Statistical Science
1990, Vol. 5, No. 1, 2–34

Quantifying Probabilistic Expressions

Frederick Mosteller and Cleo Youtz

Abstract. For 20 different studies, Table 1 tabulates numerical averages of opinions on quantitative meanings of 52 qualitative probabilistic expressions. Populations with differing occupations, mainly students, physicians, other medical workers, and science writers, contributed. In spite of the variety of populations, format of question, instructions, and context, the variation of the averages for most of the expressions was modest, suggesting that they might be useful for codification. One exception was *possible*, because it had distinctly different meanings for different people. We report new data from a survey of science writers. The effect of modifiers such as *very* or negation (not, un-, im-, in-) can be described approximately by a simple rule. The modified expression has probability meaning half as far from the appropriate boundary (0 or 100) as that of the original expression.

This paper also reviews studies that show stability of meanings over 20 years, mild effects of translation into other languages, context, small order effects, and effects of scale for reporting on extreme values.

The stem *probability* with modifiers gives a substantial range 6% to 91% and the stem *chance* might do as well if tried with *very*. The stems *frequent, probable, likely,* and *often* with modifiers produce roughly equivalent sets of means, but do not cover as wide a range as *probability*. Extreme values such as *always* and *certain* fall at 98% and 95%, respectively, and *impossible* and *never* at 1%.

The next step will be to offer codifications and see how satisfactory people find them.

Key words and phrases: Quantifying language, codifying language, meaning of qualitative expressions.

INTRODUCTION

In everyday language, people apply the expressions *always* and *certain* to events that occur in fewer than 100% of their opportunities; furthermore, on average, people regard *very high probability* as more likely than *almost certain,* a surprise to many of us. Communications that employ qualitative expressions for frequencies or rates of occurrence run the risk of being misunderstood, whether the emitters or receivers are physicians, patients, scientists, science writers, or other citizens. By associating numerical values with specific qualitative expressions we may ultimately improve communication. In the long run we plan to propose some codification. In this paper, we want to report what such expressions currently mean by summarizing results from many studies and giving new findings from science writers.

In this first treatment, our intended readers and users are scientists including statisticians, because their language often needs interpretation to wider audiences than specialists. Statisticians have special skills, interests, and stakes in communication of information about probabilities. They may also wish to participate in producing a codification of probability expressions.

Codifications have two main forms. First, in some areas of work a few standard expressions may be used for all situations to express degree of belief or relative frequency. For example Kent [33] codified some

Frederick Mosteller is Roger I. Lee Professor of Mathematical Statistics, Emeritus at Harvard University and Director of the Technology Assessment Group in the Harvard School of Public Health. Cleo Youtz is Mathematical Assistant in the Department of Statistics at Harvard University.

As this work is still in progress, readers are encouraged to send comments on the paper to Frederick Mosteller at the Department of Statistics, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

expressions in the intelligence field in terms of odds. He used as a base *almost certain, chances are good,* and *chances about even* and quoted odds of at least 9:1 in favor, at least 3:1 in favor, and 1:1 or a 50–50 chance, respectively, with the reverse odds going with corresponding statements about something not happening. Whether we like it or not, statisticians seem to have generated a similar codification that associates numbers less than 0.05 with *statistically significant* and numbers greater than 0.05 with *not statistically significant.*

A second form of codification would merely relate sets of probability expressions to certain ranges of numbers so as to offer somewhat more precision in language to those who wish to use the findings. Information in the present paper could be used for this purpose. Its presentation here invites preliminary discussion and criticism that could be the basis for additional work before firming up either form of codification.

Many people say that one cannot put a single number on a qualitative word. Actually one can put many numbers on a qualitative word, and that is one reason for pursuing such studies. For the current study, we selected 52 expressions from an initial list of about 300. Most of these 52 expressions have been studied by other investigators. Table 1 displays the average probability to the nearest whole percent for these expressions given in 19 other studies [1–19] and the current study [20] of science writers. Our references [1–20] at the end of this paper also include information about sample sizes and kinds of respondents participating in each study. As one might hope, the studies give similar, though not identical, results for the same expression when sampling and other sources of variability are considered.

