Experienced Probabilities Increase Understanding of Diagnostic Test Results in Younger and Older Adults

Bonnie Armstrong, MA, Julia Spaniol, PhD

Background. *With advancing age, the frequency of medical screening increases. Interpreting the results of medical tests involves estimation of posterior probabilities such as positive predictive values (PPVs) and negative predictive values (NPVs). Both laypeople and experts are typically poor at estimating posterior probabilities when the relevant statistics are communicated descriptively. The current study examined whether an experience format would improve posterior probability judgments in younger and older adults, relative to a description format.* ***Method.*** *Eighty younger (ages 17–34 y) and 80 older adults (ages 65–87 y) completed an experimental task in which information about medical screening tests for 2 fictitious diseases was presented either through description or experience. Participants in the descriptive format read a passage containing statistical information, whereas participants in the experience format viewed a slideshow of representative cases that illustrated the relative frequency of the disease as well as the relative frequency of positive and negative test results.* ***Results.*** *Both younger and older adults made more accurate posterior probability estimates in the experience format, relative to the description format. In the descriptive format, PPVs were overestimated and NPVs were underestimated. Regardless of format type, participants reported that they would prefer to rely on a physician to make medical decisions on their behalf compared with themselves.* ***Discussion.*** *These findings are indicative of a description-experience gap in Bayesian inference, and they suggest possible avenues for enhancing medical risk communication for both younger and older patients.* ***Key words:*** *description-experience gap; Bayesian inference; medical screening tests; older adults; numeracy; risk communication.* ***(Med Decis Making XXXX;XX:xx–xx)***

Medical decision making has shifted from a provider-centered model, in which doctors have the responsibility of making patient decisions, to a shared, patient-centered model in which patients are more involved in decisions about their medical care.[1–4] Effective use of health statistics by patients is therefore becoming increasingly important and a prerequisite for informed decision making. However, communication and interpretation of these statistics are typically fraught with problems.[5–7]

A specific difficulty involved in the interpretation of medical statistics is the estimation of posterior probabilities, such as $p(\mathrm{A}\,|\,\mathrm{B})$. According to Bayes' theorem,

$$p(A|B) = \frac{p(A) * p(B|A)}{p(B)}.$$

In the case of medical diagnosis, "A" refers to the patient's true disease status (has disease or does not have disease), and "B" refers to the patient's test result (positive or negative). Two posterior probabilities are of particular interest in medicine: the positive predictive value (PPV), which is the probability

Received 4 June 2016 from the Department of Psychology, Ryerson University, Toronto, Ontario, Canada (BA, JS). Financial support for this study was provided in part by a grant from the Natural Sciences and Engineering Research Council (DG No. 358797 to JS), by the Canada Research Chair program (JS), and an Early Researcher Award from the Ontario Ministry of Research and Innovation (JS). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. Revision accepted for publication 6 January 2017.

The online appendix for this article is available on the *Medical Decision Making* Web site at http://journals.sagepub.com/home/mdm.

Address correspondence to Bonnie Armstrong, MA, Department of Psychology, Ryerson University, 350 Victoria St., Toronto, ON M5B 2K3, Canada; e-mail: bonnie.armstrong@psych.ryerson.ca.

DOI: 10.1177/0272989X17691954

that the patient has the disease given a positive test result, and the negative predictive value (NPV), which is the probability that the patient does not have the disease given a negative test result.

Gigerenzer and others[7] reported a study in which obstetric-gynecologic physicians were tested on their ability to estimate PPVs in the context of breast cancer screening. Most respondents vastly overestimated the PPV of a positive mammogram as being greater than 80%, when in fact it was 10%. This type of overestimation may result in overdiagnosis, potentially harmful follow-up diagnostics such as biopsies, and unnecessary escalation of patient stress.[7]

In light of the difficulties associated with processing risks framed as probabilities, frequency-based risk formats have received extensive study. For example, Galesic and others[6] compared the effect of natural-frequency and conditional-probability formats on younger and older adults' comprehension of medical test results. Natural frequencies (e.g., "50 out of 10,000 people have insulin-dependent diabetes") elicited more accurate estimates than did conditional probabilities (e.g., "The probability that a person has insulin-dependent diabetes is 0.5%"). This benefit was observed for both younger and older adults, as well as those of high and low numeracy level. However, approximately 45% of participants in the natural-frequency condition still made significant estimation errors, suggesting that interpretation of test results remains challenging even when probabilities are expressed in terms of natural frequencies.

In addition to a reliance on natural frequencies, a great deal of research has focused on visualization tools such as icon arrays, whose effectiveness at improving risk comprehension may depend on the target group's graph literacy and numeracy level.[8–10] As reviewed next, a relatively less explored strategy for improving risk communication involves "experienced probabilities," that is, exposure to probability distributions over time, rather than via static numerical or graphical displays.

