Statistical Literacy in Obstetricians and Gynecologists

Britta L. Anderson, Gerd Gigerenzer, Scott Parker, Jay Schulkin

Abstract: The Obstetrician-Gynecologist Statistical Literacy Questionnaire (OGSLQ) was designed to examine physicians' understanding of various number tasks that are relevant to obstetrician-gynecologists (ob-gyns) practice. Forty-seven percent of the nationally representative, practicing ob-gyns responded. Physicians did poorly on the questions about numerical facts (e.g., number of women living with HIV/AIDS), better on questions about statistical concepts (e.g., incidence, prevalence), and best on questions about numerical relationships (e.g., convert frequency to percentage) with 0%, 7%, 36%, answering all correctly, respectively. Only 19% correctly estimated the number of U.S. women with cancer. Sixty-six percent were able to use sensitivity and specificity to choose a test option. Around 90% could translate between frequency and probability formats. Forty-nine percent of respondents were able to calculate the positive predictive value of a mammography screening test. Physicians lack some understanding of statistical literacy. It is important that we monitor physicians' statistical literacy and provide training to students and physicians.

Keywords
academic
evidence-based practice/guidelines
practice setting/focus
research
research-evaluation topics

Purpose

Numeracy or statistical literacy is the ability to use and understand numbers. Patients' numeracy skills and its effects on health decisions have been studied for years (Gigerenzer, Gaissmaier, Kurz-Mielke, Schwartz, & Woloshin, 2007; Nelson, Reyna, Fagerlin, Lipkus, & Peters, 2008; Peters et al., 2006; Reyna, Nelson, Han, & Dieckmann, 2009). The results suggest that more numerate patients make better health decisions. Unfortunately, physicians' numeracy has been overlooked as it is often thought that the problem with statistical literacy is restricted to patients. However, both patients and physicians can contribute to the problems with statistical illiteracy (Gigerenzer, 2010). Like patients, physicians have been found to make compromised judgments when statistical results are not presented in transparent formats (Covey, 2007; Forrow, Taylor, & Arnold, 1992). The term "collective statistical illiteracy" is used to describe patients' and physicians' lack of understanding health literacy (Gigerenzer, 2010).

Review of Literature

Previous studies assessing physician numeracy have used scales that were developed for the general population such as the three-question Schwartz numeracy scale (Schwartz, Wobshin, Black, & Welch, 1997) or the subjective numeracy scale (SNS; Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry, & Smith, 2007). When the Schwartz scale was administered, a large majority of physicians answer all three questions correctly (Anderson et al., 2011). The SNS requires that participants rate their ability to use numbers and their preferences for numerical information rather than solve problems. The SNS has been used in at least one study with physicians. Physicians rated themselves as having very high numeracy skills. The mean score on the SNS was 4.9 of 6 (6 being high numerate) with over one quarter of the sample scoring 6 and over 50% of the sample scoring 5 (Anderson et al., 2011). Physicians' performance on these previous scales is not ideal from the point of view of either patients or researchers; patients would prefer that every physician get all the questions correct, and researchers would prefer to have a scale that discriminates better between physicians.

In this article, we assess physicians' statistical literacy using the Obstetrician-Gynecologist Statistical Literacy Questionnaire (OGSLQ). We use the term "statistical literacy" to emphasize the inclusion of the statistical concepts that are assessed in this questionnaire. Our investigation addresses three questions: (1) How well do physicians perform on the OGSLQ and the different question types? (2) To what degree are the three types of questions associated with each other? (3) How does the OGSLQ compare to the Schwartz and Lipkus Numeracy Scales?

Study Design and Methods

Sample

Two-hundred practicing obstetrician-gynecologists (ob-gyns) who are members of the American College of Obstetricians and Gynecologists were invited to participate. Physician invitees were quasi-randomly sampled from the Collaborative Ambulatory Research Network (CARN),

Journal for Healthcare Quality
Vol. 36, No. 1, pp. 5–17

which consists of Fellows or Junior Fellows who volunteer to complete three to five survey studies each year.

Participants were sent a survey, a cover letter explaining the study, and a postage-paid return envelope. The survey contained demographic questions, the Schwartz and Lipkus Numeracy Scales, the OGSLQ (Table 1), and a list of topics where physicians indicated which clinical issues emerge in practice once per month or more (e.g., HIV/AIDS, HPV, Chagas, etc.).

Measures

Schwartz Numeracy Scale

The Schwartz Numeracy Scale (Schwartz, Woloshin, Black, & Welch, 1997) is made up of three questions: (1) a conversion from a percentage to a frequency, (2) conversion from a frequency to a percentage, and (3) an estimation of how many heads there will be in 1,000 coins flips. It has been found to have good internal reliability ($\alpha = .56–.80$) and test–retest reliability ($r = .72$; Reyna et al., 2009). Questions were scored as correct (1) or incorrect (0) and scores were summed for a total score.

Lipkus Numeracy Scale

The Lipkus Numeracy Scale (Lipkus, Samsa, & Rimer, 2001) is an expansion of the Schwartz measure. It has been found to have good internal reliability ($\alpha = .70–.75$; Reyna et al., 2009). One question (8A) was removed from the Lipkus scale to eliminate redundancy and reduce responder burden. Questions were scored as correct (1) or incorrect (0) and scores were summed for a total score.

The Obstetrician-Gynecologist Statistical Literacy Questionnaire (OGSLQ)

The OGSLQ questions were designed to represent clinically relevant numerical facts, concepts, and relations in order to examine physicians' understanding of health statistics (see Table 1). Some questions are specific to obstetrics-gynecology; others are relevant to many types of medical practice. The OGSLQ was created by the authors. Questions were grouped into one of three question types (Fact Questions, Concept Questions, or Relation Questions) based on the type of task that the question required.

