Patient Education and Counseling 100 (2017) 1882–1889

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pateducou

Aiding risk information learning through simulated experience (ARISE): Using simulated outcomes to improve understanding of conditional probabilities in prenatal Down syndrome screening

Pete Wegier[a,*], Victoria A. Shaffer[a,b]

[a] Department of Health Sciences, University of Missouri, Columbia, MO, USA
[b] Department of Psychological Sciences, University of Missouri, Columbia, MO, USA

ARTICLE INFO

Article history:
Received 19 November 2016
Received in revised form 21 April 2017
Accepted 25 April 2017

Keywords:
Simulated experience
Screening tests
Conditional probabilities
Positive predictive value
Attitudes towards screening

ABSTRACT

Objective: To determine whether the use of visually-presented simulated experiences to communicate statistical information can improve an individual's understanding of conditional probabilities—specifically the positive predictive value (PPV) of prenatal screening tests for Down syndrome.
Methods: In Experiment 1 ($N = 64$) and Experiment 2 ($N = 180$) participants were asked to estimate the PPV of a prenatal screening test for Down syndrome based on either (1) explicit statistics regarding the prevalence of Down syndrome and the sensitivity and specificity of a prenatal screening test for Down syndrome, or (2) experiencing up to 5000 simulated test results over a short time.
Results: Participants' estimates of the PPV were more accurate when they had learned via simulated experiences (79% accuracy) compared with estimates based on explicitly described statistics (14%). Participants in the simulated experience condition also reported decreased interest in screening and decreased concern with a positive test result.
Conclusion: A visual paradigm presenting simulated experiences improves PPV estimates, compared to estimates derived from explicitly provided statistics, while also shifting attitudes away from screening.
Practice implications: The use of simulated experiences may prove to be simple but powerful tool to communicate complex statistical information to patients in medical decision making situations.

1. Introduction

1.1. Difficulties in understanding conditional probabilities in screening tests

Imagine a woman pregnant with her first child who chooses to undergo noninvasive prenatal screening for Down syndrome in her first trimester. If her obstetrician informs her that the test result is positive, she will likely want to know: What is the probability that my child will have Down syndrome? To answer this question, she requires an understanding of conditional probabilities. A conditional probability is the probability that an event *A* occurs given that another event *B* has occurred. For example, what is the probability that her child will have Down syndrome, given a positive result on a prenatal screening test? A conditional probability can be calculated using Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where

P(A) and *P(B)* are the probabilities of events *A* and *B* occurring,

P(A|B) is the probability of event *A* occurring given that event *B* has occurred, and

P(B|A) is the probability of event *B* occurring given that event *A* has occurred.

Using this notation, the probability that this woman's child will have Down syndrome given a positive test result can be written: *P*(Down syndrome|Positive test result). Conditional probabilities represent a type of statistical literacy highly relevant in healthcare settings. However, research has demonstrated that both patients and physicians have difficulties understanding, calculating, and applying conditional probabilities to inferences in medicine [1–3].

One reason that people have difficulty applying conditional probabilities is the confusion about which conditional probability is appropriate to apply in which case. When evaluating the accuracy of screening tests authors typically report the *sensitivity*

* Corresponding author at: University of Missouri, 501 Clark Hall, Columbia, Missouri, 65211, USA.
E-mail address: wegierp@health.missouri.edu (P. Wegier).

http://dx.doi.org/10.1016/j.pec.2017.04.016
0738-3991/

and *specificity* of a test. The sensitivity represents the probability that a child with Down syndrome will have a positive test result, *P* (positive result|Down syndrome). The specificity represents the probability that a child without Down syndrome will have a negative test result, *P* (negative result|no Down syndrome). In the case of noninvasive prenatal screening, the sensitivity and the specificity of the test to detect Down syndrome are approximately 95%, varying slightly depending on a variety of factors [4,5].

However, in the context of prenatal testing, we are interested in the inverse of this relationship—the *positive predictive value* (PPV) of the test—the probably that an infant will have Down syndrome given a positive screening test result, *P* (Down syndrome|positive test result). A common logical fallacy is to equate a conditional probability with its inverse—believing that the probably of event *A* occurring given that event *B* has occurred is the same as the probability of event *B* occurring given that event *A* has occurred, *P*(A|B) = *P*(B|A)—and is commonly referred to as *confusion of the inverse*. In the case of screening tests, physicians commonly mistake the sensitivity of the test for its PPV [6,7]; in this case, the physician would mistakenly conclude the probability that an infant has Down syndrome given a positive test result is 95%. However, given the base rate of Down syndrome in the population is roughly 1 in 700, we can use Bayes' Theorem to calculate the actual PPV, which is approximately 2%.

Despite the fact that the PPV of a test is often much smaller than the sensitivity of the test, even highly educated individuals confuse the application of these two conditional probabilities [8,9]. Gigerenzer et al. [9] investigated how well physicians understood the results of mammography screening by providing information about the test (prevalence of breast cancer, sensitivity of the test, and false alarm rate) and asking physicians to calculate the probability that a woman has breast cancer given a positive test result (PPV of the test). They found most physicians vastly overestimated the probability the woman had breast cancer, only 21% of physicians provided the correct answer, and variability in estimates ranged from 1% to 90%.