The right-hand side of Table 1 gives the average probability reported by respondents in 20 studies, with subscripts indicating the studies or parts of studies reporting a given percentage. Because some studies used more than one set of instructions or samples for a given expression, we often list more than one response for a study. In addition to the specific result for each study, Table 1 also gives two averages across the studies and parts of studies. The first average is the unweighted (equally weighted) average for the studies, and the second gives the grand average when averages for the studies are weighted by the number of respondents. The choice of average does not make much difference in the summary of studies.

TWO EXAMPLES

Mapes [11] studied the expression *rare* in a medical context. He asked physicians to assess the probability of side effects from two drugs—a beta-blocker and an antihistamine. For the beta-blocker 59.4% of the physicians chose the category "less than 1 per thousand," whereas 20.7 percent chose this category for an antihistamine. Mapes suggested that the difference arises because the side effects from an antihistamine are much milder compared to those of a beta-blocker. His point is that the meaning of *rare* changes with context. On the other hand, physicians may actually have different perceptions of the rate of side effects of the two types of medication.

President Gerald R. Ford said that a swine flu epidemic in the 1976–1977 season was "a very real possibility." Boffey [23] got probability estimates from four experts: 2%, 10%, 35%, and "less than even." Boffey concluded from these estimates that the chances of the epidemic are "far lower than the official rhetoric . . . would lead one to expect." Thus he sees a disagreement between the meaning of the qualitative expression "very real probability" and an average of perhaps 20 to 25%. Mosteller [28] used data from Cliff [24] and Selvidge [17] to estimate a public meaning of "very real possibility" at 29%.

These two examples illustrate the variation in perception of the meaning of probability expressions.

CONTEXT

It has often been pointed out that the context employing counts, rates, or amounts can influence the estimate. For an example mentioned to one of the authors by Leo Crespi, "a handful of grapes" and "a handful of people on the beach at Coney Island" could imply very different estimates, the first in the range of 5 to 20 grapes, and the second hundreds or even thousands of people. This problem, of course, complicates matters for probabilistic expressions. We think of the studies discussed here as dealing with probabilities in the region from 1 to 99%. The probabilities for very rare events, such as atomic disasters and regional blackouts from failure of the power system, are very difficult to communicate to laymen; and, as we understand, much remains to be learned. Communicating the idea of very small risks is an important problem in the risk field; it deserves and has its own research effort (Slovic, Fischoff and Lichtenstein [31]).

Some authors have emphasized the differences that changing context can create. For example, in a study of formulation of propositions, Tversky and Kahneman [32] show very substantial changes in preference between a certainty and an equivalent gamble by formulating the same possibilities in terms of lives saved versus lives lost. Kahneman, Slovic and Tversky [27] include articles by many authors showing effects of context. Our emphasis is more on the near constancy of opinions as illustrated in Table 1, rather than the differences, though the right-hand side of

TABLE 1

Average probabilities expressed in percentages for 52 expressions from 20 studies