Five questions assess physicians' knowledge of numerical facts (Fact Questions). Questions 1 and 2 assess physicians' basic knowledge of statistics that are relevant to medical practice (e.g., number of women with cancer in the United States). Questions 3–5 assess physicians' knowledge about the increased or decreased risk of various health outcomes given common conditions.

Four questions assess physicians' knowledge of statistical concepts (Concept Questions). Questions 7–9 assess understanding of statistical terms (prevalence, incidence, specificity, and sensitivity). Question 13 assesses the ability to calculate number needed to treat.

Five questions assess physicians' understanding of numerical relationships (Relation Questions). Question 6 assesses ability to compare and convert between relative and absolute risk reductions and question 10 assesses ability to interpret a risk-reduction statement. Questions 11 and 12 assess ability to convert probabilities to frequencies and vice versa. Finally, question 14 assesses ability to calculate a positive predictive value (Bayesian posterior probability).

Each question was scored as correct (scored as 1) or incorrect (scored as 0). Blank questions were scored as incorrect. To compute the OGSLQ score, the score on all the questions were summed. The total possible scores on this scale are 0–18 (because questions 4 and 5 consist of several subquestions).

A small pilot test was conducted to gauge the difficulty and obtain feedback from obgyns. Along with completing the questionnaire, physicians were asked: (1) "Did you find any questions to be unlike numerical situations that occur in your practice environment or very difficult? If yes, which ones?" and (2) "Are there any numeric tasks or situations that you encounter on a regular basis that are not included?" Pilot responses did not indicate that any of the questions were irrelevant and/or too difficult, or that there were any relevant tasks that were not included.

Data Analysis

Data were analyzed using a personal computer-based version of SPSS 17.0 (SPSS Inc., Chicago, IL). Descriptive and frequency data were computed for primary analysis. Cronbach's alpha was used to determine each scale's internal reliability. Pearson's r was used for correlation analyses. Chi-squared was used to assess group differences.

Table 1. Obstetrician-Gynecologist Statistical Literacy Questionnaire (OGSLQ)

The OGSLQ

The following questions reflect numerical situations that occur in obstetrics and gynecology. Five questions ask about numerical facts, four ask about your understanding of numerical information, and five ask for a calculation. Some questions are hypothetical while others ask about specific medical conditions. Regardless of your specialty, please answer all of the questions to the best of your ability. Do not use the Internet or other sources to help you. Feel free to use the margins of the survey to calculate answers.

1. Roughly how many women in the United States are living with HIV/AIDS?
   A. 50,000 women [9%]
   B. 125,000 women [7%]
   C. 300,000 women [18%]
   D. 425,000 women [20%]
   E. 750,000 women [14%]
   F. 1,025,000 women [27%]
   Left Question Blank [5%]

2. Of every 100,000 U.S. women of all ages, roughly how many are diagnosed with cancer each year?
   A. 50 of 100,000 women [12%]
   B. 125 of 100,000 women [17%]
   C. 200 of 100,000 women [14%]
   D. **400 of 100,000 women [19%]**
   E. 750 of 100,000 women [13%]
   F. 900 of 100,000 women [17%]
   Left Question Blank [8%]

3. Does mammography reduce the chances of a patient getting breast cancer?
   Yes [10%] **No [90%]**

4. Do the following risks increase or decrease with the use of hormone replacement therapy?
   Risk of heart disease increase [60%] decrease [11%] **no difference [23%]** blank [6%]
   Risk of osteoporosis increase [0%] **decrease [100%]** no difference [0%] blank [0%]
   Risk of breast cancer **increase [89%]** decrease [0%] no difference [10%] blank [1%]
   Risk of colon cancer increase [1%] **decrease [81%]** no difference [18%] blank [0%]

5. Among 1,000, 55-year-old women who never smoked, within the next 10 years eight will die of heart disease and two of lung cancer. Roughly, what are the numbers for women who smoke?
   Among 1,000, 55-year-old women who smoke . . . (*circle one for each*)
   How many will die of heart disease?
   10 [3%] 14 [3%] 17 [19%] **20 [15%]** 24 [29%] 26 [22%] Left Question Blank [9%]
   How many will die of lung cancer?
   4 [5%] **6 [9%]** 8 [13%] 10 [25%] 12 [9%] 14 [31%] Left Question Blank [10%]

6. All other things equal, which drug would be best for your patient?
   A. Drug A: Reduces the likelihood of death by 25%. [1%]
   B. Drug B: Reduces the likelihood of death by 50%. [11%]
   C. Drug C: Reduces the number of deaths from eight in 1,000 to two in 1,000 [82%]
   D. Drug D: Reduces the number of deaths from eight in 1,000 to four in 1,000 [2%]
   E. Drug E: Increases the number of survivors from 992 in 1,000 to 997 in 1,000 [3%]
   F. Drug F: Increases the number of survivors from 992 in 1,000 to 996 in 1,000 [0%]
   Left Question Blank [1%]

7. One study estimated that there are 63–315 new cases of congenital Chagas in the entire United States each year. "63 to 315 new cases" is the estimated ___ of congenital Chagas disease in the United States (current population = 304,059,724).
   A. Sensitivity [1%]
   B. Specificity [0%]
   C. Prevalence [15%]
   D. Mortality [3%]
   E. Incidence [74%]
   F. There is not enough information provided (What information do you need?_____) [5%]
   Left Question Blank [2%]

Table 1. Continued

The OGSLQ

8. The following was reported in the journal *Sexually Transmitted Infections* in 2001 about genital herpes (HSV-2):
"Methods: Sera from 869 cohort members were tested using an indirect IgG enzyme-linked immunoassay specific to the HSV-2 glycoprotein G. Results: In all, 96 participants (11%) were seropositive for HSV-2, including at least 56 who seroconverted after their 21st birthday."
96 of 869 is the ____ of HSV-2 in this sample.
A. Sensitivity [6%]
B. Specificity [2%]
C. Prevalence [68%]
D. Mortality [0%]
E. Incidence [19%]
F. There is not enough information provided (What information do you need?_____) [2%]
Left Question Blank [3%]