Descriptive versus Experience-Based Risk Formats

Decisions from description are based on explicit summaries of a priori probability information.[11] Decisions from experience are based on statistical probabilities and may be influenced by recency, exploration-exploitation tradeoff, and other cognitive factors.[12,13] An example serves to illustrate the distinction: One may base the decision to carry an umbrella on the probability of rain in the weather forecast (*description*) or on the rainfall pattern in recent days (*experience*). Within the risky-choice literature, the "description-experience gap"[14–16] refers to the finding that people choose as if they overweight small described risks and underweight small experienced risks. Applied to the context of Bayesian inference in medical diagnosis, then, it is possible that the commonly observed overestimation of small posterior probabilities (e.g., the PPV of a diagnostic test) is a by-product of the descriptive format in which medical probabilities are typically communicated.[17]

Although experience-based risk formats are increasingly being studied in the medical context,[18,19] only 1 study to date has investigated how description- and experience-based risk formats affect Bayesian inference in the medical context. Fraenkel and colleagues[20] examined lung cancer risk comprehension and lung cancer screening preferences among patients attending an outpatient pulmonary practice. Patients were randomly assigned to 1 of 3 different formats representing risk information about low-dose computed tomography (LDCT) scans: 1) an experience format, involving a series of slides showing LDCT scans of 250 patients in random order, in addition to numerical information; 2) icon arrays and numbers; or 3) numbers only. Participants were tested for their objective knowledge of risk of lung cancer screening and their preference for screening. Contrary to the authors' hypothesis, results showed that icon arrays accompanied by numbers produced more accurate knowledge than the experience format with numbers and the numbers-only format. In addition, those in the experience format with numbers showed an increased endorsement of screening (even though screening produces a relatively high rate of false-positives). Overall, these findings are thus not consistent with the idea that experienced probabilities necessarily improve risk comprehension or medical decisions among patients. However, the authors note that this may have been due to the complexity and number of scans presented at relatively high speed, which may have prevented effective encoding. In summary, existing evidence of the effectiveness of experienced risks in the medical context is modest and in need of replication.

Age Differences in Medical Decision Making

The study by Fraenkel and colleagues[20] included middle-aged and older adults but did not

systematically examine the impact of age on risk comprehension or choice predisposition. In light of current demographic trends and increased rates of health problems among older adults, a closer examination of age differences is critical.

Normal aging is associated with cognitive slowing and with reductions in working memory, attentional control, and episodic long-term memory[21]—cognitive processes likely involved in Bayesian reasoning. There is also evidence of cohort differences in numeracy and statistical literacy,[22,23] with older adults sometimes faring more poorly in these domains than younger adults. On the other hand, research on frequency processing[24–26] has shown that encoding of frequency information is relatively automatic and that sensitivity to frequencies is largely preserved in old age. There is also some prior evidence of preserved sensitivity to statistical regularities[27] and relatively intact experience-based probability judgments in older adults.[28] Overall, the pattern of preserved and impaired abilities in aging suggests that frequency-based, experiential probability formats may be particularly well suited for enhancing risk comprehension in older adults.

The Current Study

The objective of the current study was to compare the impact of probability format (description v. experience) on Bayesian inference in younger and older adults. In the description condition, modeled after Galesic and others,[6] participants read a statistical summary of the relevant probabilities of a diagnostic test (i.e., base rate of disease, as well as conditional probabilities such as the sensitivity and false-alarm rate of the diagnostic test). However, in departure from Galesic et al.,[6] disease names were fictitious to minimize the influence of prior knowledge—potentially different for younger and older adults—on performance. In the experience condition, modeled after Fraenkel and others,[20] participants viewed the joint distribution of health status and test result. Specifically, participants viewed a slideshow in which they saw a series of patient profiles. Individual patients were shown on the computer screen one at a time. Each patient was characterized by his or her health status (does or does not have disease) and test result (positive or negative test result). The group of patients was a representative sample of the population, in the sense that the relative frequencies of the different disease/test result combinations in the sample equaled the population probabilities. During the slideshow, participants thus had the opportunity to update their prior beliefs about the population on the basis of experienced instances.

We predicted that the experience format would result in more accurate estimates of posterior probabilities (PPV and NPV) in comparison to the description format. Furthermore, age differences in estimation accuracy were expected to be greater in the descriptive format compared with the experience format. This hypothesis was based on the evidence for age-related declines in aspects of fluid intelligence, including speed, working memory, and executive functions.[21] These factors have been shown to play a role in text comprehension[29] and numerical reasoning,[30] both of which likely affect processing of written passages such as those used in the current study and other similar work.[6,18] In contrast, there is little support for a role of working memory in experience-based tasks.[31,32] In addition to testing hypotheses about the impact of probability format on posterior probability estimates, we also aimed to examine whether probability format would have an impact on participants' self-reported confidence as well as their preference to make their own medical decisions or to rely on a physician for making decisions for them.

METHOD

Participants

All participants gave written informed consent for the study, which was approved by the Research Ethics Board at Ryerson University in Toronto, Canada. The final sample included 80 younger adults ($\bar{x}$ = 20.86; ages 17–34 y; 18 male) and 80 older adults ($\bar{x}$ = 75.15; 65–87 y; 30 male). Younger adults were recruited through a participant database at Ryerson University and received partial course credit in exchange for their participation, whereas older adults were recruited through the Ryerson Senior Participant Pool and received $12 in cash.

