer JE, Bachmann LM, Koller M, Reitter G. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ*. 2002;324:824–6.
15. Bhatt JR, Klotz L. Overtreatment in cancer—is it a problem? *Expert Opin Pharmacol*. 2016;17(1):1–5. doi: 10.1517/14656566.2016.1115481.
16. Wegwarth O, Gigerenzer G. Trust-your-doctor: a simple heuristic in need of a proper social environment. In: Hertwig R, Hoffrage U, ABC Research Group, eds. *Simple Heuristics in a Social World*. Oxford (UK): Oxford University Press; 2013. p 67–102. doi: 10.1093/acprof:oso/9780195388435.003.0003.
17. Galesic M, Gigerenzer G, Straubinger N. Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Med Decis Making*. 2009;29(3):368–71. doi: 10.1177/0272989X08329463.
18. Obrecht NA, Anderson B, Schulkin J, Chapman GB. Retrospective frequency formats promote consistent experience-based Bayesian judgments. *Appl Cogn Psychol*. 2012;26(3):436–40. doi: 10.1002/acp.2816.
19. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol Rev*. 1995;102(4):684–704. doi: 10.1037/0033-295x.102.4.684.
20. Cosmides L, Tooby J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*. 1996;58(1):1–73. doi: 10.1016/0010-0277(95)00664-8.
21. Hoffrage U, Gigerenzer G, Krauss S, Martignon L. Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*. 2002;84(3):343–52.
22. Hertwig R, Barron G, Weber EU, Erev I. Decisions from experience and the effect of rare events in risky choice. *Psychol Sci*. 2004;15(8):534–39. doi: 10.2139/ssrn.1301100.
23. Tyszka T, Sawicki P. Affective and cognitive factors influencing sensitivity to probabilistic information. *Risk Anal*. 2011;31(11):1832–45. doi: 10.1111/j.1539-6924.2011.01644.x.
24. Zikmund-Fisher BJ, Ubel PA, Smith DM, et al. Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. *Patient Educ Couns*. 2008;73(2):209–14. doi: 10.1016/j.pec.2008.05.010.
25. 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.
26. Hawley ST, Zikmund-Fisher BJ, Ubel P, Jancovic 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.
27. 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. doi: 10.1177/0272989X05278931.

28. Garcia-Retamero R, Galesic M, Gigerenzer G. Do icon arrays help reduce denominator neglect? *Med Decis Making*. 2010;30(6):672–84. doi: 10.1177/0272989X10369000.
29. Fagerlin A, Zikmund-Fisher BJ, Ubel PA. Helping patients decide: ten steps to better risk communication. *J Natl Cancer Inst*. 2011;103(19):1436–43. doi: 10.1093/jnci/djr318.
30. Natter HM, Berry DC. Effects of active information processing on the understanding of risk information. *Appl Cogn Psychol*. 2005;19(1):123–35. doi: 10.1002/acp.1068.
31. Ancker JS, Weber EU, Kukafka R. Effects of game-like interactive graphics on risk perceptions and decisions. *Med Decis Making*. 2011;31(1):130–142. doi: 10.1177/0272989X10364847.
32. Zikmund-Fisher BJ, Dickson M, Witteman HO. Cool but counterproductive: interactive, web-based risk communications can backfire. *J Med Internet Res*. 2011;13(3):e60–17. doi: 10.2196/jmir.1665.
33. Galesic M, Garcia-Retamero R, Gigerenzer G. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychol*. 2009;28(2):210–16. doi: 10.1037/a0014474.
34. Garcia-Retamero R, Hoffrage U. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Soc Sci Med*. 2013;83:27–33. doi: 10.1016/j.socscimed.2013.01.034.
35. R Core Team. *R: A Language and Environment for Statistical Computing*. Available at: https://www.R-project.org.
36. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–9.
37. Erdfelder E, Faul F, Buchner A. GPOWER: a general power analysis program. *Behav Res Methods Instrum Comput*. 1996;28(1):1–11. doi: 10.3758/bf03203630.
38. Garcia-Retamero R, Cokely ET, Hoffrage U. Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Front Psychol*. 2015;6:713–12. doi: 10.3389/fpsyg.2015.00932.
39. Hersch J, Barratt A, Jansen J, et al. Use of a decision aid including information on overdetection to support informed choice about breast cancer screening: a randomised controlled trial. *Lancet*. 2015;385(9978):1642–52. doi: 10.1016/S0140-6736(15)60123-4.
40. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Communicating statistical information. *Science*. 2000;290(5500):2261–2.
41. Brewer NT, Richman AR, DeFrank JT, Reyna VF, Carey LA. Improving communication of breast cancer recurrence risk. *Breast Cancer Res Treat*. 2011;133(2):553–61. doi: 10.1007/s10549-011-1791-9.w