9. A study published in the *New England Journal of Medicine* was done to determine whether testing for DNA of oncogenic human papillomaviruses (HPV) is superior to the Papanicolaou (Pap) test for cervical-cancer screening. The researchers report that:
"The sensitivity of HPV testing for cervical intraepithelial neoplasia of grade 2 or 3 was 94.6% (95% confidence interval [CI], 84.2–100), whereas the sensitivity of Pap testing was 55.4% (95% CI, 33.6–77.2; p = 0.01). The specificity was 94.1% (95% CI, 93.4–94.8) for HPV testing and 96.8% (95% CI, 96.3–97.3; p = 0.001) for Pap testing."
You can only use one test and your patient wants to use the test that, according to this study, is the least likely to have a false positive result. Which test is least likely to yield a false positive result?
A. HPV testing [25%]
B. Pap testing [66%]
C. Either (they have the same likelihood of a false positive) [3%]
D. There is not enough information provided (What information do you need?_____) [4%]
Left Question Blank [2%]

10. Of 1,000 people with Disease X, 300 experience a particular severe symptom. A new treatment is found to reduce the chance that a person with Disease X will experience the particular severe symptom by 33%. This means that of 1,000 people with Disease X who use the drug about _____ will experience the particular severe symptom.
A. 33 [0%]
B. 99 [2%]
C. 100 [16%]
D. 200 [77%]
E. 333 [2%]
F. There is not enough information provided (What information do you need?_____) [0%]
Left Question Blank [3%]

11. Among your patients, 25% have a particular condition during pregnancy. Among women who have this condition during pregnancy 60% will need to remain in the hospital longer due to complications. Your patient asks you how many of 100 pregnant patients have to stay longer in the hospital for complications caused by the particular condition. What should you say?:
A. 15 [89%]
B. 20 [1%]
C. 25 [0%]
D. 30 [0%]
E. 60 [8%]
F. There is not enough information provided (What information do you need?_____) [0%]
Left Question Blank [2%]

Results

Sample Demographics

A total of 94 of the 200 sampled ob-gyns returned a completed survey, a 47% response rate. On average, participants were 52-year old ($SD = 10$) and half were female. Seventy-three percent of respondents were general ob-gyns, 18% practice gynecology only, and 9% reported other specialty.

Table 1. Continued

The OGSLQ

12. In a sample of 100 women, 30 are given a particular treatment and in addition, one in three women who are on that treatment have blood clots. What percent of that sample of 100 women are on the treatment and have blood clots? _____
A.3% [0%]
B. 6% [2%]
C.10% [93%]
D.20% [0%]
E.33% [3%]
F. There is not enough information provided (What information do you need?_____) [1%]
Left Question Blank [1%]

13. What is the number needed to treat (NNT) in the following scenario? *10,000 hypertensive patients were randomly assigned either drug (n = 5,000) or placebo (n = 5,000). There were 185 deaths in the drug group and 210 in the placebo group.*
A. About 16 [1%]
B. About 23 [2%]
C. About 25 [8%]
D. About 200 [30%]
E. About 400 [19%]
F. There is not enough information provided (What information do you need?____) [33%]
Left Question Blank [7%]

14. Ten of every 1,000 women have breast cancer
Of these 10 women with breast cancer, nine test positive
Of the 990 women without cancer, about 89 nevertheless test positive

A woman tests positive, she wants to know from you whether she has breast cancer for sure, or at least what the chances are. What is the best answer?
A. The probability that she has breast cancer is about 90%. [12%]
B. The probability that she has breast cancer is about 81%. [5%]
C. Of 10 women with a positive mammogram, about nine have breast cancer. [2%]
D. Of 10 women with a positive mammogram, about one has breast cancer. [49%]
E. The probability that she has breast cancer is about 1%. [18%]
F. There is not enough information provided (What information do you need?_____) [7%]
Left Question Blank [7%]

Note For each question, the correct answer is in bold and the percentages of physicians who chose each response alterative are given in brackets.

Participants were asked to indicate which of 10 ob-gyn–related topics occur in their practice at least once per month or more. Most topics were relevant to a majority of the sample: HIV (88%), mammography (88%), use of cigarettes (88%), likelihood of medication side effects (85%), genital herpes (84%), and cancer of any kind (78%), breast cancer, cardiovascular (73%). HIV/AIDS and Chagas were relevant to 42% and 0%, respectively. These were the topics of the OGSLQ questions and served to assess whether topics were relevant to practice. Half (67%) indicated seven or more topics and 50% indicated eight or more topics.

None of the question types (e.g., Fact, Concepts, and Relations) were correlated with the total number of topics that physicians indicated were relevant to their practice. However, for one item, the positive predictive value question (question 14), those who indicated that breast mammography occurs at least once per month were less likely to get the answer correct: 10 of the 11 (91%) for whom the topic of breast mammography does not come up in their practice answered the question correctly, compared to only 46 of the 82 (56%) for whom breast mammography is relevant (χ^2 (1) = 9, $p = .004$).