There are significant consequences to misunderstanding the PPV of a screening test. For example, mistakenly believing that the PPV is equal to the sensitivity of the test will lead patients and physicians to severely underestimate the number of false positives the test produces. This can produce acceptance of the initial positive test results without conducting follow-up tests, leading to unnecessary anxiety. Anecdotal reports have suggested that patients may terminate pregnancies that may have resulted in healthy children, due to false positives [10,11].

1.2. Learning from description or experience

Recently, psychologists have investigated an alternative paradigm for the communication of probabilistic information, termed learning from experience. While not a novel format—learning from experience has evolutionary roots—this form of information presentation has received greater attention in recent years as a way to improve decision making under uncertainty. Some psychologists have used learning from experience in decision problems represented by financial gambles (e.g., "Would you prefer $3 with certainty or an 80% chance to win $4"). Participants are either explicitly told the odds and decide which option they prefer (i.e., *a decision from description*), or they must first learn the odds by participating in the gambles repeatedly without consequence and then decide which option they prefer (i.e., *a decision from experience*). Using the above example, the certain $3 option is more often selected if the decision was made from description, but the risky $4 option is more often selected if the decision was made from experience [12]. When probabilistic information is presented through description, decisions are made as if the likelihood of rare events is overweighted; however, different decision making behavior is observed if the same information is learned through repeated experiences [13–16].

What drives this difference in information processing between description and experience is not yet agreed upon. However, evidence suggests that we have a robust ability to automatically encode frequency information over time with little effort or even intent [17]. This ability may make the understanding of complex statistical information, such as conditional probabilities, through experienced information more manageable. Thus, the use of an experience-based paradigm may elicit more accurate understandings of probabilistic information than the traditional descriptive approach.

However, gaining such firsthand experiences is not feasible in a medical decision making context—for example, it is not possible to have a patient experience the outcomes of a surgery many times to learn the probabilities of success, failure, and side effects occurring—so experience must be gained in other ways. Shaffer and colleagues have shown that narratives from individuals who have experienced the possible outcomes of a medical decision are a useful tool for communicating experiential information, representing a type of *experience by proxy* [18–20]. However, it is difficult to communicate experiences of rarely occurring outcomes via narratives due to the potential for overrepresentation. For example, if a disease occurs in only 1% of a population but only 1 of 5 presented narratives describe a case where the disease is present, one may assume that the prevalence of the disease is closer to 20% than to 1%. A more representative set of narratives should be presented—potentially difficult due to the length of narrative information—or additional information about the underlying prevalence is required.

As learning from experience has led to improved understanding about probabilistic information, this paradigm may be employed to improve understanding of probabilities in medical contexts. For example, Tyszka and Sawicki [21] found that the use of an experience-based paradigm to communicate prevalence information resulted in participants showing greater sensitivity to the underlying prevalence of a disease and decreased subjective ratings of worry about the condition. Relatedly, Fraenkel et al. [22] used an experience-based method to communicate results of lung cancer screening, which increased participants' understanding of the false alarm rate; but, attitudes towards screening remained unchanged. Armstrong and Spaniol [37] demonstrated that presenting information from experience improved Bayesian inferences, and therefore PPV estimates, across both younger and older adults. However, these studies were limited as each sequentially presented single experiences to participants. As the presentation of a single simulated experience takes several seconds, this makes the presentation of enough experiences to accurately communicate low base rate events—such as the prevalence of Down syndrome—difficult. Taken together, simulating experiences appears a viable approach to communicating probabilistic information in medical decision making; however, an approach is required which allows for the rapid communication of large amounts of simulated experiences.

1.3. The present research

The goal of the present research was to test the hypothesis that simulated experiences improve communication about conditional probabilities, specifically the PPV of screening tests, using a format allowing for the communication of a large number of simulated experiences. We report the results of two experiments investigating the use of an experience-based approach to improve patient understanding of diagnostic screening tests compared to providing descriptive information only. We introduce a novel paradigm—