Design

The study employed a 2 × 2 × 2 mixed design, with age group (younger v. older) and format (description v. experience) as between-subjects factors and disease type (polykronisia, zymbosis) as a within-subjects factor. Half of the participants in each age group were randomly assigned to either the descriptive format or the experience format,

Table 1 Disease Properties

	Polykronisia	Zymbosis
Disease prevalence (%)	2.00	1.00
Test properties (%)		
Sensitivity	100.00	100.00
Specificity	91.84	89
False-alarm rate	8.16	11.11
PPV	20.00	8.33
NPV	100.00	100.00
Frequency (experience format)		
Disease/positive test	2	1
Disease/negative test	0	0
No disease/positive test	8	11
No disease/negative test	90	88

Note: PPV = positive predictive value; NPV = negative predictive value; disease/positive test = number of patients with disease who get positive test result; frequency = number of patients (out of a total of 100) representing each combination of disease status (has disease/does not have disease) and test result (negative/positive).

respectively. On each trial of the task, participants received information about a fictitious disease and its diagnostic test before providing a series of probability judgments. Two trials were administered, corresponding to 2 fictitious diseases, polykronisia and zymbosis (see Table 1). Trial order was random.

Stimuli and Apparatus

E-Prime 2.0 (Psychology Software Tools, Inc.) was used for stimulus presentation and response collection on a 16.0-inch LCD display running 32-bit Windows 7 Enterprise Edition. Viewing distance was approximately 50 cm. Text instructions for the descriptive format appeared in black against a white background, and task stimuli for the experience format appeared in red and blue 18-point Times New Roman font against a white background.

Procedure

In the description format, participants read passages that included information about the prevalence of the disease as well as the sensitivity and the false-positive rate of the test (see the supplemental appendix). All probabilities were presented in natural frequency format (e.g., ''2 out of every 100 people''), as frequencies are easier to understand than probabilities.[6] Participants had up to 7 min to read the statistical summary. Once time was up or they indicated their readiness, participants continued to the test phase. They were asked to estimate disease prevalence, sensitivity, false-positive rate, specificity, PPV, and NPV using a natural frequency response format for each response. Only the estimates of PPV and NPV are relevant to the objectives of the current study.

In the experience format (see Figure 1), participants viewed a series of slides showing a representative sample of 100 fictitious patients who had undergone a screening test for the disease. Patients were presented one at a time for 3 s, separated by a 1-s blank white screen. Including an introductory screen, the slide show of 100 patients lasted approximately 7 min, making the overall exposure time similar to the maximum time available to participants in the description format. Patients were characterized by a combination of disease status (disease or no disease) and test result (positive or negative). The words ''Has Disease'' and ''Positive Test Result'' were presented in red font, whereas the words ''Does Not Have Disease'' and ''Negative Test Result'' were presented in blue font. The colored font served as an additional cue to increase the salience of the different combinations of disease status and test result. The number of patients representing the combinations of disease status (has disease v. does not have disease) and test result (positive v. negative) differed for the two diseases (see Table 1). Participants were prohibited from taking written notes during the slideshow and were discouraged from using rote memorization or other mnemonic techniques. Instead, they were instructed to simply pay attention to the information on the screen. The test phase was the same as in the description condition.

After the computer task, participants completed a battery of background measures and cognitive tests, as well as a self-assessment questionnaire, which prompted participants to rate, using a 5-point scale, their level of confidence and comfort working with numbers, the difficulty of the estimation task, and their belief that their estimates were close to the correct answers. Participants were also asked to indicate on a 10-point scale whether they would prefer to rely on themselves (lowest value on scale: 1) or on a physician (highest value on scale: 10) in making final decisions about their medical care.

Data Analysis

In a first step, the PPV and NPV estimation errors were determined as the absolute (unsigned) difference between participants' estimates and the corresponding correct values. For example, if the correct

Table 2 Sample Characteristics

	Description		Experience	
	Younger (n = 40)	Older (n = 40)	Younger (n = 40)	Older (n = 40)
Sex	8 male	15 male	10 male	15 male
Age (y)	21.40 (4.18)	75.75 (6.05)	20.33 (2.81)	74.55 (6.31)
Age range (y)	17–34	66–87	17–27	66–89
Education (y)[a]	14.55 (2.23)	17.05 (2.34)	13.58 (1.50)	16.60 (2.93)
MMSE	29.30 (0.82)	29.10 (1.11)	29.70 (0.46)	28.60 (4.71)
Digit-symbol[a]	86.47 (14.93)	59.28 (14.27)	86.35 (15.78)	61.22 (12.64)
Numeracy[a]	9.60 (1.81)	8.27 (2.35)	9.48 (1.81)	8.75 (2.46)
Positive mood[a]	24.70 (7.81)	33.73 (6.12)	24.03 (7.36)	31.85 (8.77)
Negative mood[a]	14.62 (4.17)	12.33 (4.26)	14.00 (6.46)	11.57 (2.11)
Depression[a,b]	7.85 (5.03)	4.80 (4.17)	5.15 (5.53)	2.55 (3.27)
Anxiety[a]	7.65 (4.64)	3.25 (3.59)	5.50 (4.92)	3.35 (3.72)
Stress[a,b]	10.60 (4.92)	8.75 (5.02)	8.95 (6.75)	5.25 (5.33)

Note: MMSE = score on the Mini-Mental State Examination[33]; numeracy = score on a scale that included the 11-item numeracy scale[34] and 1 coin-toss item[35]; positive mood and negative mood: scores on the Positive and Negative Affect Schedule[36]; depression, anxiety, and stress: scores on the Depression Anxiety Stress Scales (DASS-21[37]). Standard deviations are shown in parentheses.
[a]Significant age difference ($P < 0.01$).
[b]Significant format effect ($P < 0.01$).