The OGSLQ

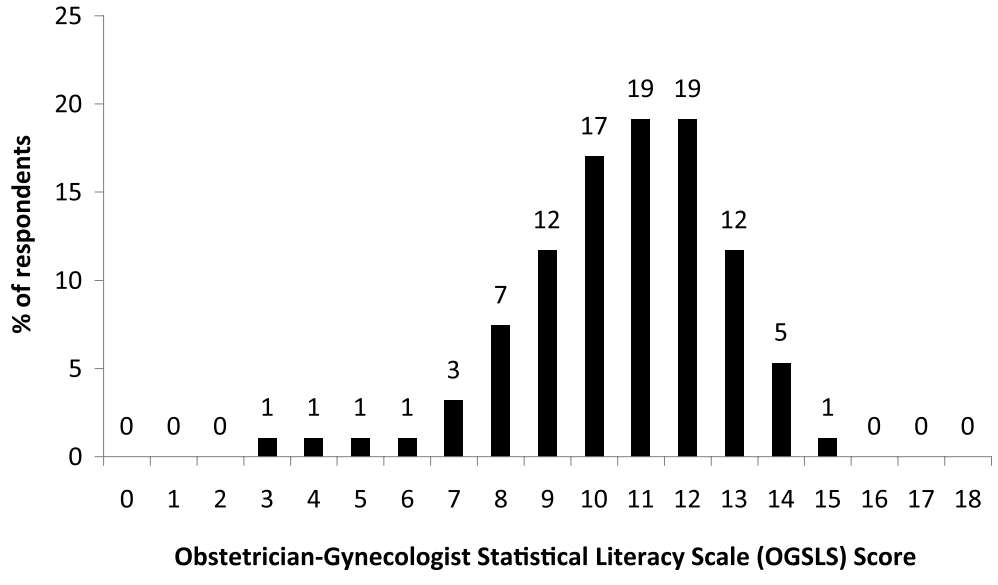
Table 1 lists each question and the percent of responders that answered each question correctly. Table 2 lists each of the 14 questions and shows the percent of respondents who answered each item correctly. Figures 1–4 show the distribution of scores for the OGSLQ,

Table 2. Performance on the Obstetrician-Gynecologist Statistical Literacy Questionnaire (OGSLQ)

Question number	Question type	Percent correct ($n = 94$)	Question topic
1	Fact	7	Number of U.S. women with HIV/AIDS
2	Fact	19	Number of U.S. women diagnosed with cancer per year
3	Fact	90	Whether mammography reduces chances of getting cancer
4a	Fact	23	Direction of change in risk of heart disease due to HRT*
4b	Fact	100	Direction of change in risk of osteoporosis due to HRT*
4c	Fact	89	Direction of change in risk of breast cancer due to HRT*
4d	Fact	81	Direction of change in risk of colon cancer due to HRT*
5a	Fact	15	Amount of increased risk of heart disease due to smoking
5b	Fact	9	Amount of increased risk of lung cancer due to smoking
6	Relation	82	Comparing absolute and relative risk statements
7	Concept	74	Recognizing a statement as the incidence of a disease
8	Concept	68	Recognizing a statement as the prevalence of a disease
9	Concept	66	Using specificity and sensitivity to choose a test option
10	Relation	77	Interpreting a risk reduction
11	Relation	89	Translating a conditional probability to a frequency
12	Relation	93	Translating a frequency to a joint probability
13	Concept	30	Calculating a number needed to treat (NNT)
14	Relation	49	Calculating positive predictive value (PPV) of screening result

Note For each question, the table lists the question number, the question type, percent of correct answers, and the topic of the question. See Table 1 for the actual questions and response options.
*HRT, hormone replacement therapy.

Figure 1. Distribution of Total OGSLQ Scores



Fact Questions, Concept Questions, and the Relation Questions, respectively. Cronbach's alpha for the OGSLQ showed a reliability of $\alpha = .52$.

Questions 7, 8, 10, 11, 12, 13, and 14 all had the response option "There is not enough information provided (What information do you need?___)." Except for question number 13, only a small number of ob-gyns selected this option on each question ($n = 5, 2, 0, 0, 1, 31$, and 7, respectively). Their comments varied widely; for example, on question 13 some of the written comments included the following: "*p* value," "knowledge," "cause of death,"

Figure 2. Distribution of Fact Questions

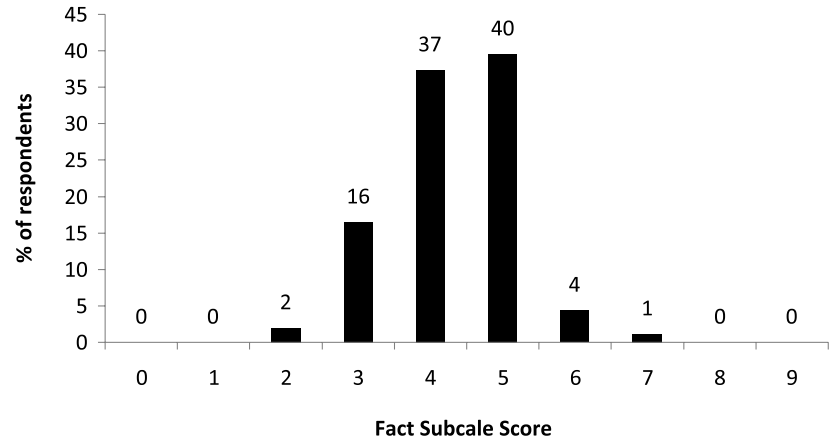
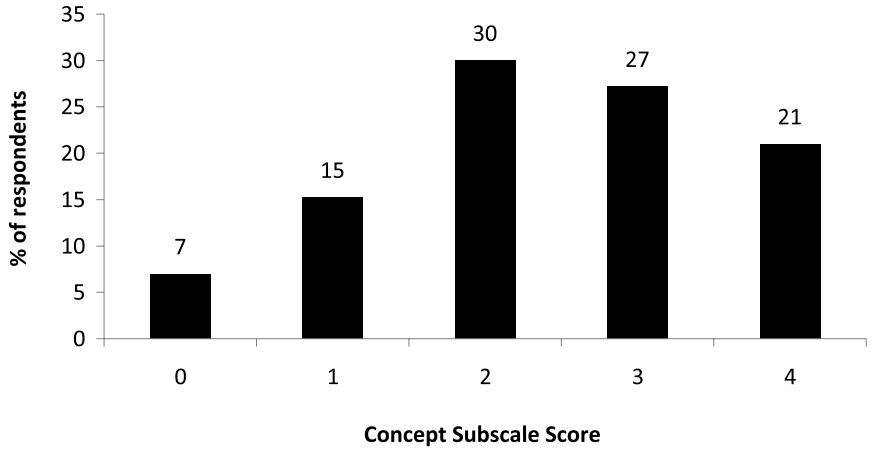


Figure 3. Distribution of Concept Questions



"number needed to treat what?," "what is number needed to treat?," "confidence interval standard deviation," "I don't understand the question," and "endpoint." Few respondents left any given question blank.