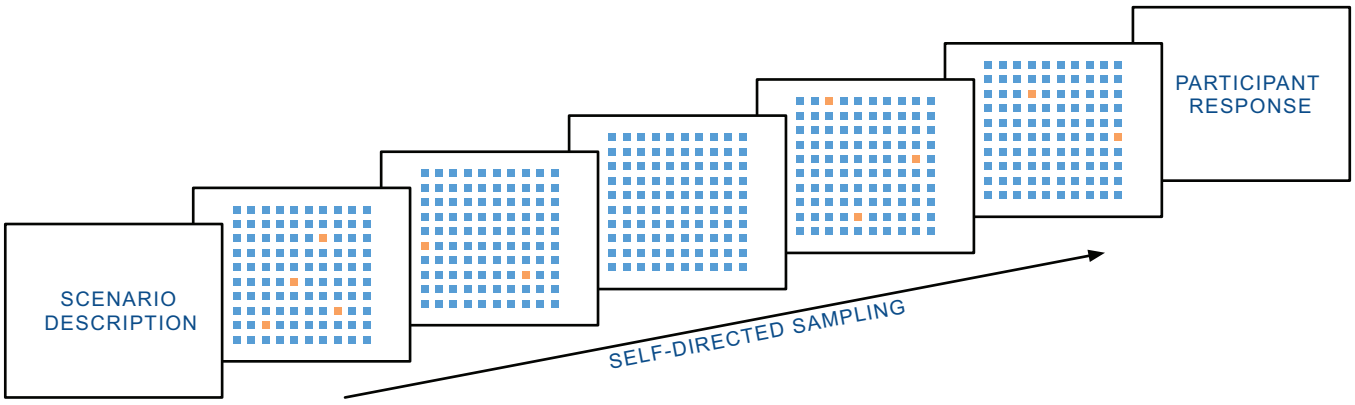

Fig. 1. Experience condition visualized. A hundred screening test results are presented at a time, with blue squares denoting false positives (positive test result but does not actually have Down syndrome) and orange squares denoting true positives (positive test result and does have Down syndrome). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Aiding Risk Information learning through Simulated Experience, or ARISE—which allows participants to experience large amounts of possible outcomes of prenatal Down syndrome screening through sequentially presented grids of colored squares, with each grid representing 100 possible outcomes of screening tests. This paradigm allows individuals to experience very low base rate events in reasonable lengths of time. In Experiment 1, we investigated whether undergraduate participants were more accurate in determining the PPV of a screening test when the statistical information was provided in a description format or a simulated experience format. Experiment 2 replicated Experiment 1 in an online sample, and additionally investigated whether the different learning formats (description vs. simulated experience) impacted attitudes towards screening.

2. Experiment 1

2.1. Methods

2.1.1. Design & procedure

Participants were first provided with basic information about Down syndrome and prenatal Down syndrome screening. To examine the differences in understanding about the PPV of the screening test, participants were randomly assigned to two experimental conditions: learn about the accuracy of the tests either through description (explicitly provided statistical information) or through simulated experience. The experiment was conducted via an interactive web app we built in JavaScript. In the description condition, participants were provided with the prevalence of Down syndrome (0.12%), and the specificity (95%) and sensitivity (95%) of the screening test. This information was based on statistics from the California Department of Public Health's Prenatal Screening Program Handbook [5]. In the experience condition, participants were shown grids of 100 colored squares, with each square representing the result of a fictional patient that had undergone prenatal Down syndrome screening and received a positive test result (see Fig. 1). Blue squares represented patients who received a positive test result but whose children did not have Down syndrome (false positives) while orange squares represented patients who received a positive test result and whose children did have Down syndrome (true positives). Search effort was self-directed; participants could sample up to 50 of these grids—representing a total of 5000 simulated screening test results.

The base rates of blue and orange squares in the grids were based on a probability distribution designed to mimic the underlying true PPV of our simulated screening test. For every five grids sampled (500 cases observed) participants would see 11 true positives (representing the underlying PPV of ~2.2%). Any given grid had between 0 and 5 true positives (orange squares).

After participants had studied the provided statistical information to their satisfaction, either from description or simulated experience, they were asked to estimate the PPV of the screening test (i.e., "Given a positive test result, what is the likelihood that a fetus has Down syndrome?"). As evidence exists that the use of natural frequencies may improve understanding of conditional probabilities [23–27], participants gave their estimates in one of two formats (assignment to format was random)—probabilities or natural frequencies (where participants provided both the numerator and denominator). This resulted in a 2 (learning format: description vs. simulated experience) × 2 (format of elicitation: probabilities vs. natural frequencies) between-subjects design.[1] Once participants had provided their estimates, they completed the 8-item Subjective Numeracy Scale [28] and the non-adaptive version of the 4-item Berlin Numeracy Test [29].

2.1.2. Analyses

In this experiment, we were interested in how participants' PPV estimates were affected by the format in which they learned the statistical information (description vs. simulated experience). We hypothesized that participants in the simulated experience condition would provide more accurate PPV estimates than the participants in the description condition. First, we used Fisher's exact test to investigate if the proportion of participants providing a correct PPV estimate was greater in the simulated experience condition than the description condition. We defined a correct estimate as one that was within ±5 percentage points of the correct PPV of 2.2% or 0.022 (i.e., any estimate greater than zero and less than or equal to 7.2%). Second, we calculated the magnitude of error in participants' PPV estimates by subtracting the correct PPV estimate from participants' estimates. We conducted a regression on the magnitude of error to examine which factors predict the amount of error in the PPV responses. Predictors of the regression model included the method of learning (description vs. simulated experience), format of elicitation (probability vs. natural frequency), and participant scores on the Berlin Numeracy Test and

[1] The statistics for the prevalence of Down syndrome, and the sensitivity and specificity of the screening test were provided as natural frequencies to participants in the description condition who were assigned to answer using natural frequencies.

Table 1
Participant characteristics.

	Description	Simulated experience
Experiment 1		
N	31	33
Gender	16 M/15 F	21 M/12 F
Age—*M, (SD)*	19.4 (2.1)	19.4 (1.3)
Age range	18–29	18–24
In first year university—*N, %*	23 (74%)	20 (61%)
Berlin Numeracy Test—*M, (SD)*	1.1 (1.0)	1.2 (1.1)
Subjective Numeracy Scale—*M, (SD)*	4.2 (0.8)	4.3 (0.8)
Experiment 2		
N	92	88
Gender	47 M/45 F	49 M/39 F
Age—*M, (SD)*	36.4 (10.9)	34.3 (11.2)
Age range	19–69	18–63
Berlin Numeracy Test—*M, (SD)*	1.7 (1.4)	1.8 (1.4)
Subjective Numeracy Scale—*M, (SD)*	4.7 (0.8)	4.7 (1.0)

Subjective Numeracy Scale. Analyses were conducted using R [30].

