

A New Scale for Assessing Perceptions of Chance:

A Validation Study

STEVEN WOLOSHIN, MD, MS, LISA M. SCHWARTZ, MD, MS,
STEPHANIE BYRAM, PhD, BARUCH FISCHHOFF, PhD,
H. GILBERT WELCH, MD, MPH

Background. Clinicians and researchers often wish to know how patients perceive the likelihoods of health risks. Little work has been done to develop and validate scales and formats to measure perceptions of event probabilities, particularly low probabilities (i.e., <1%). **Objective.** To compare a new visual analog scale with three benchmarks in terms of validity and reliability. **Design.** Survey with retest after approximately two weeks. Respondents estimated the probabilities of six events with the new scale, which featured a "magnifying glass" to represent probabilities between 0 and 1% on a logarithmic scale. Participants estimated the same probabilities on three benchmarks: two linear visual analog scales (one labeled with words, one with numbers) and a "1 in x" scale. **Subjects.** 100 veterans and family members and 107 university faculty and students. **Measures.** For each scale, the authors assessed: 1) validity—the correlation between participants' direct rankings (i.e., numbering them from 1 to 6) and scale-derived rankings of the relative probabilities of six events; 2) test-retest reliability—the correlation of responses from test to retest two weeks later; 3) usability (missing/incorrect responses, participant evaluation). **Results.** Both the magnifier and the two linear scales outperformed the "1 in x" scale on all criteria. The magnifier scale performed about as well as the two linear visual analog scales for validity (correlation between direct and scale-derived rankings = 0.72), reliability (test-retest correlation = 0.55), and usability (2% missing or incorrect responses, 65% rated it easy to use). 62% felt the magnifier scale was a "very good or good" indicator of their feelings about chance. The magnifier scale facilitated expression of low-probability judgments. For example, the estimated chance of parenting sextuplets was orders of magnitude lower on the magnifier scale (median perceived chance 10^{-5}) than on its linear counterpart (10^{-2}). Participants' assessments of high-probability events (e.g., chance of catching a cold in the next year) were not affected by the presence of the magnifier. **Conclusions.** The "1 in x" scale performs poorly and is very difficult for people to use. The magnifier scale and the linear number scale are similar in validity, reliability, and usability. However, only the magnifier scale makes it possible to elicit perceptions in the low-probability range (<1%). **Key words:** patient perceptions; perception measurement scale (Med Decis Making 2000;20:298–307)

Received November 29, 1999, from the VA Outcomes Group, White River Junction, Vermont (SW, LMS, HGW); the Center for the Evaluative Clinical Sciences, Dartmouth Medical School, Hanover, New Hampshire (SW, LMS, HGW); Norris Cotton Cancer Center, Lebanon, New Hampshire (SW, LMS); and Carnegie Mellon University, Pittsburgh, Pennsylvania (SB, BF) Drs. Woloshin and Schwartz are joint first authors—the order of their names is arbitrary. They are supported by Veterans Affairs Career Development Awards in Health Services Research and Development and a New Investigator Award from the Department of Defense Breast Cancer Research Program. Grant DAMD17-96-MM-6712. The views expressed herein do not necessarily represent the views of the Department of Veterans Affairs or the United States Government.

Address correspondence and reprint requests to Dr. Woloshin: VA Outcomes Group (111B), Department of Veterans Affairs Medical Center, White River Junction, VT 05009.

Clinicians, researchers, and policymakers have a strong interest in learning how patients perceive the chances of disease. If people at high risk of disease are unaware of their elevated risks, they may fail to appropriately consider potentially beneficial interventions. If those whose risks of disease are low have a falsely heightened sense of risk, they may experience undue health-related anxiety or may pursue interventions that offer more harm than benefit. Educational efforts and the informed consent process must begin with not only an understanding of people's beliefs regarding the nature of disease but also their perceptions about the chances of disease.

Surprisingly little work has been done to validate methods of eliciting such perceptions. Only one

published study formally compares the psychometric properties of scales commonly used to quantify patients' perceptions about health hazards. Diefenbach et al. concluded that format importantly influenced the validity, reliability, and usability of a scale.¹ Two important problems limit the application of the study results. First, the study sample consisted of college students enrolled in an introductory psychology course. Consequently, it is not known how these scales would perform among those less educated. Second, the one scale intended to elicit small probabilities (i.e., <1%)—the range of many diseases screened for in the general population—performed very poorly.

We developed a new "magnifier" visual analog scale, designed to make it easier for people to quantify event probabilities—particularly small probabilities. This scale, shown in figure 1, features a magnifying glass to represent probabilities between 0 and 1% on a logarithmic scale. Although similar scales have been used previously,²⁻⁴ they have not been subject to formal evaluation. We evaluated the magnifier scale's performance for events varying widely in their probabilities and compared its validity, reliability, and usability with those of three benchmark scales

in a diverse group of people sampled from a university and outpatient clinic setting. The respondents' main task was to estimate the probabilities of six events using each of the four scales. In order to test reliability, the survey was repeated at two weeks.