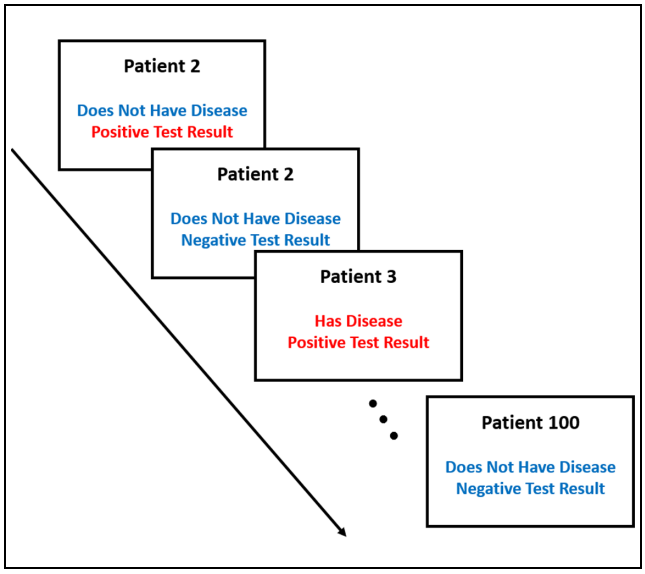

Figure 1 Schematic of the slideshow used in the experience format. Participants view a sequence of 100 patient cases representing combinations of disease status (has disease/does not have disease) and test result (negative/positive). The frequency of each combination is shown in Table 2.

PPV was 20% and the participant's estimate was 14%, then the estimation error was 6%. We used unsigned, rather than signed, errors in light of the restricted range of possible errors for PPVs and NPVs. We also ran the analyses on mean squared errors. To foreshadow, these analyses yielded the same pattern of results (direction and statistical significance levels of the main effects and interactions), so for simplicity, we report only the absolute errors.

Separate mixed analyses of variance (ANOVAs) were conducted on PPV and NPV estimation errors, with age group (younger v. older) and format (description v. experience) as between-subjects factors and disease (polykronisia v. zymbosis) as a within-subject factor. In light of violations of normality in the data, we also conducted Mann-Whitney U tests on the estimation errors. Each U test assessed differences between the levels of one factor (e.g., younger v. older adults), collapsing across the levels of the other factors. To foreshadow, the results of these pairwise nonparametric tests mirrored the main effects found in the ANOVAs and are therefore not reported separately.

RESULTS

Participant Characteristics

Demographic, cognitive, and affective characteristics are shown in Table 2. A series of between-subjects ANOVAs on these characteristics, with factors age group (young v. old) and format (description v. experience), revealed several significant effects of age group, some significant effects of format, and no significant interactions.

Table 3 Posterior Probability Estimation Errors and Self-Assessment Responses

	Description		Experience	
	Younger	Older	Younger	Older
PPV error				
Polykronesia	58.30 (29.01)	51.67 (31.42)	19.15 (26.53)	7.78 (18.07)
Zymbosis	64.53 (37.77)	57.93 (39.51)	10.81 (21.22)	9.29 (19.64)
NPV error				
Polykronesia	26.03 (39.03)	26.03 (39.03)	5.51 (16.15)	8.45 (18.48)
Zymbosis	15.83 (29.18)	27.97 (39.13)	2.84 (4.92)	8.39 (16.70)
Self-assessment				
Confidence	2.75 (1.03)	2.87 (1.20)	3.43 (0.93)	3.28 (1.28)
Difficulty	3.75 (0.87)	3.75 (0.81)	3.05 (1.01)	3.10 (1.11)
Belief in accuracy	3.20 (0.85)	3.20 (0.82)	3.55 (1.06)	3.50 (0.78)
Self v. physician	6.45 (2.72)	6.63 (2.68)	6.38 (2.39)	5.75 (2.44)

Note: PPV error = absolute difference between estimated and true positive predictive value; NPV error = absolute difference between estimated and true negative predictive value; confidence = rated confidence in working with numbers (on a scale of 1–5); difficulty = rated difficulty of the estimation task (on a scale of 1–5); belief in accuracy = belief that estimates were correct (on a scale of 1–5); self v. physician = preference for self- v. physician-made medical decisions (on a scale of 1–10).

The age effects largely mirrored typical patterns for cognitive and affective measures reported in the psychological literature on healthy aging.[21,38] Even though older adults ($\bar{x}$ = 16.83) had more years of education than younger adults ($\bar{x}$ = 14.06), $F(1, 156) = 57.29$, $P < 0.01$, $\eta_p^2 = 0.27$, they scored lower on the digit-symbol coding test than younger adults ($\bar{x}$ = 60.25 v. $\bar{x}$ = 86.41), $F(1, 156) = 131.07$, $P < 0.01$, $\eta_p^2 = 0.46$, and had lower numeracy scores than younger adults ($\bar{x}$ = 8.51 v. $\bar{x}$ = 9.54), $F(1, 156) = 9.27$, $P < 0.01$, $\eta_p^2 = 0.06$. Older adults reported higher positive mood than younger adults ($\bar{x}$ = 32.79 v. $\bar{x}$ = 24.36), $F(1, 156) = 49.50$, $P < 0.01$, $\eta_p^2 = 0.24$; reported lower negative mood than younger adults ($\bar{x}$ = 11.95 v. $\bar{x}$ = 14.31), $F(1, 156) = 10.93$, $P < 0.01$, $\eta_p^2 = 0.07$; scored lower on depression than younger adults ($\bar{x}$ = 3.68 v. $\bar{x}$ = 6.50), $F(1, 156) = 15.22$, $P < 0.01$, $\eta_p^2 = 0.09$; scored lower on anxiety than younger adults ($\bar{x}$ = 3.30 v. $\bar{x}$ = 6.58), $F(1, 156) = 23.72$, $P < 0.01$, $\eta_p^2 = 0.13$; and scored lower on stress than younger adults ($\bar{x}$ = 7.00 v. $\bar{x}$ = 9.78), $F(1, 156) = 9.99$, $P < 0.01$, $\eta_p^2 = 0.06$.