2.2. Results

2.2.1. Participants

Participants were recruited from an undergraduate psychology course at a university in the Midwestern region of the United States. The study was approved by the university's research ethics board. All students provided consent to participate and received course credit for their participation. Participant characteristics are presented in Table 1.

2.2.2. Positive predictive value estimates

We found a difference of 70 percentage points between conditions—76% of participants in the simulated experience condition provided an accurate PPV estimate; however, only 6% of participants in the description condition could do the same. Thus, learning by simulated experience significantly improved the accuracy of PPV estimates compared to learning by description, $\chi^2(1) = 28.70$, $p < .001$. Further, most participants in the description condition (81%) provided a PPV estimate of 95%, which was the sensitivity/specificity of the test. Learning by simulated experience also significantly decreased the amount by which participants erred. The average PPV estimate in the learning by simulated experience condition was 12%, while the average PPV estimate provided by participants learning through description was 84%, $t(52.5) = 10.98$, $p < .001$, $d = 2.76$ (recall that the correct PPV estimate is ~2%). However, neither the format of elicitation (probabilities or frequencies) nor participant numeracy (measured either objectively or subjectively) was a significant predictor of the magnitude of error in PPV estimates. See Table 2 for details about the parameter estimates from the regression model.

2.3. Discussion

Participant estimates of the PPV for the prenatal screening test for Down syndrome were much more accurate when they learned about the screening test through ARISE, our novel simulated experience condition, than through explicitly described statistics, which is how this type of information is typically presented to patients. Why was simulated experience a more effective method of communicating this information? Of the participants in the description condition, 81% ($N = 25$) provided a PPV estimate of 95%, which was equal to the sensitivity/specificity of the prenatal Down syndrome test. However, in the simulated experience condition no participants provided a PPV estimate of 95%. It is likely that without the simulated learning environment, participants either mistook the sensitivity of the test for the PPV—committing the fallacy of the inverse—or confusing the PPV with the specificity of the test. We believe the former to be more likely, as the phrasing of the sensitivity statistic—"*if a fetus has Down syndrome, then the screening test will return a positive result 9500 out of 10,000 times*"—is almost identical to the question asking for a PPV estimate—"*if the test result of your prenatal screening was positive, what is likelihood that your fetus has Down syndrome?*"—with only the order of clauses varying. While the phrasing for the specificity of the test—"*If a fetus does not have Down syndrome, then the screening test will return a negative result 9500 out of 10,000 times.*"—is negated. However, as both the sensitivity and specificity of the test were equal, we cannot definitively conclude participants were committing the fallacy of the inverse.

3. Experiment 2

The goal of Experiment 2 was to replicate our previous results with a more representative sample and to investigate whether the use of simulated experience to communicate conditional probability information would also shift attitudes towards Down syndrome screening.

3.1. Methods

3.1.1. Design & procedure

Experiment 2 was identical in design to Experiment 1, with the following exceptions. First, rather than undergraduate students, we recruited an online sample of participants via Amazon's

Table 2
Regression results on PPV estimates.

	β	*SE*	*p*
Experiment 1			
Intercept		18.77	<.001
Learning Format (Description vs. Simulated Experience)	−0.82	6.71	<.001
Format of Elicitation (Probability vs. Frequency)	−0.03	6.81	.721
Berlin Numeracy Test	0.03	3.50	.690
Subjective Numeracy Scale	0.03	4.32	.758
Experiment 2			
Intercept		12.04	<.001
Learning Format (Description vs. Simulated Experience)	−0.77	4.34	<.001
Berlin Numeracy Test	0	1.55	.973
Subjective Numeracy Scale	−0.04	2.55	.440

Table 3
Experiment 2—Regression results on responses regarding attitudes towards screening.

	β	*SE*	*p*
Likelihood of undergoing screening			
Intercept		0.44	<.001
Time of Response (Pre vs. Post)	−0.01	0.21	.909
Learning Format (Description vs. Simulated Experience)	−0.27	0.22	<.001
Time of Response × Learning Format	0.25	0.31	.005
Berlin Numeracy Test	0.14	0.06	.008
Subjective Numeracy Scale	0.001	0.09	.979
Concern regarding a positive test result			
Intercept		0.31	<.001
Time of Response (Pre vs. Post)	0.05	0.15	.469
Learning Format (Description vs. Simulated Experience)	−0.36	0.15	<.001
Time of Response × Learning Format	0.33	0.22	<.001
Berlin Numeracy Test	0.11	0.04	.038
Subjective Numeracy Scale	0.05	0.06	.363
Likelihood of recommending screening			
Intercept		0.45	<.001
Time of Response (Pre vs. Post)	−0.01	0.22	0.887
Learning Format (Description vs. Simulated Experience)	−0.25	0.23	<.001
Time of Response × Learning Format	0.22	0.32	.012
Berlin Numeracy Test	0.05	0.06	.354
Subjective Numeracy Scale	−0.06	0.09	.273

Mechanical Turk, a service in which individuals can complete short tasks for monetary compensation. The service has become popular in the social sciences for the rapid collection of general samples of participants [31]. Participants were compensated $1.50 USD for their participation, paid upon completion of the study.