We compared the magnifier with the following three benchmark scales:

- *Linear word scale*—a visual analog scale labeled with words, based on the scale that performed "best" in the Diefenbach et al. study.¹ Given the inherent variability in how word labels are interpreted (e.g., "very unlikely," "moderate chance"),⁵⁻⁹ this scale was designed to assess the relative order of events rather than to quantify actual probabilities. Since ordering events is less demanding than estimating their probabilities, we selected this scale as a benchmark to help define a reasonable upperbound of scale performance.
- *"1 in x" scale*—this format is widely used to quantify chance. For example the "1 in 8" statistic used by the American Cancer Society to describe a woman's chance of developing breast cancer.¹⁰
- *Linear number scale*—a visual analog scale labeled with numbers (i.e., magnifier scale minus the "magnifier"). This quantitative benchmark allowed us to isolate the effect of the "magnifying glass."

Methods

OVERVIEW

The main goal of this study was to compare the performance of the new magnifier scale with the performances of three benchmark scales (figure 2)

DESIGN AND SAMPLE

We recruited a convenience sample of 207 men and women to complete our initial survey. Of these, 107 faculty and students at Carnegie Mellon Univer-

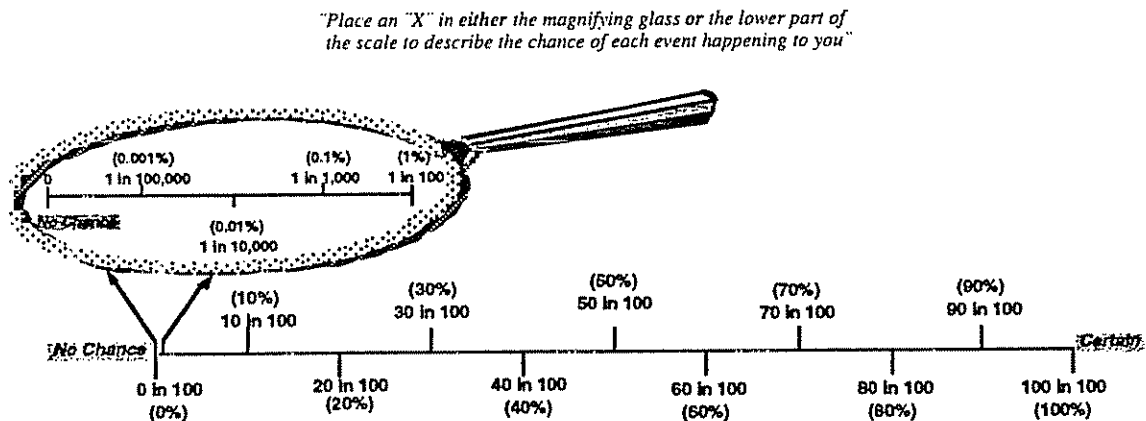


FIGURE 1. The magnifier scale for measuring perceptions of event probability. Prior to the section of the survey with the magnifier scale, we included the following explanation:

On the next few pages you will find questions about how likely it is that various things will happen. We will ask you to put your answers on scales like the ones you see here. The scale is a line which goes from "no chance" (0%) to "certain" (100%). It has a magnifying glass for the smallest chances. For the first example, we have marked with an "X" the chance of an average person being killed by lightning in the next 10 years. Fortunately, this chance is very low so it goes in the magnifying glass. (Scale shown with "X" marked slightly below 1 in 100,000). In example 2, we have marked with an "X" the chance of getting junk mail in the next year. Unfortunately, this chance is **very high**. (Scale shown with "X" marked between 90 and 100%).