Unexpectedly, the ANOVAs also revealed 2 significant effects of format. Depression scores were higher for participants in the description condition ($\bar{x}$ = 6.33) than for participants in the experience condition ($\bar{x}$ = 3.85), $F(1, 156) = 11.68$, $P < 0.01$, $\eta_p^2 = 0.07$. Similarly, stress scores were higher for participants in the description condition ($\bar{x}$ = 9.68) than for participants in the experience condition ($\bar{x}$ = 7.10), $F(1, 156) = 8.60$, $P < 0.01$, $\eta_p^2 = 0.05$. Including depression and stress as covariates in the analyses of task performance outcomes did not change the pattern of results, so we report only the covariate-free analyses for simplicity.

Task Performance

PPV

There were no significant main effects of age or disease on PPV estimation errors (see Table 3). However, there was a significant effect of format, $F(1, 156) = 131.02$, $P < 0.01$, $\eta_p^2 = 0.46$, indicating that estimation errors were significantly smaller in the experience format ($\bar{x}$ = 12.64%) than in the description format ($\bar{x}$ = 58.11%). The format effect was qualified by a significant Disease × Format interaction, $F(1, 156) = 9.13$, $P < 0.001$, $\eta_p^2 = 0.06$. Follow-up ANOVAs, conducted separately for each disease, showed that the format effect was significant for polykronisia, $F(1, 156) = 88.39$, $P < 0.001$, $\eta_p^2 = 0.36$, but that it was larger for zymbosis, $F(1, 156) = 114.65$, $P < 0.001$, $\eta_p^2 = 0.42$.

Because absolute estimation errors do not indicate whether the errors reflect over- or underestimation, we also examined the distributions of raw (signed) PPV estimation errors (see panels A and B in Figure 2). The distributions showed a tendency to underestimate the true PPVs in the experience format and a tendency to overestimate PPVs in the description format. In addition, errors were tightly clustered around the true value in the experience format, whereas they ranged more widely in the description format, for both diseases.