Unweighted average	Weighted average	
98	99	Always $94_{1,2}$ 96_{5} $97_{15a,15b}$ $98_{16a,16b}$ $99_{7,18a^*,18b^*,20}$ $100_{3^*,4a,4b,4c,16c,19}$
91	91	Almost always 88_{5} 89_{19} 90_{15b} 91_{20} 93_{15a} 94_{7}
95	97	Certain 91_{10} $95_{1,8a}$ $98_{14^*,20}$
86	86	Almost certain 78_{8c} 82_{8b} $90_{8a,13^*,20}$
81	81	Very frequent 81_{20}
55	61	Frequent 36_{11b} 39_{11a} 56_{8b} 66_{12b} $67_{12a,20}$
45	45	Not infrequent 45_{20}
17	17	Infrequent 16_{12a} 17_{20} 19_{12b}
7	7	Very infrequent 7_{20}
91	91	Very high probability 90_{17a^*} $91_{17b^*,20}$
84	81	High probability 80_{20} 87_{1}
56	52	Moderate probability 51_{20} 61_{1}
16	16	Low probability 16_{20} 17_{1}
6	6	Very low probability 6_{20}
82	85	Very likely 74_{6} 79_{10} $85_{8a,13^*,20}$ 87_{9}
69	69	Likely 61_{6} $63_{8b,8c}$ 65_{10} $70_{13^*,20}$ 72_{9} 73_{1} 74_{2} 80_{14^*}
17	16	Unlikely $10_{17a^*,17b^*}$ 14_{8b} 15_{13^*} 17_{20} 18_{9} $20_{1,2,14^*}$ 23_{6} 25_{10}
11	8	Very unlikely 6_{20} 9_{9} 10_{13^*} 14_{10} 15_{6}
82	85	Very probable $79_{6,10}$ 80_{13^*} $87_{9,20}$
70	69	Probable 62_{6} 64_{8c} 65_{8b} $70_{13^*,20}$ $71_{9,10}$ 72_{2} 77_{1} 80_{14^*}
16	15	Improbable 12_{9} 13_{8b} 15_{13^*} $16_{6,20}$ 17_{2} 18_{10} 20_{14^*}
7	6	Very improbable 5_{13^*} 6_{20} 11_{10}
85	87	Very often $82_{4b,20}$ 87_{3^*} $88_{18a^*,18b^*}$
65	69	Often 50_{17b^*} 57_{16c} $59_{15b,16a,19}$ 60_{17a^*} $61_{7,16b}$ 62_{12a} 64_{12b} $70_{4a,20}$ 71_{15a} $73_{2,18b^*}$ 74_{3^*} 75_{5} 78_{18a^*}
62	61	More often than not 59_{20} 62_{18b^*} 64_{7}
50	50	As often as not $50_{5,18b^*,20}$
38	38	Less often than not 38_{20}
15	14	Not often 11_{18b^*} 13_{18a^*} 15_{5} 16_{3^*} 19_{20}
13	13	Not very often 13_{20}
42	37	Possible 27_{8b} 33_{20} 37_{9} 38_{2} 40_{13^*} 43_{1} 51_{6} $55_{10,14^*}$
1	1	Impossible $1_{14^*,20}$
78	81	High chance 72_{10} 80_{13^*} 82_{20}
58	58	Better than even chance 58_{20}
50	50	Even chance $50_{13^*,20}$
41	41	Less than an even chance 41_{20}
14	13	Poor chance 13_{20} 14_{10}
15	13	Low chance 10_{20} 20_{13^*}
66	66	Liable to happen 66_{20}
36	36	Might happen 36_{20}
77	79	Usually 70_{15b} 71_{19} 74_{20} 76_{7} 77_{9} 78_{2} 79_{3^*} 80_{18b^*} 84_{15a} 85_{18a^*}
19	19	Unusually 19_{20}
28	26	Sometimes 19_{15b} $20_{12a,18a^*}$ $22_{12b,15a}$ 27_{18b^*} 28_{20} 29_{3^*} 32_{1} 33_{7} 34_{16a} 36_{4c} 37_{16c} 38_{16b}
19	17	Once in a while 15_{18a^*} 16_{18b^*} 18_{20} 22_{3^*} 24_{5}
39	37	Not unreasonable 32_{8b} 39_{20} 47_{1}
22	22	Occasionally 17_{15b} $20_{7,18a^*}$ $21_{12a,12b,15a,19,20}$ 23_{18b^*} 24_{4b} $28_{3^*,4a}$
26	23	Now and then 18_{20} 20_{18a^*} 25_{18b^*} 32_{5} 34_{3^*}
12	12	Seldom 7_{15b} 8_{15a} 9_{3^*} $10_{4a,18a^*,18b^*}$ 13_{20} $16_{2,9}$ 18_{5}
7	6	Very seldom $6_{18a^*,18b^*,20}$ 7_{3^*} 12_{5}
9	7	Rarely $5_{3^*,7,15a,18a^*,18b^*,19}$ 7_{20} 8_{2} 9_{15b} 12_{16b} 15_{16a} 23_{16c}
4	4	Very rarely 4_{20}
4	3	Almost never $2_{3^*,8d}$ $3_{18a^*,18b^*}$ 4_{20} 7_{5}
1	1	Never $0_{2,3^*,4a,4b,7,18a^*,19,20}$ $1_{5,15a,15b,18b^*}$ 3_{8d}