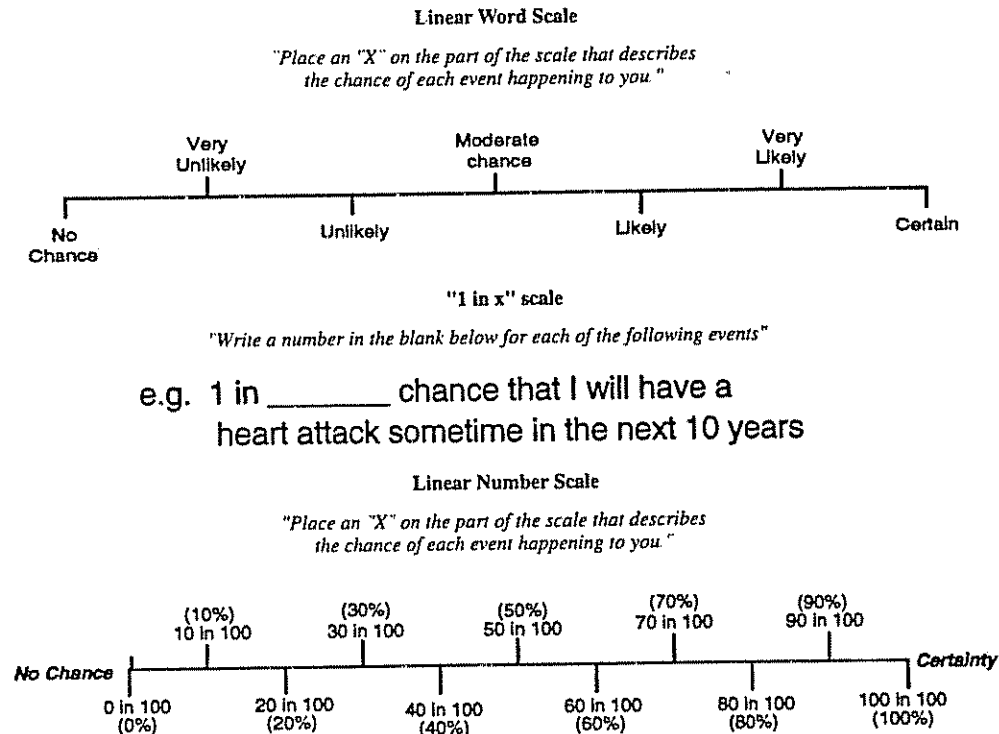


FIGURE 2 The three benchmark scales that the respondents also used to indicate the chances of six events

sity were recruited via an electronic bulletin board. The university participants received and returned their surveys through campus mail. The 100 veterans and family members at the White River Junction VA Medical Center were recruited from the clinic waiting area, where they completed the survey. Respondents were paid \$5 for each survey completed.

QUESTIONNAIRE DESIGN

The structure of our questionnaire and the basic analytic approach built on those of Diefenbach et al.¹ We selected six familiar events whose probabilities ranged from extremely rare to extremely common. We asked respondents to rate the likelihoods of the six events in different ways: 1) *Direct ranking*: ordering the events by their chances of occurrence (i.e., numbering them from 1 to 6) and 2) *scale-derived rankings*: here the order was inferred from how the respondents estimated the probabilities of each of the six events on the magnifier and the three benchmark scales (i.e., 6 events \times 4 scales = 24 separate questions).

Direct ranking. The first survey question asked:

Please rank the following in order of how **likely** it is that each will happen to you. Use a "1" to indicate the *most likely* event, a "2" for the second most likely . . . and so on up to "6" for the *least likely*

event. Please use each number only once to rank the following 6 events in order of their likelihood:

- _____ **Die from any cause** within the next 10 years.
- _____ **Sustain a minor injury in a car crash** sometime in the next 10 years.
- _____ **Have a heart attack** sometime in the next 10 years.
- _____ **Be the parent of "sextuplets"** sometime in the next 10 years.
- _____ **Be diagnosed with breast cancer** sometime in the next 10 years.
- _____ **Catch a cold** sometime in the next *one* year.

We considered this direct ranking to be the "gold standard" assessment, for two reasons. First, the ordering of probability judgments is preserved even when absolute values are sensitive to the type of scale used.^{11,12} Second, direct ranking makes no assumptions about how respondents interpret numbers or words.

Scale-derived ranking. Respondents assessed the likelihoods of the same six events on the magnifier and on the three benchmark scales. That is, for each scale, we determined the scale-derived ranking of the probabilities of the six events to compare against the "gold standard" direct ranking. The scale-derived rankings are simply the orders of the six events

based on the respondents' probability estimates using a particular scale. Standard ranking procedure were followed; a missing values (i.e., the respondent failed to mark a particular scale) was not assigned a rank, and a tie was assigned the average of the tied ranks.

To reduce any effects of event or scale order, we randomly distributed four versions of the survey that differed only in the ordering of the events and scales. Since we observed no important ordering effect, we report the aggregated results.

MEASURING VISUAL ANALOG SCALE RESPONSES

Linear scales. We measured the distance in millimeters from the subject's mark to the scale's leftmost anchor (i.e., "no chance"). Dividing this value by the total length of the scale converts the measured distance into a probability measure and allowed us to interpolate responses. For example, if the scale were 100 mm long and the subject placed a mark 75 mm from the leftmost anchor, the corresponding probability is 75 mm / 100 mm, or 0.75.

Magnifier scale. Respondents were asked to place a mark in either the magnifying glass or the linear portion of the scale. Marks on the linear portion were interpolated as described above. Marks on the magnifying glass portion were converted in the same way after interpolating their positions on the log scale (STATA subroutine available on request).

SCALE EVALUATION CRITERIA

We judged the performance of each scale using three criteria adapted from prior work²: validity, usability, and satisfaction, and test-retest reliability.

Validity. For each scale, validity was assessed by correlating each respondent's direct and scale-derived rankings of the chances of the six events. These individual-level Spearman correlation coefficients were calculated from six data pairs. For example, for each person the validity of the magnifier scale was the correlation of his or her direct and magnifier-derived ranks for the six events. A scale's validity is the average of these 207 individual-level coefficients.

Usability and satisfaction. After completing the probability judgments, participants rated each scale from "very hard to use" to "very easy to use" and on how well it represented their feelings about chance from "very good indicator of my feelings" to "very poor indicator of my feelings." Each participant also picked the one scale that best represented his or her feelings about small chances. We also tallied item non-response for each scale and the number of unusable responses.

Test-retest reliability. All respondents were asked

to complete the same survey approximately two weeks later—178, or 88%, did so (VA sample 71%; Carnegie Mellon University 100%). All retest surveys were returned by mail.