Fact Questions

Cronbach's reliability index showed an internal reliability of $\alpha = .01$, indicating that the Fact Questions do not measure a single underlying competence, but that factual knowledge is heterogeneous. Neither age nor gender was associated with performance on the Fact Questions.

Concept Questions

Seven percent of responders answered all four of the Concept Questions incorrectly. Cronbach's alpha for the Concept Questions showed a reliability of $\alpha = .55$. Age was negatively correlated with the Concept Question scores (older physicians did more poorly when controlling for gender) ($r = -.25$, 95% CI $= -0.46$ to -0.06), but there was no association with gender when controlling for age.

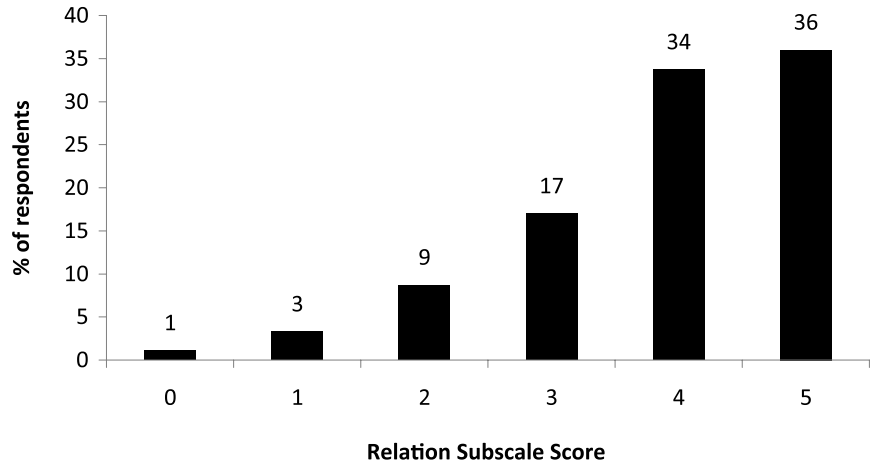
Relation Questions

A total of 36% answered all of the Concept Questions correctly. Cronbach's alpha for the Relation Questions showed a reliability of $\alpha = .53$. Neither age nor gender was associated with performance on the Relation Questions.

Concept–Relation Questions

Because the Fact Questions had low internal consistency compared to the Concept and

Figure 4. Distribution of Relation Questions



Relation Questions, we investigated combining the Concept and Relation Questions without the Fact Questions as a measure of numeracy. The distribution of scores when combining the two is shown in Figure 5. Cronbach's alpha for the Concept–Relation Questions showed a reliability of $\alpha = .61$. Age was negatively correlated with the Concept–Relation Questions (older physicians did more poorly when controlling for gender; $r = -.28$, 95% CI $= -0.46$ to -0.07), but there was no association with gender when controlling for age.

Comparison of OGSLQ with Schwartz and Lipkus Numeracy Scales

Seventy-eight percent of respondents answered all three items on the Schwartz Numeracy Scale correctly (possible scores are 0–3). The distribution of scores is shown in Figure 6. Cronbach's alpha for the Schwartz Numeracy Scale showed a reliability of $\alpha = .12$.

Sixty-one percent of respondents answered all 10 items on the Lipkus Numeracy Scale correctly (possible scores are 0–10). Cronbach's alpha for the Lipkus Numeracy Scale showed a