The subscripts indicate studies listed in the reference list. a, b, c and d indicate different instructions or sample sizes.

* Indicates median.

Table 1 illustrates differences owing to samples, instructions or context.

In comparing context results with no-context (in isolation) results for 22 expressions, Selvidge [17] found that differences rounded to the nearest 5% for medians were 0 for 10 expressions, 5 for 7 expressions, 10 for 4 expressions, and 15 for 1 expression. The largest difference was for *appreciably,* which went from 25% in isolation to 10% in context.

Pepper and Prytulak [30] offer an elaborate investigation for studying effect of context by creating low-, moderate- and high-frequency contexts for framing the same expressions, as well as evaluation in isolation. The main source of divergence was that probabilities in the context of airplane crashes and of the occurrence of earthquakes were far removed from the numerical values given the same expressions in remarks about students missing breakfasts or the proportion of men that Miss Sweden thought found her attractive. The paper illustrates that context can push the meaning a good way, but that for ordinary events the differences are modest. We emphasize again that we are not considering numerical evaluations of probabilistic expressions in such contexts as very rare events occurring in short time periods. Context takes several forms in addition to the substantive topic, such as type of scale used and order of presentation. We discuss these matters further in the section on special topics.

SCIENCE WRITERS

We gathered data on meanings of the 52 qualitative expressions from science writers through a mail questionnaire. For a situation without context, the writers gave estimates for probabilities for the 52 expressions and lower and upper limits they thought their readers would set for each expression. The response rate was about 37%. The average responses are similar to those from other studies reported in Table 1. The relation of the ranges reported for the estimates is informative in spite of the low response rate. Some background for the special study of the science writers and the selection of the 52 expressions appears in the Appendix.

RESULTS

Median, Quartiles and Variability

In Table 2 we give the median, quartiles, and interquartile range for the distribution of the science writers' own point estimates for each expression. We computed medians and quartiles for frequency distributions from Form A and Form B (see Appendix) and then averaged results for the two forms.

As one expects, extreme expressions like *always* and *never* have small variation as measured by the interquartile range (IQR), the distance between the quartiles of the cumulative distribution. As a measure of variability of frequency distributions of the science writers' "own estimates," we use the IQR, because it is less sensitive than the standard deviation to extreme values. More centrally located expressions (nearer 50%) usually have broader variation.

Examples of Distributions

Figure 1 shows relative frequency distributions (using class intervals of length 10 centered on certain multiples of 5) for the science writers' personal choices for *possible, almost always,* and *unlikely.* We chose these expressions to give a notion of the variety of distributions encountered. To get a feeling for the behavior and systematic movement of distributions as the modifiers for a stem change, running down the median and quartiles in Table 2 for a given stem quickly shows how the middle 50% of the distribution moves. Figure 1 shows that *possible* has a bimodal distribution. This bimodality already suggests that *possible* is unsatisfactory as a qualitative expression. We return to it below. Other expressions were not bimodal (except for such minor variations for a few expressions as expected in histograms).

Relation of Variability to Level

Figure 2 shows a plot of the IQR against the median M. Because of anchoring at the end-points, we expect the IQR to rise as the median moves from the extremes toward 50%, and thus we expect a cap shape. This idea is borne out fairly well in the intervals $0 \le M \le 33$ and $67 \le M \le 100$, but not in the central interval $33 < M < 67$, as we now discuss.