To calculate test-retest reliability, we began by calculating the event-level (e.g., chance of getting a cold) Pearson correlation coefficient. This coefficient was calculated for the 178 pairs of probability estimates for that event at test and retest. A scale's test-retest reliability was the average of these six event-level correlation coefficients.

NUMERACY

In addition to basic demographic information for all respondents (age, sex, educational attainment, total household income), we assessed numeracy—basic facility with numerical concepts and probability. Numeracy was scored as the total number of correct answers to the following three-item scale¹³: 1) estimate the number of heads in 1,000 flips of a fair coin; 2) convert a proportion (1 in 1,000) into a percentage (0.1%); and 3) convert a percentage (1%) into a proportion (10 in 1,000). Missing answers were treated as incorrect.

ANALYSIS

Chi-square tests and Kruskal-Wallis tests were used to compare scale evaluation criteria across the four scales. All comparisons were two-sided and were considered statistically significant at $p < 0.05$. In order to assess whether the differences in validity or reliability for the four scales were statistically significant, we created multiple linear regression models. For the validity model, the dependent variable was the Fisher Z-transformed validity correlation coefficient and independent variables were indicators for scale and person (i.e., 207 individual-level correlations \times 4 scales = 828 observations). For the reliability model, the dependent variable was the Fisher Z-transformed reliability correlation coefficient and independent variables were indicators for scale and events (i.e., 4 scales \times 6 events = 24 event-level observations). We used STATA 5.0 (College Station, TX) for all analyses.

Results

Table 1 shows that our samples of university faculty and students (Carnegie Mellon University—CMU, Pittsburgh, PA) and veterans and their families (White River Junction VAMC and ROC, Vermont) represented both ends of the sociodemographic spectrum. The CMU sample was younger, more affluent, more likely to be employed or in school, and had more formal education than the VA sample. Dif-

Table 1 • Characteristics of the Study Sample by Site and Combined

	Carnegie Mellon University Pittsburgh, PA	VA Medical Center White River Junction, VT	Combined
Sample frame	Faculty and students	Veterans and their families	
Number	107	100	207
Completed retest survey (%)	100%	71%	88%
Median age (25%ile, 75%ile)	34 years (22, 45)	56 years (44, 68)	44 years (30, 58)
Sex (women)	61%	52%	57%
Household income			
<\$10,000	11%	21%	16%
\$10,000-24,999	22%	41%	32%
\$25,000-49,999	32%	33%	32%
\$50,000-74,999	18%	3%	10%
≥\$75,000	17%	2%	10%
Highest level of education			
<high school graduate	0%	21%	10%
High school degree	51%	55%	53%
College degree	31%	23%	27%
Postgraduate degree	18%	0%	9%
Employment			
Student	23%	2%	13%
Employed	68%	19%	44%
Unemployed	0%	4%	2%
Retired	4%	45%	24%
Homemaker	0%	10%	5%
Numeracy score*			
0	4%	35%	19%
1	6%	21%	13%
2	36%	30%	33%
3	54%	14%	35%

*Sum of number of correct answers to the following three numeracy questions: best guess of the number of heads in 1,000 flips of a fair coin; convert a proportion (1 in 1,000) into a percentage (0.1%); and convert a percentage (1%) into a proportion (10 in 1,000).

ferences in education were particularly striking. While all CMU respondents had graduated from high school (and almost half had college or postgraduate degrees), only a fourth of the VA respondents had continued their formal education beyond high school. Not surprisingly, numeracy scores differed markedly across the samples: the median score was 3 at CMU, but only 1 at the VA.

Tables 2 and 3 show the performance criteria for the magnifier and the three benchmark scales. Table 2 reports on validity, usability, and satisfaction; table 3 presents test-retest reliability. We now consider how the magnifier compares to each benchmark.

MAGNIFIER VS LINEAR WORD SCALE

As expected, the linear word scale had the highest validity, usability, and reliability. In essence, this scale established the upper bound of performance that could be reasonably expected based on cur-

rently used methods. The magnifier did almost as well as the linear word scale on most performance criteria. At the same time, however, the magnifier was designed to perform an even more challenging task: actually quantifying probability.

MAGNIFIER VS. "1 IN x" SCALE

The "1 in x" scale is commonly used to present the chances of diseases under the assumption that people can use it well. Despite its apparently simple format, the "1 in x" scale had the lowest validity ($r = 0.64$) and test-retest reliability ($r = 0.45$). Participants described this scale as much harder to use and a poorer indicator of their feelings than the other scales.

Consistent with these poor ratings, 14% of the respondents skipped the "1 in x" scales altogether (compared with 2% for the magnifier scale). Figure 3 shows that proportions of missing responses differed particularly for low-probability events (i.e., 26%

vs 1% on the magnifier). Usability also differed according to respondent numeracy. Participants were called "innumerate" if they answered none or one of our three numeracy questions correctly. These 66 individuals (32% of the total sample) often left the "1 in x" scale blank (ranging from 18% to 50% across questions) but seldom failed to complete any of the magnifier questions (never more than 5% missing).