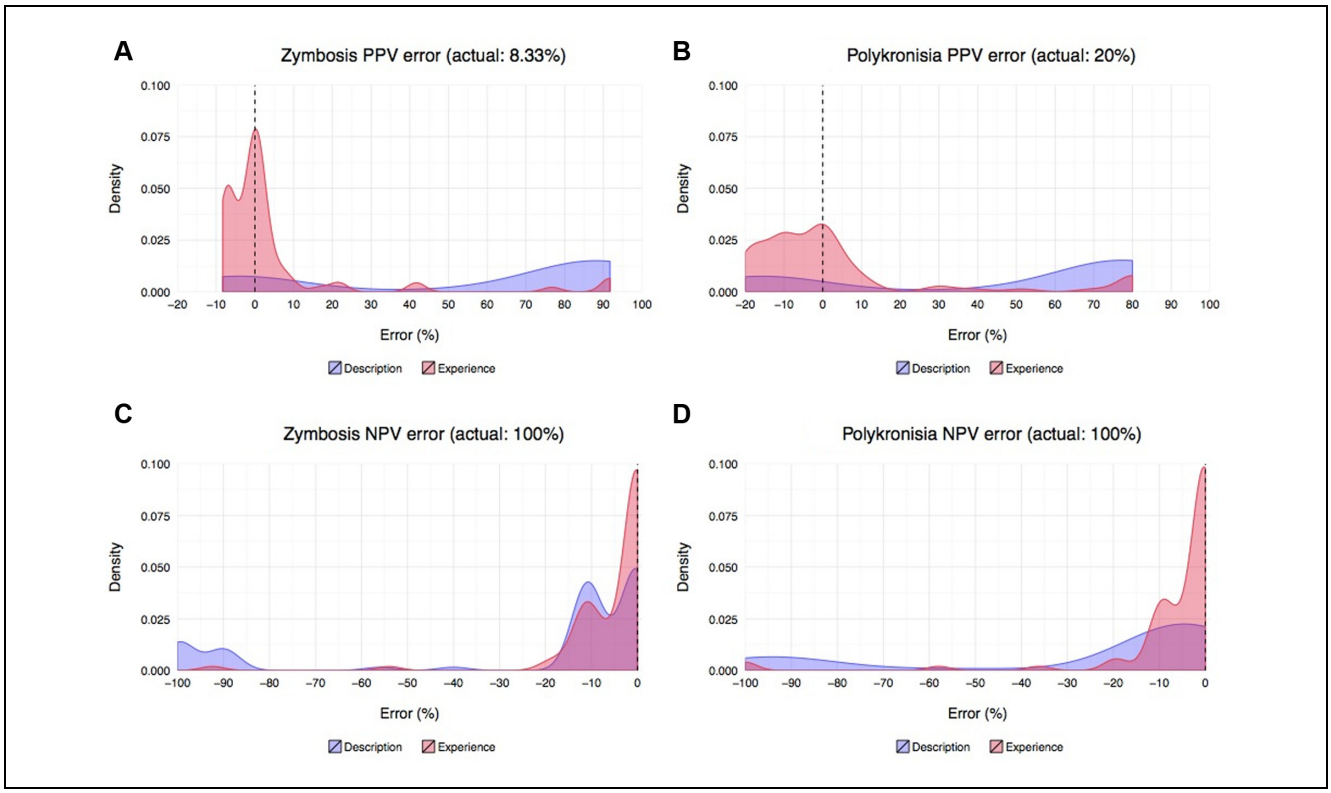

Figure 2 Visualization of the distribution of estimation errors for both the positive predictive value (PPV) and the negative predictive value (NPV), shown separately for the two fictitious diseases (zymbosis, polykronisia). Zero on the x-axis indicates no error (i.e., the estimate was identical to the actual PPV or NPV). Negative values indicate underestimation of the actual PPV or NPV, and positive values indicate overestimation of the actual PPV or NPV. Two older adults in the description condition gave out-of-range estimates (e.g., 111/100), which were not included here.

NPV

There were no significant main effects of age or disease on NPV estimation errors nor any significant interactions (see Table 3). However, there was a significant effect of format, $F(1, 156) = 23.31$, $P < 0.01$, $\eta_p^2 = 0.13$, indicating that estimation errors were significantly smaller in the experience format ($\bar{x} = 6.30\%$) than in the description format ($\bar{x} = 22.87\%$). Figure 2c, d illustrates these patterns.

Self-assessment

ANOVAs on self-assessment responses (see Table 3) with factors age group (young v. old) and format (description v. experience) revealed significant main effects of format but no effects of age group and no significant interactions. Participants in the description format ($\bar{x} = 2.81$) reported less confidence working with numbers compared with those in the experience format ($\bar{x} = 3.35$), $F(1, 156) = 9.22$, $P < 0.01$, $\eta_p^2 = 0.056$. Participants in the description format ($\bar{x} = 3.75$) also reported more difficulty estimating properties of a diagnostic test compared with those in the experience format ($\bar{x} = 3.08$), $F(1, 156) = 19.95$, $P < 0.001$, $\eta_p^2 = 0.113$. In addition, those in the experience format ($\bar{x} = 3.53$) expressed a stronger belief in the accuracy of their estimates than those in the description format ($\bar{x} = 3.20$), $F(1, 156) = 5.37$, $P < 0.01$, $\eta_p^2 = 0.033$. There were no effects of age group or format on participants' ratings of their preference for a decision maker (self or physician), with a slight overall tendency to prefer relying on a physician ($\bar{x} = 6.30$).

Influence of Numeracy

In light of prior reports of numeracy effects on Bayesian inference in the medical context,[6] we examined nonparametric correlations between numeracy and PPV and NPV estimation errors,

separately for each participant group. None of the correlations reached significance, $Ps > 0.05$.

DISCUSSION

The goal of this study was to assess the effect of different probability formats on estimates of the PPVs and NPVs of medical test results in healthy younger and older adults. We hypothesized that an experience format, involving sequential encoding of representative patient cases, would result in more accurate estimates of predictive values than a description format involving verbal summaries of relevant statistics.[6] Consistent with this hypothesis, we found a significant format effect on estimation errors, which were significantly smaller in the experience format, compared with the description format. Younger and older adults showed similar effects of probability format (and similar overall performance levels). There was no evidence for a relationship between numeracy and estimation accuracy. Finally, despite their superior performance on the estimation task and higher self-reported belief in the accuracy of their estimates, participants in the experience condition did not indicate a stronger preference for making their own medical decisions than participants in the description condition.

Description-Experience Gap in Bayesian Inference

The results in our description condition are in line with prior studies that have shown Bayesian inference to be difficult when relevant information is presented descriptively.[6,7] A direct comparison between our results and those of Galesic and others[6] is not possible because the authors[6] did not report exact PPV estimates and did not assess NPV estimates. However, our results were qualitatively similar to those of Galesic and others,[6] as we observed significant estimation errors in the description condition, in both age groups.

The current results depart somewhat from those reported by Fraenkel and others,[20] who presented patients with information about lung cancer screening and found that a descriptive format (icon arrays) produced better comprehension and choice preference outcomes than an experience format. However, as noted by Fraenkel and others,[20] the slideshow employed in their study may have been too complex to permit effective encoding (250 slides at a rate of 1 s/slide, each slide featuring text, an unfamiliar CT scan, and 1 of 3 colors that represented patient disease status and test result). In contrast, in the current study, participants viewed 100 slides at a rate of 3 s/slide, and each slide included only 2 pieces of information (patient disease status and test result). It is possible that encoding of the relative frequencies of specific disease/diagnosis combinations in Fraenkel and others'[20] experience condition was hindered by the fast presentation rate and the high attentional demands of the slideshow, which may have undermined effective encoding. In this sense, the experience format employed in the current study may have provided more "description" than that in the study by Fraenkel and others.[20]

Consistent with observations in the risky-choice literature showing that rare outcomes affect choices as if they are overweighted in decisions from description,[14] participants in the description condition tended to overestimate the true PPVs. In contrast, participants in the experience condition tended to underestimate the true PPVs (although to a far lesser extent). However, it should be noted that we use the term *estimation* loosely. Posterior probability judgments can be made using strategies and heuristics (e.g., anchoring) that fall short of a strict definition of probability estimates.[39,40] Future research should adopt a more fine-grained approach to shed light on the specific processes younger and older adults use to arrive at posterior probability judgments.