Figure 5. Distribution of Concept–Relation Questions

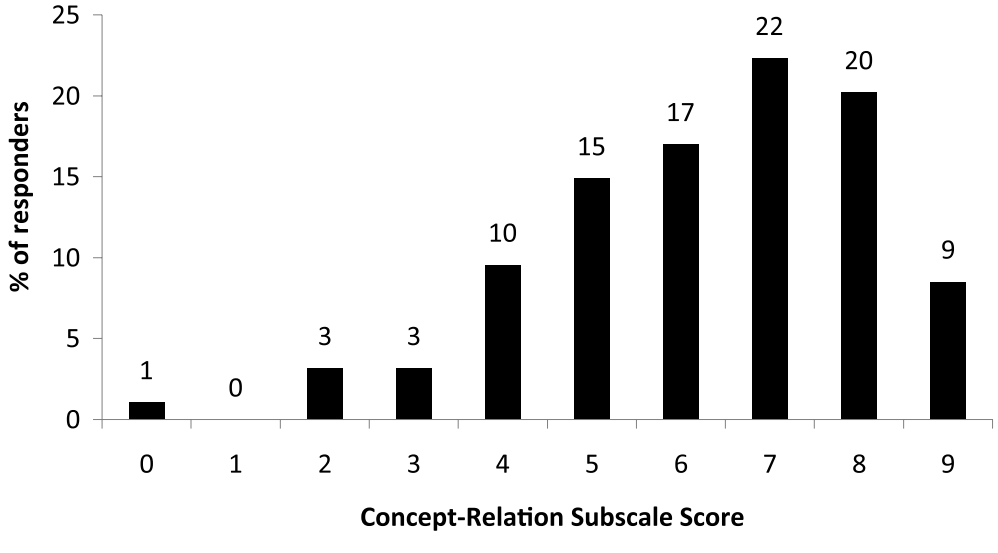


Figure 6. Distribution of Schwartz Numeracy Scale Scores

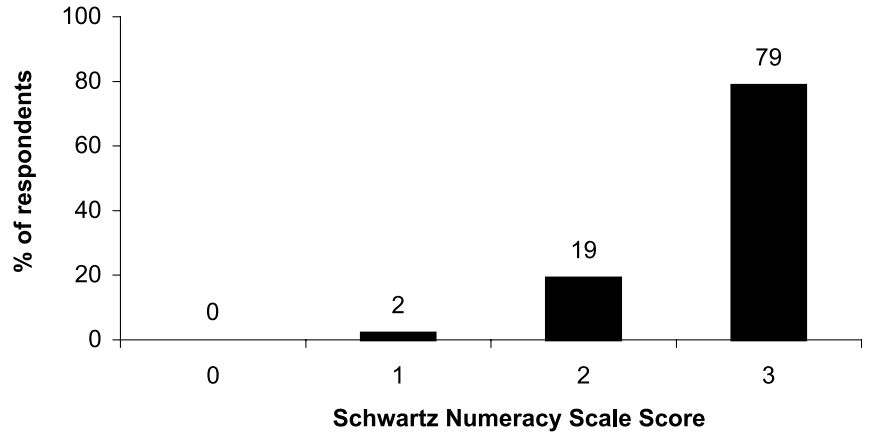
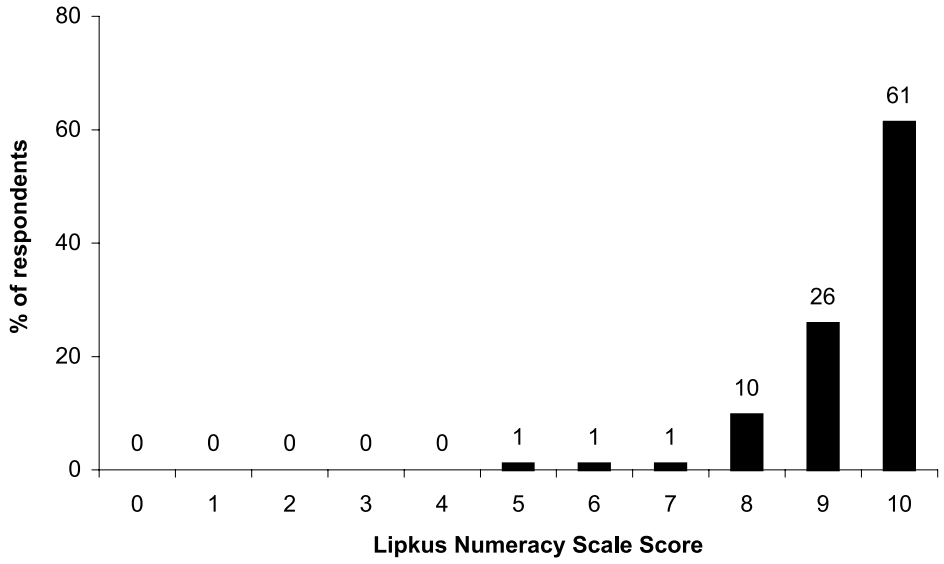


Figure 7. Distribution of Lipkus Numeracy Scale Scores



reliability of $\alpha = .30$. The distribution is shown in Figure 7.

The Relation and Concept Questions were correlated (about $r = .3$) with the Schwartz Numeracy Scale but less so (about $r = .2$) with the Lipkus Numeracy Scale. The Concept–Relation Questions correlated more strongly with both the Schwartz and Lipkus Scales. The total OGSLQ score was correlated with the Schwartz Numeracy Scale ($r = .4$) and the Lipkus Numeracy Scale ($r = .3$; see Table 3).

Limitations

This study's limitation is a possible self-selection bias. The physician invitees were randomly sampled, but those who chose to respond may be different from those who did not. For example, it is possible that some of the obgyns who had difficulty answering questions or were not confident that they could answer the questions did not return the survey. Compared with other physician samples (Estrada, Barnes, Collins, & Byrd, 1999; Gigerenzer, 2010; Gigerenzer et al., 2007), respondents to this survey performed slightly better on the Schwartz measure; other studies using convenience samples at grand rounds reported that 72% (Gigerenzer et al., 2007) and 60% (Estrada et al., 1999) answered all three questions correctly. Using a subjective scale such as the

Table 3. Correlations of the OGSLQ, Fact Questions, Concept Questions, Relation Questions, Concept–Relation Questions, Schwartz Numeracy Scale, and Lipkus Numeracy Scale

	Fact Questions	Concept Questions	Relation Questions	Schwartz Numeracy Scale	Lipkus Numeracy Scale
Fact Questions		$r = .15$ CI = −0.05–0.35	$r = .16$ CI = −0.04–0.36	$r = .19$ CI = −0.01–0.38	$r = .14$ CI = −0.06–0.34
Concept Questions	$r = .15$ CI = −0.05–0.35		$r = .28$ CI = 0.08–0.46	$r = .29$ CI = 0.09–0.46	$r = .18$ CI = −0.02–0.38
Relation Questions	$r = .16$ CI = −0.04–0.36	$r = .28$ CI = 0.08–0.46		$r = .30$ CI = 0.10–0.47	$r = .19$ CI = −0.01–0.38
Concept–Relation Questions	$r = .19$ CI = −0.02–0.38			$r = .37$ CI = 0.18–0.53	$r = .23$ CI = 0.03–0.42
Total OGSLQ				$r = .39$ CI = 0.21–0.56	$r = .25$ CI = 0.05–0.43

Note Pearson's r and 95% CIs are reported ($n = 94$).