Many people also seemed to have a difficulty with the "1 in x" scale at both probability extremes. For rare events, the "1 in x" scale had a higher proportion of participants defaulting to zero (e.g., about half estimated the chance of parenting sextuplets to be zero compared with 31% of the same respon-

dents on the magnifier). The "1 in x" scale significantly limited expression of high event probabilities. Given the awkwardness of writing something other than a whole number in the blank, all respondents defaulted to either one in two (0.5) or one in one (1) for common events (e.g., 51% estimated the chance of catching a cold to be "1" and 26% estimated it to be "0.5"). On the magnifier, only 3% of the same individuals estimated "1" and 7% estimated "0.5."

MAGNIFIER VS LINEAR NUMBER SCALE: MAGNIFIER EFFECT

The comparison of the magnifier with the linear

Table 2 • Performance Criteria for the Magnifier Scale and Three Benchmark Scales for Measuring Perceptions of Event Likelihood ($n = 207$)

	Magnifier Scale	Benchmark Scales		
		Linear Scales		"1 in x" Scale
		Numbers	Word	
Validity				
Correlation of scale-derived ranks with direct ranks*	0.72	0.74	0.78	0.64
Usability				
Missing or unusable responses	2%	<1%	1%	14%
"How easy was the scale to use?"				
Very easy/easy	65%	68%	75%	39%
Very hard	6%	6%	4%	18%
Satisfaction with scale				
"How well did the scale reflect your feelings about each chance?"				
very good/good indicator	62%	61%	77%	36%
very poor indicator	7%	8%	4%	16%
"Indicate the scale you liked best to represent small chances, such as the chance of being the parent of sextuplets	39%	14%	40%	6%

*In a multiple linear-regression model (where the dependent variable was the Fisher Z-transformed correlation coefficient) controlling for person, the "1 in x" scale was significantly lower than the other three scales ($p < 0.001$).

Table 3 • Test-Retest Reliabilities of the Magnifier Scale and Three Benchmark Scales.*

	Magnifier Scale	Benchmark Scales		
		Linear Scales		"1 in x" scale
		Numbers	Word	
Parent of sextuplets	0.46	0.42	0.69	0.22
Minor injury in car crash	0.41	0.44	0.45	0.32
Die from any cause	0.53	0.61	0.78	0.47
Heart attack	0.53	0.64	0.70	0.55
Catching a cold	0.70	0.64	0.69	0.57
Being diagnosed with breast cancer	0.67	0.66	0.86	0.58
Mean†	0.55	0.57	0.69	0.45

*The number in each cell is the Pearson correlation coefficient for perceptions on identical surveys two weeks apart ($n = 178$).

†To put these results in context, the correlation between direct rankings at test and retest was 0.83. Because test-retest reliability varied across the six events, we used multiple linear regression to test for statistical differences between the scales independent of these differences across events. For this analysis, the data set consisted of a 4 "scale" \times 6 "event" table where the value in each cell was the Z-transformed Pearson's correlation coefficient. In this model, the dependent variable was the Z-transformed Pearson correlation coefficient and indicator variables for scale type and event were the independent variables.

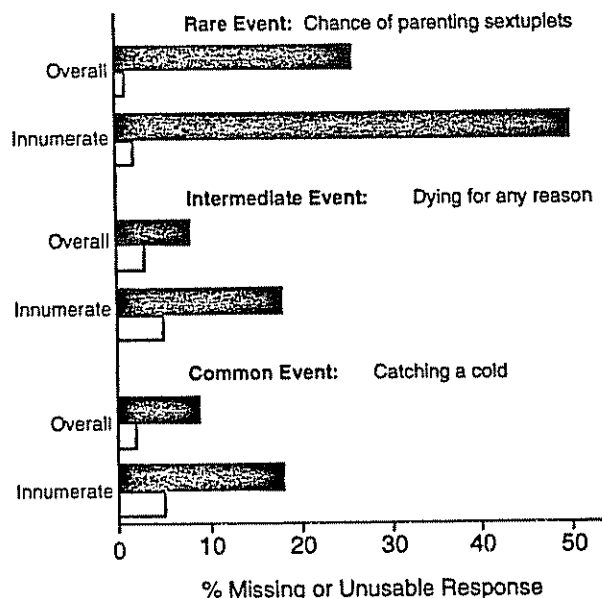


FIGURE 3. Percentages of missing or incorrect responses for perceptions of rare, intermediate, and common events on the "1 in x" scale and magnifier scales. Since there were so few missings for the linear word and number scales (1% or less), we could not examine the effect of numeracy for these scales. The results are presented for the entire sample ($n = 207$) and for the 66 respondents deemed "innumerate" (defined as answering only 0 or 1 of 3 numeracy questions correctly). ■ = "1 in x" scale; □ = magnifier scale.

number scale helps to gauge the costs and benefits of adding the magnifying glass. The cost proved negligible. These two scales had the same validity (mean correlations between direct and scale-derived rankings of the six events were 0.72 vs 0.74 for the magnifier and linear scale, respectively), usability (65% vs 68% rated the scales "easy to use"), and test-retest reliability (0.55 vs 0.57).