Second, we included several questions to assess participants' attitudes towards prenatal Down syndrome screening. After participants had been provided with the basic information about Down syndrome and prenatal Down syndrome screening, they were asked: (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?"*. Participants provided their responses on 6-point Likert scales ranging from "Extremely Unlikely" to "Extremely Likely" for questions 1 and 3, and "Extremely Unconcerned" to "Extremely Concerned" for question 2.

Third, as no difference in PPV estimate accuracy due to format of elicitation was observed in Experiment 1, we only asked participants to provide their PPV estimates as natural frequencies in Experiment 2. Finally, after participants had provided their PPV estimates, they were asked about their attitudes towards screening once again to investigate any changes that may have occurred as a result of learning about the statistical properties of the screening test. This gave us a final design in which: (1) participants were introduced to Down syndrome screening; (2) participants' attitudes to screening were recorded; (3) screening statistics were provided either via description or simulated experience; (4) participants estimated the PPV of the screening test; and (5) participants' attitudes toward screening were recorded again.

3.1.2. Analyses

Based on the findings from Experiment 1, we hypothesized that participants would provide more accurate PPV estimates in the simulated experience condition than in the description condition. We used Fisher's exact test to investigate this and conducted a regression on the magnitude of error in participants' PPV estimates to examine which factors impacted estimate accuracy. Additionally, we hypothesized that participants' attitudes towards screening would shift after learning from simulated experience, based on prior findings [21]. Specifically, we expected participants to be less likely to undergo screening, less concerned with a positive test result, and less likely to recommend screening, if they had learned about the screening test through our simulated experience paradigm (ARISE) than through the more traditional descriptive approach. To test these hypotheses, we conducted a repeated measures regression on participants' ratings before and after they had learned the statistical information about the screening test, to investigate any differences in attitudes towards screening between the two learning formats. Analyses were conducted using R [30].

3.2. Results & discussion

3.2.1. Participants

Participant characteristics are presented in Table 1.

3.2.2. Positive predictive value estimates

We observed a large difference in PPV estimate accuracy of 63 percentage points between conditions, where 80% of participants in the simulated experience condition provided an accurate PPV estimate; however, only 17% of participants in the description condition could do the same. These findings replicate Experiment 1, indicating that accuracy of PPV estimates was again significantly greater when learning by simulated experience than learning by description, $\chi^2(1) = 68.75$, $p < .001$. Further, most participants in the description condition (76%) provided a PPV estimate of 95%, which was the sensitivity/specificity of the test. Learning by simulated experience again significantly decreased the amount by

Table 4
Experiment 2—Attitudes toward prenatal screening for Down syndrome, M (SD).

Item	Description (*N* = 92)					Simulated experience (*N* = 88)				
	Pre	Post	Mean diff.[a]	*p*[b]	*d*[c]	Pre	Post	Mean diff.[a]	*p*[b]	*d*[c]
1. Likelihood of undergoing screening	4.6 (1.4)	4.6 (1.5)	0	.628	–	4.7 (1.3)	3.8 (1.6)	−1.0	<.001	0.65
2. Concern regarding a positive test result	5.3 (0.9)	5.2 (1.1)	−0.1	.146	–	5.4 (0.8)	4.4 (1.2)	−1.1	<.001	0.68
3. Likelihood of recommending screening	4.3 (1.5)	4.3 (1.6)	0	.671	–	4.3 (1.4)	3.5 (1.6)	−0.9	<.001	0.48

Note: Ratings were made on 6-point Likert 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.

[a] Mean differences were calculated from unrounded pre/post values.

[b] Denotes the results of paired-sample *t*-tests on the pre/post differences in attitudes for each learning format.

[c] Cohen's *d* for the effect size of the paired-sample *t*-tests on pre/post differences.