SNS might help to increase response rates; however, little is known about how well subjective numeracy correlates with objective numeracy, especially among physicians. The response rate to this study was similar to other studies conducted with this sample (Anderson, Silverman, Loewenstein, Zinberg, & Schulkin, 2007; Anderson et al., 2008).

Discussion

In sum, this study is the first study to assess a variety of statistical literacy skills in a specialty group of practicing physicians using tasks that are relevant to their practice.

Our first question was as follows: how well do ob-gyns perform on statistical literacy tasks that are relevant to their own medical practice? Knowledge of basic facts such as the number of women with cancer or HIV/AIDS is important base-rate information that physicians need in order to place their patients' risks into context. However, physicians do not appear to have accurate perceptions about these things. For example, 79% of the respondents overestimated the prevalence of HIV, even though 42% reported that the issue of HIV comes up in their practice at least monthly. It also appeared that many of the respondents significantly over- or underestimated. In fact, for two questions (questions 1 and 5b) the most frequent answer is the one furthest from the correct response. The large number of incorrect responses to the question about increased risk of heart disease due to hormone replacement therapy (HRT) is likely due to the evidence that HRT affects different aspects of heart disease in opposite ways; HRT is thought to increase risk of stroke but decrease risk of heart attack (Schulkin, 2008).

Physicians answered a large number of the Concepts Questions incorrectly, with 52% answering only two or fewer of the four correctly. These findings are in line with a study that assessed residents' knowledge of biostatics concepts (Windish, Huot, & Green, 2007). Residents in this study were administered a survey, which required that they demonstrate understanding and ability to interpret biostatical methods, study design, and data interpretation (unlike our study, no calculations were required). On average, residents only answered eight of 20 questions (40%) correctly.

We found that younger ob-gyns performed better on the Concept Questions than older ob-gyns. These results may reflect that medical schools have improved their statistical training in the past decade or two; though it is also possible that younger ob-gyns performed better because they had been trained more recently. The positive predictive value question in the OGSLQ (question 14) had been used in a previous study (Gigerenzer et al., 2007) with gynecologists so we compared our results to the earlier one. In our sample, 49% got the answer correct when information was presented in the frequency format, whereas only 21% gave the correct answer in the previous study (Gigerenzer et al., 2007) when information was presented in the conditional probability format. However, when those gynecologists were then taught how to convert conditional probabilities into absolute frequencies, 87% provided the correct answer. Another study

(Obrecht, Anderson, Schulkin, & Chapman, 2011) found that physicians are most accurate at estimating positive predictive values when they estimate event frequencies rather than single-event frequencies using their past experiences.

Our second question was whether or not the different types of questions (e.g., Fact, Concept, and Relations) were correlated with each other. The Concept and Relation Questions showed a small correlation ($r = .28$) with each other, and their correlations with the scores on the Fact Questions were even smaller ($r = .15, .16$). One might think that numerical skills build on one another (e.g., elementary skills such as addition and subtraction need be mastered before one can solve a complex equation), but statistical facts, concepts, and relations appear to be relatively independent from each other. We believe this to be the first attempt to measure different aspects of physicians' statistical literacy in one scale. Further research will be needed to investigate how this measure can be used and improved to predict patient outcomes.

Third, we examined how the OGSLQ relates to two existing numeracy scales. The first difference is that the OGSLQ goes beyond testing the ability to perform simple calculations, and includes facts, concepts, and relations relevant to ob-gyn practice. The second difference concerns discriminatory power. While the Schwarz and Lipkus scales can differentiate between individuals and nations (Galesic & Garcia-Retamero, 2010), they appear to provide relatively little discrimination power among physicians, as well as the ob-gyns in our sample. In contrast, the OGSLQ shows that there is a considerable spread of performance among physicians. The Concept and Relation Questions also have higher internal consistency than do the Schwartz and Lipkus scales in our physician sample. It would be easy to adjust these scales for other medical specialties by changing the topics of the questions.

Implications for Practice

In general, numeric information is more transparent in the format of frequencies (rather than single-event probabilities), absolute risks (rather than relative risks), mortality rate (rather than survival rates), and natural frequencies (rather than conditional probabilities; Gigerenzer, 2002; Gigerenzer & Edwards, 2003). Some good news is that many ob-gyns were able to convert information to and from these transparent formats. For example, 93% were able to translate a frequency into a probability and 89% were able to translate a probability into a frequency. It has been suggested the physicians should use the simplest mathematical constructs when communicating numerical concepts to patients (Apter et al., 2008). Our results show that they are good at converting between formats. The remaining question is whether physicians are aware of which formats are best for their patients. If physicians are aware of which formats are less transparent, they will likely be able to make simple conversions (as demonstrated here) to more transparent formats. Our results also show that many ob-gyns are unable to calculate a positive predictive value when provided with the information needed (prevalence, specificity, and sensitivity). The implications of this finding may be that patients are misinformed about their risks or that they are left to perform this advanced calculation on their own. Inaccurate knowledge about the base rate of common diseases and conditions may also lead some physicians to provide inaccurate risk estimates to their patients. More research is needed to establish whether these findings have such implications and whether these implications have a significant impact on the quality of care. Efforts to increase physician statistical literacy through training in medical school and continuing medical education would be useful. In order to address and improve collective statistical literacy, it is important that we monitor physicians' statistical literacy, to which end this scale can help.

Acknowledgment

This study is funded by grant, UA6MC19010, through the U.S. Department of Health and Human Services, Health Resources and Services Administration, Maternal, and Child Health Research Program.

References

Anderson, B. L., Obrecht, N. A., Chapman, G., Driscoll, D. A., & Schulkin, J. (2011). Physicians' communication of down syndrome screening test results: The influence of physician numeracy. *Genetics in Medicine, 13*, 744–749.