DID THE MAGNIFIER ALLOW PEOPLE TO EXPRESS SMALL CHANCES?

Ideally, the magnifier scale allows people to indicate small chances when that is what they mean to do. As can be seen in table 4, the magnifier-scale responses were much smaller for the two events generally acknowledged to be rare: the ten-year chance of parenting sextuplets and a man's ten-year chance of developing breast cancer. For both events, perceptions were orders of magnitude lower on the magnifier than on the linear scale (median perceived chance 10^{-5} vs 10^{-2}).

DOES THE MAGNIFIER BIAS RESPONSES TOWARD ZERO?

One concern about the magnifying glass is that it

will artifactually bias all responses downward (i.e., toward zero) because it may suggest that we expect small values. If the magnifying glass biased responses downward, we would expect to see reductions in all estimates. On the other hand, if the scale worked as intended, we would expect the magnifying glass to reduce estimates for rare events but not for common events.

While the estimates of rare events were lower, we observed no downward shift in the responses for estimates of the high-probability events on the magnifier scale. The distributions of probabilities for the chance of catching a cold in the next year were almost identical with the two scales (median (interquartile range) for magnifier: 0.87 (0.50, 0.97); for linear number 0.88 (0.49, 0.99)).

The intermediate-probability event, the ten-year chance of dying from any cause, is age-dependent. Therefore, we stratified the subjects by age (< 30, 30–59, 60+ years). On both scales, the median probability estimates appropriately and similarly increased with age. While the median estimates for the two scales were nearly identical, we did find a shift in the distributions of the responses. For the youngest participants, the low end of the interquartile range was (appropriately) an order of magnitude lower on the magnifier compared with the linear scale (25%ile: 10^{-3} vs 10^{-2}), while the upper end of the range was unchanged (75%ile: 0.22 and 0.25). Thus, only respondents who thought their chances of dying were low (i.e., marked the low end of the linear number scale) moved into the magnifying glass.

The same phenomenon was seen for participants aged 30–59 [i.e., a tenfold difference in the 25%ile: 10^{-2} (magnifier) vs 0.10 (linear)]. For those aged 60 and older, the shift was smaller, a threefold difference (25%ile: 0.10 vs 0.29).

CAN RESPONDENTS WITH LOW NUMERACY USE THE MAGNIFIER SCALE?

The magnifier scale received usability ratings similar to those of its simple linear counterpart, even among innumerate respondents. Innumerate respondents rated the magnifier scale as "very easy or easy" to use (66% gave this rating to the magnifier and the linear number scale) and as a good indicator of their feelings about chance (59% gave this rating to the magnifier scale vs 62% giving this rating to the linear number scale). Moreover, table 5 shows that participants with low numeracy were also able to use the magnifier scale sensibly. As with the entire sample, probability estimates for rare events were again orders of magnitude lower on the magnifier than on the linear scale, with little evidence to suggest that the magnifier biased responses toward zero.

Table 4 • Perceptions of Rare, Intermediate, and Common Event Probabilities Expressed on the Magnifier and Linear Number Scales

	Magnifier Scale Median (25%ile, 75%ile)	Linear Number Scale Median (25%ile, 75%ile)
Rare events		
Being the parent of sextuplets in the next ten years ($n = 207$)	10^{-5} (0, 10^{-5})	10^{-2} (0, 10^{-2})
Being diagnosed with breast cancer in the next ten years Men ($n = 88$)	10^{-5} (0, 10^{-4})	10^{-2} (0, 10^{-2})
Intermediate event		
Dying from any cause in the next ten years		
20-29 yo ($n = 49$)	0.10 (10^{-3} , 0.22)	0.10 (10^{-2} , 0.25)
30-59 yo ($n = 107$)	0.20 (10^{-2} , 0.50)	0.28 (0.10, 0.49)
60-82 yo ($n = 50$)	0.50 (0.10, 0.60)	0.50 (0.29, 0.69)
Common event		
Catching a cold in the next year ($n = 203$)	0.87 (0.50, 0.97)	0.88 (0.49, 0.99)

Table 5 • Perceptions of Rare, Intermediate, and Common Event Probabilities Expressed on the Magnifier and Linear Number Scales for the 66 "Innumerate" Participants (i.e., Numeracy Score of 0 or 1)

	Magnifier Scale Median (25%ile, 75%ile)	Linear Number Scale Median (25%ile, 75%ile)
Rare events		
Being the parent of sextuplets in the next ten years ($n = 66$)	10^{-5} (0, 10^{-2})	10^{-2} (0, 10^{-2})
Being diagnosed with breast cancer in the next ten years Men ($n = 24$)	10^{-5} (0, 10^{-2})	10^{-2} (0, 10^{-2})
Intermediate event		
Dying from any cause in the next ten years		
39-59 yo ($n = 32$)	0.31 (10^{-4} , 0.50)	0.32 (0.10, 0.50)
60-82 yo ($n = 30$)	0.44 (0.10, 0.79)	0.50 (0.24, 0.68)
Common event		
Catch a cold in the next year ($n = 66$)	0.61 (0.30, 0.97)	0.60 (0.30, 0.95)

Discussion

Despite its initially daunting appearance, the magnifier scale was rated easy to use—even among respondents with low numeracy. The magnifier allowed respondents to better communicate perceptions of low-probability events—estimates of events such as the chance of parenting sextuplets were orders of magnitude lower on the magnifier than on the linear number scale, without evidence of a systematic bias toward zero. The benefit of the magnifying glass comes at little cost: overall its validity, reliability, and usability approximated those of the less demanding (but non-quantitative) linear word scale, far outperformed the "1 in x" scale, and were consistently as good as those of its simple counterpart, the linear number scale.