The Middle Expressions

The 11 expressions with median probabilities between 33 and 67 sort themselves into groups whose variability differs.

Group 1 contains numbers that either compare directly with the 50–50 situation (*less than an even chance, even chance, better than even chance*) or compare the chance of the event with that of the nonevent (*less often than not, as often as not, more often than not*). These expressions lead to small interquartile ranges. The six dots at the base of the middle panel of Figure 2 represent them. Essentially the value 50% is anchoring the responses, especially with *even chance* and *as often as not.* We might get similar anchoring effects elsewhere if we asked how often respondents thought "about once in *n* times" occurred. Group 1 expressions offer high precision and are therefore potentially attractive for codification.

TABLE 2

Quartiles, median and interquartile range for the science writers' own preferred estimates for 52 *probability expressions, pooled from distributions produced by Form A and Form B*

Expression	25%	Median	75%	IQR
Always	99.6	99.7	99.8	.3
Almost always	89.7	91.7	95.2	5.5
Certain	98.7	99.6	99.8	1.1
Almost certain	87.5	90.2	95.0	7.5
Very frequent	75.3	82.6	89.7	14.5
Frequent	60.0	72.2	75.3	15.2
Not infrequent	32.7	49.6	57.3	24.6
Infrequent	10.1	17.3	22.6	12.5
Very infrequent	3.6	5.2	10.0	6.4
Very high probability	89.8	92.5	95.2	5.4
High probability	77.1	82.3	87.2	10.1
Moderate probability	40.1	52.4	58.7	18.5
Low probability	7.8	15.0	22.3	14.5
Very low probability	1.9	4.9	7.6	5.7
Very likely	80.1	87.5	90.2	10.1
Likely	62.6	71.1	77.6	15.0
Unlikely	9.8	17.2	22.7	13.0
Very unlikely	2.7	5.0	9.8	7.1
Very probable	81.5	89.7	90.4	8.9
Probable	64.7	70.2	77.7	13.0
Improbable	7.6	12.5	22.3	14.7
Very improbable	1.5	4.8	7.5	5.9
Very often	77.5	82.8	89.9	12.4
Often	65.0	72.5	75.4	10.4
More often than not	57.1	59.8	60.4	3.3
As often as not	49.8	50.0	50.3	.6
Less often than not	34.8	40.0	42.7	7.9
Not often	10.3	19.7	24.8	14.5
Not very often	5.3	10.1	19.6	14.3
Possible	7.5	38.5	50.2	42.7
Impossible	.2	.3	.5	.3
High chance	77.5	80.4	89.1	11.7
Better than even chance	53.3	57.6	60.2	6.9
Even chance	49.7	50.0	50.2	.5
Less than an even chance	39.6	40.2	45.0	5.4
Poor chance	8.4	10.3	19.7	11.3
Low chance	5.0	9.8	12.8	7.8
Liable to happen	59.8	68.2	77.7	17.9
Might happen	19.9	37.6	50.1	30.2
Usually	65.6	75.1	82.2	16.7
Unusually	9.9	17.4	26.1	16.3
Sometimes	17.5	25.0	35.0	17.5
Once in a while	9.9	15.3	22.4	12.5
Not unreasonable	23.5	37.6	52.6	29.1
Occasionally	12.5	20.0	27.7	15.2
Now and then	9.8	15.1	25.0	15.1
Seldom	7.4	10.2	17.5	10.1
Very seldom	3.2	4.9	7.7	4.5
Rarely	3.6	7.2	10.0	6.5
Very rarely	1.2	3.0	5.0	3.8
Almost never	1.2	2.9	4.6	3.4
Never	.1	.3	.4	.3