Anderson, B. L., Silverman, G. K., Loewenstein, G. F. Zinberg, S., & Schulkin, J. (2007). Factors associated with physicians' reliance on pharmaceutical sales representatives. *Academic Medicine, 84*, 994–1002.

Anderson, B. L., Schulkin, J., Ross, D. S., Rasmussen, S. A., Jones, J. L., & Cannon, M. J. (2008). Knowledge and practices of obstetricians and gynecologists regarding cytomegalovirus infection during pregnancy—United

States, 2007. *Morbidity and Mortality Weekly Report, 57*, 65–68.
Apter, A. J., Paasche-Orlow, M. K., Remillard, J. T., Bennet, I. M., Pearl Ben-Joseph, E., Batista, R. M., et al. (2008). Numeracy and communication with patients: They are counting on us. *Journal of Gene Medicine, 23*, 2117–2124.
Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making, 27*, 638–654.
Estrada, C., Barnes, V., Collins, C., & Byrd, J. C. (1999). Health literacy and numeracy. *Journal of the American Medical Association, 282*, 527.
Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*, 672–680.
Forrow, L., Taylor, W. C., & Arnold, R. M. (1992). Absolutely relative: How research results are summarized can affect treatment decisions. *The American Journal of Medicine, 92*, 121–124.
Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine, 170*, 462–468.
Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.
Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal, 327*, 741–744.
Gigerenzer, G., Gaissmaier, W., Kurz-Mielke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*, 53–96.
Gigerenzer, G. (2010). Collective statistical illiteracy: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine, 170*, 468–469.
Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37–44.
Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., & Peters, E. (2008). Clinical implications of numeracy: Theory and practice. *Annals of Behavioral Medicine, 35*, 261–274.
Obrecht, N. A., Anderson, B. L., Schulkin, J., & Chapman, G. B. (2011). Retrospective frequency formats promote consistent experience-based Bayesian judgments. *Applied Cognitive Psychology*, DOI: 10.1002/acp2816.
Peters, E., Vasfjall, D., Slovic, P., Merzt, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*, 407–413.
Reyna, V. F., Nelson, W. L., Han, P., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*, 943–973.
Schulkin, J. (2008). *Medical decisions, estrogen, and aging*. New York: Springer.
Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*, 966–972.
Windish, D. M., Huot, S. J., & Green, M. L. (2007). Medicine residents' understanding of the biostatistics and results in the medical literature. *Journal of the American Medical Association, 298*, 1010–1022.

Appendix A
Statistical Terminology Glossary

Absolute risk reduction (ARR) (question 6): The term ARR is a measure of the efficacy of a treatment in terms of the absolute number of people saved. It is the numerical difference between the rates of the adverse outcome in the control group and the intervention group, with ARR = 1/NNT.

Relative risk reduction (RRR) (question 6): The term RRR is a measure of the efficacy of a treatment in terms of the relative number of people saved. It refers to ARR divided by the rate of the adverse outcome in the control group, with RRR = ARR/rate of the adverse outcome in the control group.

Incidence (question 7): The term incidence rate refers to the number of new cases of a disease within a specified time interval (e.g., over 1 year)

Prevalence (question 8): The term prevalence (synonymous with the term "base rate") refers to the total number of cases of a disease at a certain point in time (e.g., the day that the sample was tested).

Sensitivity (question 9): The term sensitivity refers to the proportion of individuals with a positive test result among those who have the disease; in other words, those who are correctly classified as having the disease.

Specificity (question 9): Specificity is the proportion of individuals with a negative test result among those without the disease; in other words, those who are correctly classified as not having the disease.

False positive rate (question 9): The term false positive rate refers to the proportion of individuals who test positive but do not have the disease; in other words, those who are incorrectly classified as having the disease.

Conditional probability (question 11): A conditional probability is the probability that an event A occurs given event B; a joint probability is the probability of A and B cooccurring.

Frequency (question 11): A frequency is the number of observations in a class of events.

Number needed to treat (NNT) (question 13): The term NNT is a statistical measure of the efficacy of a treatment; it is defined as the number of people who need to receive an intervention in order to save one person from an adverse outcome (e.g., death). The NNT = 1/(rate of the adverse outcome in the control group—rate of the adverse outcome in the intervention group).

Positive predictive value (PPV) (question 14): The term PPV is the proportion of people among all those who test positive who actually do have the disease (or condition). In other words, PPV is defined as the number of true positives divided by the total number of those

who test positive, with PPV = numbers who are correctly identified has having the disease or total number who tests positive.

Authors' Biographies

Britta L. Anderson, PhD, works as a Research Associate at the American College of Obstetricians and Gynecologists. She graduated with a PhD in psychology from American University. Her research interests include literacy and numeracy, medical decision making, behavioral economics, women's health, and tobacco and caffeine use and abuse.

Gerd Gigerenzer, PhD, is Director at the Max Planck Institute for Human Development and Director of the Harding Center for Risk Literacy in Berlin. His research interests include bounded rationality and social intelligence, decisions under uncertainty and time restrictions, competence in risk and risk communication, decision-making strategies of managers, judges, and physician. In one of his recent books, Better Doctors, Better Patients, Better Decisions, he shows how better informed doctors and patients can improve healthcare while reducing the costs.

Scott Parker, PhD, is Professor of Psychology at American University. He is also an Affiliated Faculty of the Department of Mathematics and Statistics at American University. His research interests include the quantitative aspects of our experiences and how we make evaluations and choices based on them.

Jay Schulkin, PhD, is Director of Research at the American College of Obstetricians and Gynecologists. He is also Research Professor in the Department of Neuroscience at Georgetown University and Research Associate in the Behavioral Endocrinology Branch of the National Institute of Mental Health. Some of his research interests include medical decision making and the neuroscience of behavior.

For more information on this article, contact Britta L. Anderson at anderson.britta.l@gmail.com.