The responses to the linear number and magnifier scales present different pictures of peoples' expectations regarding low-probability events. The linear number scale may have forced these respondents to overestimate the probabilities of such events because it effectively prohibits re-

sponses between 0 and 1%. Various studies have found that people dramatically overestimate the chances of rare outcomes.¹⁴⁻¹⁶ The present results suggest that the problem may not be thinking about small probabilities, but expressing them in the response modes that investigators offer.

BENCHMARK PERFORMANCE: "1 IN x" AND LINEAR WORD SCALES

The performance of the "1 in x" merits special comment. This format is frequently used to present disease risk, for example, the American Cancer Society uses it in reports of SEER data such as the often cited "1 in 8 lifetime chance of getting breast cancer."¹⁰ A recent paper in the *New England Journal of Medicine* went so far as to suggest that clinicians use the "1 in x" to help patients put their breast cancer risks in perspective.¹⁷ Despite its widespread use, we found that the "1 in x" scale performed substantially worse than all the other scales examined.

The explanation for the poor performance of the "1 in x" scale probably relates to three cognitive

challenges posed by the scale. Its open-ended format (i.e., fill in the blank) requires respondents to generate their own response options without suggestive anchors. Second, the awkwardness of filling in the blank with anything other than a whole number tends to constrain responses at the high end to 33% (1 in 3); 50% (1 in 2), or 100% (1 in 1). Expressing any other probability requires substantial mental effort (e.g., 75% = 1 in 1 1/3). Third, this format invites confusion because smaller chances are expressed with larger numbers in the denominator (e.g., 1 in 200 vs 1 in 500).

We want to be clear that our study assessed the performance of "1 in x" for eliciting perceptions of chance only—not presenting information. Nonetheless, we believe our results should raise questions about the usefulness of this format for presenting data. In fact, a recent study suggests that people have trouble understanding information presented to them in this way. This study asked patients to select the larger of two risks when the risks were expressed as "1 in x" or as a number per 1,000.¹⁸ When the risks were expressed in the "1 in x" format, 17% fewer patients were able to correctly judge the larger of them.

The performance of the linear word scale also merits comment. It has been shown that people tend to prefer expressing their perceptions of chance using words as opposed to numbers.¹⁹ Not surprisingly, the linear word scale did as well as or better than the other scales in terms of validity, reliability, and usability. The fundamental limitation of the word scale is that absolute numerical probabilities can only be inferred from responses to this scale. The prior literature on the variable interpretations of verbal probability labels demonstrates that "likely" may mean 50% to me and 1% to you. Second, even within subjects the quantitative meanings of words such as "likely" can vary dramatically depending on the context. Thus, we think it is important to be extremely cautious before imputing absolute probabilities from responses on the word scale. If one's goal is to determine the perceived relative order of events, our data suggest that the linear word scale may be the best choice. However, if the goal is to elicit absolute probability estimates, we think the word scale is inappropriate, and would argue for the magnifier scale.

FUTURE WORK

Our study raises two important questions. First, did our subjects really understand the log scale represented in the magnifier? Almost all respondents put answers inside the magnifier when estimating the chances of events that most people would accept as being rare. Furthermore, the distribution of re-

sponses within the magnifier was appropriately shifted depending on the rarity of the event. For example, the distribution for the chance of parenting sextuplets was centered over the leftmost extreme of magnifier (i.e., 1 in 100,000), while responses for 20-year-olds' chance of dying in the next ten years were correctly distributed around 1 in 1,000.²⁰ However, using the magnifier sensibly—to express the belief that an event has a small but non-zero probability of occurring—is not the same as using the scale literally (i.e., understanding the interval properties of a logarithmic scale). While we can infer that the respondents understood the magnifier to mean "rare," and that moving toward its leftmost anchor meant increasingly rare, whether they understood that 10^{-5} is precisely 100 times smaller than 10^{-3} , for example, is an open question. This question, however, should not undermine what the magnifier contributes, and in fact, this same question about literal understanding could be raised about any of the scales examined.