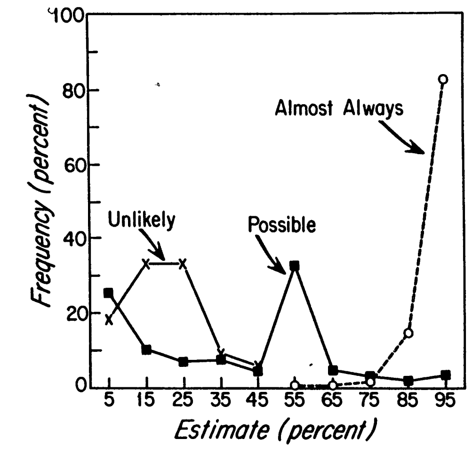

FIG. 1. *Frequency distributions of science writers' own estimates for three expressions, gouped by* 10% *intervals centered at multiples of* 5%.

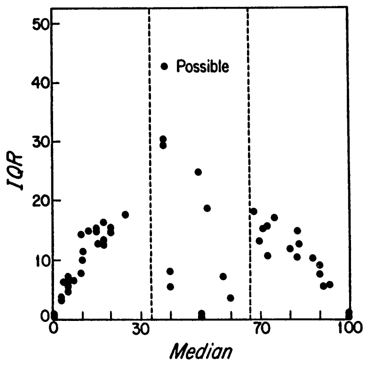

FIG. 2. *Plot of interquartile range versus median of science writers' own estimates for the* 52 *expressions.*

Group 2 (*not infrequen't, moderate probability*) has an IQR about the size one would expect from extrapolation from the two outer thirds of Figure 2.

Group 3 offers extra variation, increasing from *not unreasonable* and *might happen* to *possible.* Because it seemed a logical rather than a quantitative expression, we did not regard the first of these as a probabilistic expression, nor was *might happen* a cheerful choice, but they were included for examination because of the scarcity of expressions in the interval 33 to 67. Finally *possible* seems excellent, at first blush, for making a very vague statement, but recall the bimodality from Figure 1.

The double negative in *not unreasonable* and *not infrequent* makes them unattractive for codification, and we will see undesirable quantitative features of them in the next section.

Acceptability Functions for the Range Data

Although Table 2 and Figure 2 give an idea of the central value and variability among writers of the science writers' personal estimates, they do not make allowance for the variability (the ranges) that the writers report for these expressions. Therefore we introduce a notion of an acceptability function (used in Reagan, Mosteller and Youtz [13]) for each expression, ranging over the possible percents 0 to 100 as before. The value of the function at a given percent is the proportion of respondents who include that percent in their range for the expression. (We use *proportion* in speaking of the fraction of respondents and *percent* for the probability of the expression to distinguish more readily between two measures that could both be called percentages or both be called proportions.)

Although ranges of estimates for probability expressions have been gathered before [1,6,7,9,16], we analyze them in more detail than usual. For each expression, the respondent gave a lower bound and an upper bound that estimated the extremes for readers of the respondent's work. To combine these to give a notion of overall acceptability for any expression, we computed for each whole percentage from 0 to 100 the proportion of respondents who included that percentage in the interval for the expression. This produced graphs—we give points only at multiples of 5%—such as Figure 3. The graph then represents an acceptability function, and it shows the region where respondents concentrate their belief in the appropriateness for a given expression. The graph also shows how strong and how concentrated their belief is by the height and narrowness of the figure.

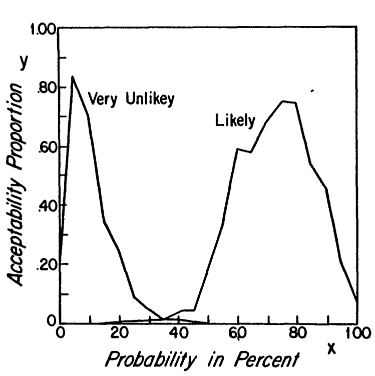

FIG. 3. *Acceptability functions for* likely *and* very unlikely. *The height of the acceptability function at x is the proportion of respondents who include the percentage x in their interval.*

As they stand, these graphs are not relative frequency distributions. To obtain medians and quartiles for these distributions, we first standardized them so that the total mass associated with the percents is 100. Then we treated the standardized results as if they were distributions and obtained medians and quartiles for the graphs for Form A and Form B, and finally averaged the results for the two forms to get a single pooled estimate for the quartiles and median of the combined result.