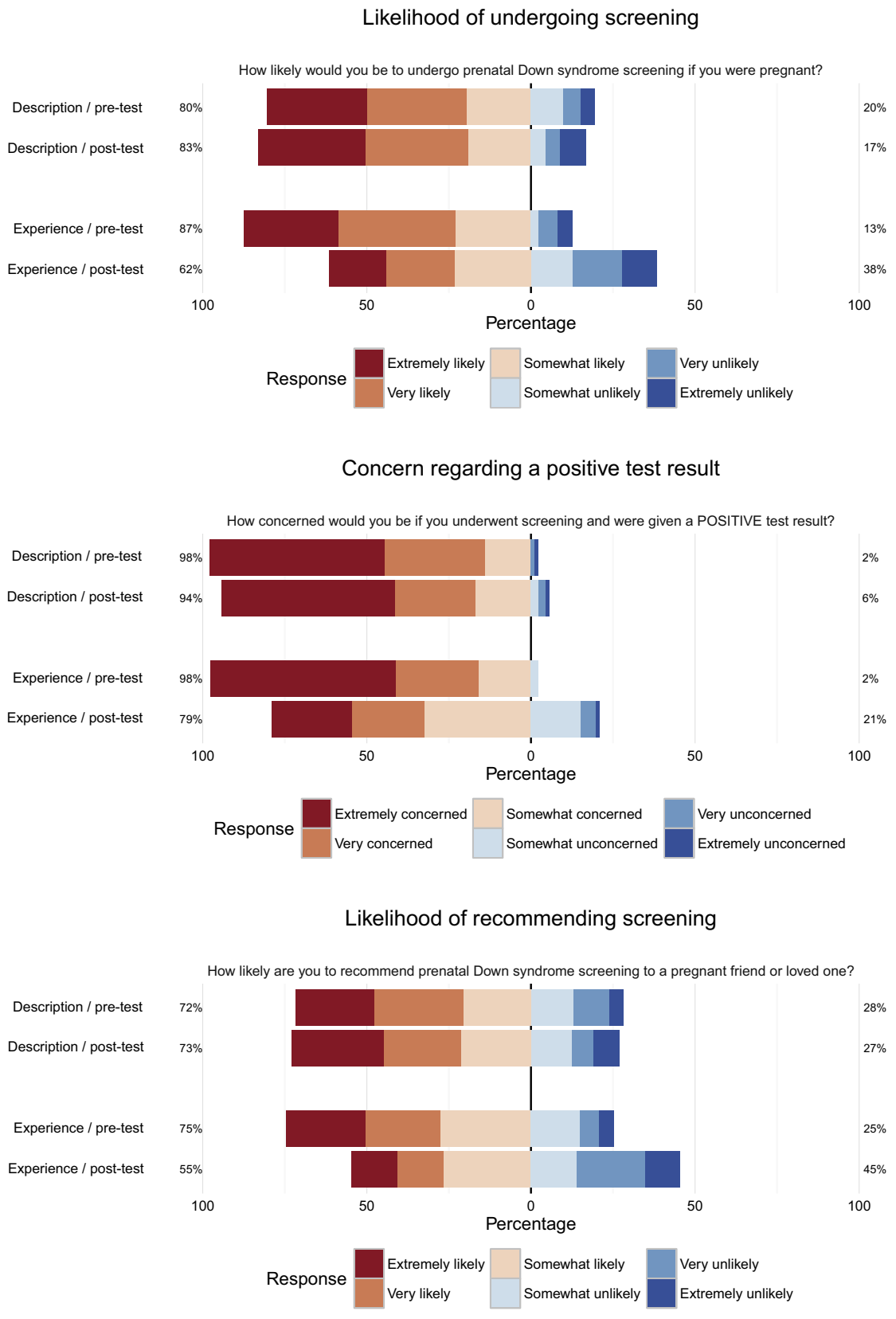

Fig. 2. Experiment 2–Participant attitudes towards screening.

which participants erred. The average PPV estimate in the learning by simulated experience condition was 10%, while the average PPV estimate provided by participants learning through description was 79%, $t(138.33) = 16.09$, $p < .001$, $d = 2.41$ (recall that the correct PPV estimate is ~2%). Consistent with Experiment 1, participant numeracy did not significantly predict the magnitude of error in PPV estimates. See Table 2 for parameter estimates from the regression model.

3.2.3. Attitudes towards screening

Participants were asked to rate their attitudes towards screening on a 6-point Likert scale both before learning any statistical information about the screening test and after they had provided their estimates of the PPV (referred to as pre- and-post learning, respectively). For all three questions, we found a significant interaction between time (pre- vs. post-learning of the statistical information) and learning format (description vs. simulated experience). See Table 3 for parameter estimates from these models, and see Table 4 for means and standard deviations of the attitude measures. Specifically, attitudes toward screening were the same before and after learning by description. However, attitudes toward screening became more negative after learning through simulated experience (Fig. 2). This figure visualizes the distributions of participant responses to the questions regarding their attitudes towards screening across the two learning formats (description vs. simulated experience) and time of response (pre- and post-test).

4. Discussion & conclusion

4.1. Discussion

In this paper, we introduced ARISE, a novel simulated-experience paradigm designed to better communicate probabilistic information—specifically low base rate events—during medical decision making. The paradigm presents probability information using color-coded grids, allowing for the rapid simulation of one hundred experiences at a time, providing a more accurate representation of rare events. Across two experiments, we demonstrated that the use of ARISE significantly improved participant understanding of conditional probabilities in diagnostic screening tests. When participants learned statistical properties of the screening test from simulated experience, their estimates of the PPVs were significantly more accurate compared with participant estimates derived from the same information presented as explicitly described statistics. When we assessed attitudes towards screening before and after participants had learned about the screening, we found no significant changes in the participants' ratings on any of these items before and after learning in the description condition. However, participants reported 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 after learning through simulated experience. These shifts in attitudes towards screening are likely the result of the improved understanding about the PPV of the screening test learned through simulated experience.

4.2. Limitations

Only two outcomes (true positives and false positives) were presented to participants in the experience condition. While this was appropriate given the medical context, it is difficult to say how successful the ARISE paradigm would be at communicating multiple categories of possible outcomes—although evidence exists that the facilitation effect of presenting information as a natural frequency versus a probability has generalized from simple to complex tasks [26]. Future research should attempt to replicate the findings presented here in a context with more complex outcome possibilities.

Alternative formats also exist for the communication of PPV information that were not included in this study—e.g., icon arrays—which may prove successful. A comparison between these alternative formats, description, and simulated experience is needed in future work. Other limitations included the use of an online sample of participants, the use of a hypothetical scenario rather than patients seeking information about prenatal Down syndrome screening, and the selection of identical values for the sensitivity and specificity of test, which prevented us from concluding which of these two values participants were confusing for the PPV. Finally, explicitly providing participants with the PPV, rather than having participants estimate it described statistics, may result in different findings. This comparison of ARISE to explicitly provided PPV information should be examined in future research.