Second, what constitutes validation for a scale measuring perception of event probability? Unfortunately, there is no accepted external criterion (i.e., "gold standard") for validating how well our scale really captures a person's perception of chance. In fact, none of the scales considered in our study—even the linear scale labeled with numbers—has undergone formal study, with the exception of Diefenbach's study.¹ We considered several approaches to validation. First, we considered using an objective calculation of the respondent's actual risk. Comparison with this standard, however, would measure how accurately people estimate event probabilities rather than how well the scale captures their *perceptions* of those probabilities. Second, we considered using behavior—whether or not a respondent took action to reduce his or her risk (e.g., had mammography). This standard also seemed inappropriate, since behavior reflects not only perceived probability of an outcome but also its utility (i.e., how a person values an event, whether it is perceived as particularly dreadful, avoidable, or controllable²¹⁻²³). Instead, we based our approach on the work of Diefenbach et al.¹ While this approach is not perfect, we believe it has strong face validity. We used an individual's direct ranking to establish his or her relative ordering of the likelihoods of six events. Each scale was judged by how well this ordering was preserved when the individual quantified each probability on the scale.

Perceptions of chance—the likelihood of disease, the probability of cure—are of fundamental importance to patients and clinicians. Valid, reliable, and usable methods for assessing patients' perceptions of chance are needed to gauge whether their decisions are informed by realistic perceptions and to

evaluate the effects of educational interventions on these perceptions. Such methods are also important to quantitative decision researchers, since valid probability judgments are a fundamental input to utility assessment. Our results indicate that the new magnifier scale has the same validity, reliability, and usability as the linear number scale in helping people quantify their perceptions of probability. The magnifier scale has the important advantage of allowing people to express perceptions of low-probability events in a range practically inaccessible on standard linear scales.

An analysis of the data collected at Carnegie Mellon University was presented in Stephanie Byram's PhD dissertation at Carnegie Mellon University, Department of Social and Decision Sciences, April 1998. The authors thank Therese Stukel, PhD, and George Wolford, PhD, for their statistical expertise and Mary W. Heath for technical assistance.

References

- 1 Diefenbach MA, Weinstein MD, O'Reilly J. Scales for assessing perceptions of health hazard susceptibility. *Health Educ Res* 1993;8:1281-92.
- 2 Fischhoff B, Bruine WD. Fifty/fifty = 50%? *J Behav Decis Making* 1999;12:149-63.
- 3 Linville PW, Fischer GW, Fischhoff B. AIDS risk perceptions and decision biases. In: Pryor J, Reeder G (eds). *The Social Psychology of HIV Infection*. Hillsdale, NJ: Erlbaum; 1993:5-38.
- 4 Quadrel MJ, Fischhoff B, Davis W. Adolescent (in)vulnerability. *Am Psychol* 1993;48:102-16.
- 5 Budescu DV, Rapoport A, Zwick R, Forsyth B. Measuring the vague meanings of probability terms. *J Exp Psychol Gen* 1986;115:348-65.
- 6 Woloshin KK, Ruffin MT, Gorenflo DW. Patient's interpretation of qualitative probability statements. *Arch Fam Med* 1994;3:961-6.
- 7 Bryant GD, Norman GR. Expressions of probability: words and numbers [letter]. *N Engl J Med* 1980;302:411.
- 8 Toogood JH. What do we mean by "usually"? *Lancet* 1980; 1:1094.
- 9 Kong A, Octo-Barnett G, Mosteller F, Youtz C. How medical professionals evaluate expressions of probability. *N Engl J Med* 1986;315:740-4.
- 10 Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics, 1998. *CA-Cancer J Clin* 1998;48:6-30.
- 11 Lichtenstein S, Slovic P, Fischhoff B, Layman M, Combs B. Judged frequency of lethal events. *J Exp Psychol: Hum Learn Mem* 1978;4:551-78.
- 12 Poulton EC. *Bias in Quantifying Judgement*. Hillsdale, NJ: Erlbaum; 1989.
- 13 Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. *Ann Intern Med* 1997;127:966-72.
- 14 Black WC, Nease RF, Tosteson ANA. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. *J Natl Cancer Inst* 1995;87:720-31.
- 15 Dominitz J, Manski CF. *Perceptions of economic vulnerability*. Institute for Poverty Research (DP-1069-95). Madison, WI: University of Wisconsin, 1995.
- 16 Viscusi WK. *The risk of smoking*. Cambridge, MA: Harvard University Press; 1993.
- 17 Phillips K, Glendon G, Knight JA. Putting the risk of breast cancer in perspective. *N Engl J Med* 1999;340:141-4.
- 18 Grimes D, Snively G. Patients' understanding of medical risks: implications for genetic counseling. *Obstet Gynecol* 1999;93:910-4.
- 19 Budescu DV, Wallsten TS. Processing linguistic probabilities: general principles and empirical evidence. *Psychology of Learning and Motivation* 1995;32:275-318.
- 20 National Center for Health Statistics. Unpublished mortality data (<http://www.cdc.gov/nchswww/datawh/statab/upubd/mortabs.htm>) 1999.
- 21 Fischhoff B, Slovic P, Lichtenstein S, Read S, Combs B. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences* 1978; 8:127-52.
- 22 Slovic P. Perception of risk. *Science* 1987;236:280-5.
- 23 Fischhoff B, Bostrom A, Quadrel M. Risk perception and communication. In: Detels R, McEwen J, Omenn G (eds). *Oxford Textbook of Public Health*. London, U.K.: Oxford University Press; 1997:987-1002.