An interesting question is whether this description-experience gap in probabilistic inference affects subsequent attitudes and decisions. Responses to the self-assessment questions revealed a description-experience gap in participants' confidence in their own estimates but not in their preference to rely on a physician. Future studies should examine whether subjective decision-making competence can be enhanced by providing participants with informative feedback following experience-based training trials.

The Role of Age and Numeracy

Contrary to our second hypothesis, we observed no significant age differences in the accuracy of estimated PPVs and NPVs. On the assumption that processing the verbal summaries in the description condition would tap cognitive abilities such as working memory and executive control, abilities known to decline with age, we had expected older adults to perform more poorly than younger adults in this condition. However, the null effect of age on Bayesian inference in the description condition

mirrors prior findings by Galesic and colleagues,[6] who also found no significant age differences in predictive-value estimation from description, despite an age difference in numeracy. It should be noted that a common limitation in both studies is the reliance on convenience samples of healthy older adults with relatively high levels of education. More diverse samples may be needed to demonstrate age-related declines in the ability to process described probabilities.

Limitations and Future Directions

The current study had several limitations. First, despite random assignment of participants to the description and experience conditions, there were unexpected differences between these groups. Specifically, participants in the description condition scored higher on measures of depression and stress compared with those in the experience condition. Since the cognitive and affective measures were assessed at the end of the session, *after* the probability estimation task, the group differences in negative affect may have resulted from the experimental manipulations. However, without preexperiment baseline measures, this possibility could not be tested directly. In future studies, it would be important to assess pre- and posttask affect to examine whether exposure to described probabilistic information induces stress and negative mood. Given the detrimental effect of stress on cognitive function, including decision0-making performance,[41] this question has obvious clinical relevance.

A second limitation concerns the differences between description and experience formats. Although the 2 format conditions were matched on many relevant aspects (e.g., response modalities used during the test phase), there were also several differences that may have affected the results. For example, the time spent reading the verbal summaries in the description condition was self-paced (with an upper limit), whereas the slideshow in the experience condition was experimenter paced. By necessity, there were also differences in the physical properties of the stimuli (verbal summaries and slideshows composed of words in varying colored font), and it is unclear which of these may have affected performance. In addition, participants in the description condition were presented with marginal (base-rate) and conditional probability information, whereas participants in the experience condition were presented with joint distributions (i.e., disease status and test result). Therefore, participants did not have access to the same information. To increase the similarity of information across conditions, it would be useful in future studies to add a 2×2 table with the disease status and test result in the description condition to present joint distributions that were also provided in the experience condition. Pinpointing the ''active ingredient'' underlying the format effect on Bayesian inference will require more fine-grained manipulations of encoding conditions in future studies.

Another limitation of the current study design involved the use of only 2 fictitious diseases. A deeper understanding of the mechanisms underlying the ''experience advantage,'' as well as its practical applicability, will require testing a broader range of denominators (e.g., 500 v. 100 patients in slideshow), disease prevalences, test sensitivities, and test specificities. To increase the ecological validity of the current findings, it will also be important to establish their replicability with real diseases and among patients facing actual medical decisions.

CONCLUSION

This study is the first to provide evidence for a significant increase in comprehension of medical test results (positive and negative predictive values) following exposure to experienced probabilities. It supports Hogarth and Soyer's[17] suggestion that sequential observation of representative instances can serve as an attractive complement to description-based methods and that this approach holds promise for younger as well as older decision makers.

ACKNOWLEDGMENTS

We thank Dr. Pete Wegier for his comments and suggestions throughout this project. We also thank Ryan Marinacci and Ryan S. Williams for their assistance with data entry and management.

REFERENCES

1. O'Connor AM, Bennett C, Stacey D, et al. Do patient decision aids meet effectiveness criteria of the international patient decision aid standards collaboration? A systematic review and meta-analysis. Med Decis Making. 2007;27:554–74.

2. Reyna VF, Nelson WL, Han PK, Dieckman NF. How numeracy influences risk comprehension and medical decision making. Psychol Bull. 2009;135:943–73.

3. Sheridan SL, Harris RP, Woolf SH. Shared decision making about screening and chemoprevention: a suggested approach from the U.S. Preventive Services Task Force. Am J Prev Med. 2004;26:56–66.

4. Simon D, Loh A, Harter M. Measuring (shared) decision-making: a review of psychometric instruments. Z Arztl Fortbild Qualitatssich. 2007;101:259–67.

5. Broadbent E, Petrie KJ, Ellis CJ, Anderson J, Gamble G, Anderson D, Benjamin W. Patients with acute myocardial infarction have an inaccurate understanding of their risk of a future cardiac event. Intern Med J. 2006;36:643–7.

6. 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;29:368–71.

7. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychol Sci Public Interest. 2007;8:53–96.