4.3. Conclusion

Attempts have been made to improve patient understanding of numerical information in a variety of ways, such as converting probabilities into natural frequencies [9], pictographs [32–34], and other visual aids [35,36]. The use of experience-based paradigms have also been found to improve understanding of probabilistic information [13–16]. The ARISE paradigm presented here combines elements from each of these formats to improve communication about tests for low prevalence conditions (e.g. Down syndrome). Participants gain a more accurate understanding of the underlying PPV of a screening test if they experience the outcomes than when presented with statistical information from description only. Coupled with past findings of improved understanding of statistical information [21,22,37], the use of simulated experience can be a powerful—but very simple and easy to use—tool in patient education.

4.4. Practice implications

The ARISE paradigm would allow a physician to easily communicate numerical information to patients with a simple tool, without the need for a deep understanding of statistics by either party. Additionally, the use of a grid presentation format for simulated experience information provides the ability to present large quantities of information very rapidly. We presented participants with up to 5000 simulated screening test results, allowing individuals to experience events with very low probabilities of occurrence (e.g., 1 in 3000) that may not have been otherwise possible to convey. This research suggests that simulated experience is a simple and quick way to communicate complex numerical information to patients quickly and effectively.

Conflicts of interest

The authors have no conflicts of interest to report.

Role of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

We are grateful to Haylee P. Cromer, Sean X. Duan, & Jordan A. Keitt for assistance with data collection.