8. Galesic M, Garcia-Retamero R. Graph literacy: a cross-cultural comparison. Med Decis Making. 2011;31:444–57.

9. Garcia-Retamero R, Cokely ET. Communicating health risks with visual aids. Curr Dir Psychol. 2013;22:392–9.

10. Garcia-Retamero R, Galesic M. Who profits from visual aids: overcoming challenges in people's understanding of risks. Soc Sci Med. 2010;70:1019–25.

11. Knight F. Risk, Uncertainty and Profit. Boston: Houghton Mifflin; 1921.

12. Gonzalez C, Dutt V. Instance-based learning: integrating sampling and repeated decisions from experience. Psychol Rev. 2011;118:525–51.

13. Hogarth RM, Einhorn HJ. Order effects in belief updating: the belief-adjustment model. Cogn Psychol. 1992;24:1–55.

14. Hertwig R, Erev I. The description-experience gap in risky choice. Trends Cogn Sci. 2009;13:517–23.

15. Erev I, Ert E, Roth AE, et al. A choice prediction competition: choices from experience and from description. J Behav Dec Making. 2010;23:15–47.

16. Hertwig R, Barron G, Weber EU, Erev I. Decisions from experience and the effect of rare events in risky choice. Psychol Sci. 2004;15:534–39.

17. Hogarth R, Soyer E. Providing information for decision making: contrasting description and simulation. J Appl Res Mem Cogn. 2015;4:221–8.

18. Tyszka T, Sawicki P. Affective and cognitive factors influencing sensitivity to probabilistic information. Risk Anal. 2011;31:1832–45.

19. Lejarraga T, Pachur T, Frey R, Hertwig R. Decisions from experience: from monetary to medical gambles. J Behav Decis Mak. 2015;29:67–77.

20. Fraenkel L, Peters E, Tyra S, Oelberg D. Shared medical decision making in lung cancer screening: experienced versus descriptive risk formats. Med Decis Making. 2015;36:518–25.

21. Park DC. The basic mechanisms accounting for age-related decline in cognitive function. In:Park DC, Schwarz N, eds. Cognitive Aging: A Primer. New York: Psychology Press; 2000. p 3–21.

22. Bruine de Bruin W, Vanderklaauw W, Downs JS, Fischhof B, Topa G, Armantier O. Expectations of inflation: the role of demographic variables, expectation formation, and financial literacy. J Consum Aff. 2010;44:381–402.

23. Weller J, Dieckman NF, Tusler M, Mertz CK, Burns WJ, Peters E. Development and testing of an abbreviated numeracy scale: a Rasch analysis approach. J Behav Decis Mak. 2013;26:198–212.

24. Hasher L, Zacks RT. Automatic and effortful processes in memory. J Exp Psychol Gen. 1979;108:356–8.

25. Hasher L, Zacks RT. Working Memory, Comprehension, and Aging: A Review and a New View. San Diego (CA): Academic Press; 1988.

26. Zacks RT, Hasher L. Frequency processing: a twenty-five year perspective. In:Sedlmeier P, Betsch T, eds. Frequency Processing and Cognition. New York: Oxford University Press; 2002. p 21–36.

27. Campbell KL, Healey MK, Lee MMS, Zimerman S, Hasher L. Age differences in visual statistical learning. Psychol Aging. 2012;27:650–6.

28. Spaniol J, Bayen UJ. Aging and conditional probability judgments: a global matching approach. Psychol Aging. 2005;20:165–81.

29. Federmeier KD, Kutas M. Aging in context: age-related changes in context use during language comprehension. Psychophysiology. 2005;42:133–41.

30. Hodzik S, Lemaire P. Inhibition and shifting capacities mediate adults' age-related differences in strategy selection and repertoire. Acta Psychol. 2011;137:335–44.

31. Frey R, Mata R, Hertwig R. The role of cognitive abilities in decisions from experience: age differences emerge as a function of choice set size. Cognition. 2015;142:60–80.

32. Wulff DU, Hills TT, Hertwig R. How short- and long-run aspirations impact search and choice in decisions from experience. Cognition. 2015;144:29–37.

33. Folstein M, Folstein SE, McHugh PR. ''Mini-Mental State'': a practical method for grading the cognitive state of patients for the clinician. J Psychiat R. 1975;12:189–98.

34. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. Med Decis Making. 2001;21:37–44.

35. Shwartz LM, Woloshin S, Black WC, Welch GH. The role of numeracy in understanding the benefit of screening mammography. Ann Intern Med. 1997;127:966–71.

36. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol. 1988;54:1063–70.

37. Lovibond SH, Lovibond PF. Manual for the Depression, Anxiety, Stress Scales. 2nd ed. Sydney (Australia): School of Psychology, University of New South Wales, Psychology Foundation; 1995.

38. Scheibe S, Cartensen LL. Emotional aging: recent findings and future trends. J Gerontol B Psychol Sci Soc Sci. 2010;2:135–44.

39. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. Psychol Rev. 1995;102:684–704.

40. Achtziger A, Alos-Ferrer C, Hugelschafer S, Steinhauser M. The neural basis of belief updating and rational decision making. Soc Cogn Affect Neurosci. 2014;9:55–62.

41. Starcke K, Brand M. Decision making under stress: a selective review. Neurosci Biobehav Rev. 2012;36:1228–48.