References

[1] I.M. Lipkus, G. Samsa, B.K. Rimer, General performance on a numeracy scale among highly educated samples, Med. Decis. Mak. 21 (2001) 37–44, doi:http://dx.doi.org/10.1177/0272989X0102100105.
[2] V.F. Reyna, W.L. Nelson, P.K. Han, N.F. Dieckmann, How numeracy influences risk comprehension and medical decision making, Psychol. Bull. 135 (2009) 943–973, doi:http://dx.doi.org/10.1037/a0017327.
[3] G. Gigerenzer, M. Galesic, Why do single event probabilities confuse patients? Br. Med. J. 344 (2012) e245, doi:http://dx.doi.org/10.1136/bmj.e245.
[4] M.L. MacDonald, R.M. Wagner, R.N. Slotnick, Sensitivity and specificity of screening for Down syndrome with alpha-fetoprotein, hCG, unconjugated estriol, and maternal age, Obstet. Gynecol. 77 (1991) 63–68.
[5] California Department of Public Health, The California Prenatal Screening Program Provider Handbook, (2009) .
[6] D.M. Eddy, Probabilistic reasoning in clinical medicine: problems and opportunities, in: D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgement Under Uncertainty: Heuristics and Biases, 1982, pp. 249–267.
[7] J. Steurer, J.E. Fischer, L.M. Bachmann, M. Koller, G. ter Reit, Communicating accuracy of tests to general practitioners: a controlled study, Br. Med. J. 324 (2002) 824–826.
[8] J.A. Paulos, Innumeracy Mathematical Illiteracy and Its Consequences, Hill and Wang, New York, NY, 1988.
[9] G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L.M. Schwartz, S. Woloshin, Helping doctors and patients make sense of health statistics, Psychol. Sci. Public Interest 8 (2007) 53–96, doi:http://dx.doi.org/10.1111/j. 1539-6053.2008.00033x.
[10] B. Daley, Oversold and misunderstood, Prenatal Screening Tests Prompt Abortions, New England Center for Investigative Reporting, 2014 http://eyenecir.org/2014/12/13/prenatal-testing/ (Accessed 6 September 2016).
[11] S.W. Cheung, A. Patel, T.Y. Leung, Accurate description of DNA-based noninvasive prenatal screening, N. Engl. J. Med. 372 (2015) 1675–1677, doi:http://dx.doi.org/10.1056/NEJMc1412222.
[12] R. Hertwig, G. Barron, E.U. Weber, I. Erev, Decisions from experience and the effect of rare events in risky choice, Psychol. Sci. 15 (2004) 534–539, doi:http://dx.doi.org/10.2139/ssrn.1301100.
[13] R. Hertwig, I. Erev, The description–experience gap in risky choice, Trends Cogn. Sci. 13 (2009) 517–523, doi:http://dx.doi.org/10.1016/j.tics.2009.09.004.
[14] R. Hau, T.J. Pleskac, R. Hertwig, Decisions from experience and statistical probabilities: why they trigger different choices than a priori probabilities, J. Behav. Decis. Mak. 23 (2010) 48–68, doi:http://dx.doi.org/10.1002/bdm.665.
[15] R.M. Hogarth, E. Soyer, Sequentially simulated outcomes: kind experience versus nontransparent description, J. Exp. Psychol. Gen. 140 (2011) 434–463, doi:http://dx.doi.org/10.1037/a0023265.
[16] T.T. Hills, R. Hertwig, Information search in decisions from experience: do our patterns of sampling foreshadow our decisions? Psychol. Sci. 21 (2010) 1787–1792, doi:http://dx.doi.org/10.1177/0956797610387443.
[17] L. Hasher, R.T. Zacks, Automatic processing of fundamental information: the case of frequency of occurrence, Am. Psychol. 39 (1984) 1372–1388, doi:http://dx.doi.org/10.1037/0003-066X.39.12.1372.
[18] V.A. Shaffer, B.J. Zikmund-Fisher, All stories are not alike: a purpose-, content-, and valence-based taxonomy of patient narratives in decision aids, Med. Decis. Mak. 33 (2013) 4–13, doi:http://dx.doi.org/10.1177/0272989X12463266.
[19] V.A. Shaffer, L. Hulsey, B.J. Zikmund-Fisher, The effects of process-focused versus experience-focused narratives in a breast cancer treatment decision task, Patient Educ. Couns. 93 (2013) 1–10, doi:http://dx.doi.org/10.1016/j.pec.2013.07.013.
[20] V.A. Shaffer, E.S. Focella, L.D. Scherer, B.J. Zikmund-Fisher, Debiasing affective forecasting errors with targeted, but not representative, experience narratives, Patient Educ. Couns. (2016) 1–9, doi:http://dx.doi.org/10.1016/j.pec.2016.04.004.
[21] T. Tyszka, P. Sawicki, Affective and cognitive factors influencing sensitivity to probabilistic information, Risk Anal. 31 (2011) 1832–1845, doi:http://dx.doi.org/10.1111/j.1539-6924.2011.01644.x.
[22] L. Fraenkel, E. Peters, S. Tyra, D. Oelberg, Shared medical decision making in lung cancer screening: experienced versus descriptive risk formats, Med. Decis. Mak. 36 (2016) 518–525, doi:http://dx.doi.org/10.1177/0272989X15611083.
[23] R. Bramwell, H. West, P. Salmon, Health professionals and service users interpretation of screening test results: experimental study, Br. Med. J. 333 (2006) 284–286, doi:http://dx.doi.org/10.1136/bmj.38884.663102.AE.
[24] G. Gigerenzer, Risk Savvy: How to Make Good Decisions, Viking, New York, NY, 2014.
[25] U. Hoffrage, G. Gigerenzer, Using natural frequencies to improve diagnostic inferences, Acad. Med. 73 (1998) 538–540.
[26] U. Hoffrage, S. Krauss, L. Martignon, G. Gigerenzer, Natural frequencies improve Bayesian reasoning in simple and complex inference tasks, Front. Psychol. 6 (2015) 214–226, doi:http://dx.doi.org/10.3389/fpsyg.2015.01473.
[27] M. Galesic, G. Gigerenzer, N. Straubinger, Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests, Med. Decis. Mak. 29 (2009) 368–371, doi:http://dx.doi.org/10.1177/0272989X08329463.
[28] A. Fagerlin, B.J. Zikmund-Fisher, P.A. Ubel, A. Jankovic, H.A. Derry, D.M. Smith, Measuring numeracy without a math test: development of the Subjective Numeracy Scale, Med. Decis. Mak. 27 (2007) 672–680, doi:http://dx.doi.org/10.1177/0272989X07304449.
[29] E.T. Cokely, M. Galesic, E. Schulz, S. Ghazal, R. Garcia-Retamero, Measuring risk literacy: the Berlin numeracy test, Judgm. Decis. Mak. (2012) 25–47.
[30] R Core Team R: A Language and Environment for Statistical Computing, (n.d.). https://www.R-project.org.
[31] M. Buhrmester, T. Kwang, S.D. Gosling, Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. 6 (2011) 3–5, doi:http://dx.doi.org/10.1177/1745691610393980.
[32] B.J. Zikmund-Fisher, A.M. Angott, P.A. Ubel, The benefits of discussing adjuvant therapies one at a time instead of all at once, Breast Cancer Res. Treat. 129 (2010) 79–87, doi:http://dx.doi.org/10.1007/s10549-010-1193-4.
[33] A.R. Tait, T. Voepel-Lewis, B.J. Zikmund-Fisher, A. Fagerlin, The effect of format on parents' understanding of the risks and benefits of clinical research: a comparison between text, tables, and graphics, J. Health Commun. 15 (2010) 487–501, doi:http://dx.doi.org/10.1080/10810730.2010.492560.
[34] R. Garcia-Retamero, U. Hoffrage, Visual representation of statistical information improves diagnostic inferences in doctors and their patients, Soc. Sci. Med. 83 (2013) 27–33, doi:http://dx.doi.org/10.1016/j.socscimed.2013.01.034.
[35] M. Galesic, R. Garcia-Retamero, G. Gigerenzer, Using icon arrays to communicate medical risks: overcoming low numeracy, Health Psychol. 28 (2009) 210–216, doi:http://dx.doi.org/10.1037/a0014474.
[36] R. Garcia-Retamero, E.T. Cokely, B. Wicki, A. Joeris, Improving risk literacy in surgeons, Patient Educ. Couns. 99 (2016) 1156–1161, doi:http://dx.doi.org/10.1016/j.pec.2016.01.013.
[36] R. Garcia-Retamero, E.T. Cokely, B. Wicki, A. Joeris, Improving risk literacy in surgeons, Patient Educ. Couns. 99 (2016) 1156–1161, doi:http://dx.doi.org/10.1016/j.pec.2016.01.013.
[37] B. Armstrong, J. Spaniol, Experienced probabilities increase understanding of diagnostic test results in younger and older adults, Med. Decis. Mak. 37 (2017) 670–679, doi:http://dx.doi.org/10.1177/0272989X17691954.