It is useful to know how tall the original graphs are: that is, what is an expression's modal acceptability. To smooth away local fluctuations and number preferences, we wanted a measure for an interval, not just the percent with the highest acceptability. To measure this, we looked at intervals of length 6% between and including adjacent multiples of 5. For most expressions we used the intervals 5–10, 10–15, ..., 90–95. We averaged the acceptability proportion for the six percentages in an interval and took as our measure of modal acceptability the largest average of the proportions among the sets. Near the extremes it seemed better to give the average for 1–5 or for 95–99 in a few instances (namely, *very rarely* and *almost never*); just the value at 0 or 100 for *always, never, certain,* and *impossible;* and the value at 50 for *even chance.* We did this for Form A and Form B and then averaged the modal acceptabilities.

When plotted against the median, the modal acceptability (not shown) looks like an upside-down version of Figure 2. Similarly, a plot of the IQR of the acceptability functions against their medians gave a picture (not shown) like Figure 2.

Some expressions with exceptionally high modal acceptability or exceptionally high IQR are shown in Table 3, together with their modal acceptability and their IQR. The expressions with highest modal

TABLE 3
Expressions with unusually large acceptability mode or large IQR

		Acceptability	
	Median	Mode	IQR
Large acceptability mode			
Even chance	50	99.4	5.9
Always	96	98	6.3
Never	2	97	5.4
Impossible	4	93	6.9
Certain	95	92	8.1
Large IQR			
Liable to happen	71	64	24.1
Sometimes	31	58	25.6
Not infrequent	47	51	27.2
Not unreasonable	44	42	31.7
Might happen	43	51	35.0
Possible	47	55	42.4

acceptability are those that are well anchored either in the middle or at the extremes: *even chance, always, certain, impossible,* and *never.* The largest IQR goes with *possible,* and *might happen* is the nearest competitor. The lowest modal acceptability values go with *not unreasonable, might happen, not infrequent,* and *possible;* we note that this set includes the two double negatives.

We studied *possible* more deeply by plotting a respondent's range for *possible* against the midrange (average of the respondent's bounds). The resulting clusterings for *possible* show that not only do many respondents feel that *possible* corresponds to any number between 0 and 100 (a reasonable literal interpretation), but others associate it with an event so rare that it can scarcely occur (another reasonable interpretation, as in *barely possible*), whereas a few assign it a 50–50 chance exactly (somewhat a surprise), others about 25 percent with a large range, and still others are scattered all over the lot. Such distinct clustering seems to say that interpretations of *possible* differ substantially and that *possible* has distinct meanings for different groups—at least at a given moment. Thus it does not have a homogeneous meaning of broad range, and so its use may mislead some readers, and what looked good for vagueness "at first blush" has not survived closer examination. We made range-versus-midrange plots for all the expressions but found little notable about the others.

Earlier and Later Responses

With large non-response, it is customary in sample surveys to check for changes in response between early and late respondents (Mosteller [29], page 215; Bartholomew [21]; and Hansen and Hurwitz [26]). For each expression, we compared the average probability for those who responded before the second mailing with the average probability for those responding afterward. To look for some systematic trend, we divided expressions into high (average probability over 50%) and low (the rest). Then we scored each expression as going "up" or going "down" from early to late. The resulting 2 × 2 table exhibited a slight tendency for "high" expressions to go "down" and "low" expressions to go "up." A chi-squared test corrected for continuity fell almost exactly at the 0.05 level, 3.86 instead of the tabled 3.84. Although this might sound like a regression effect, the grouping "high" versus "low" is chosen on the basis of all respondents, and so the selection explanation would not be very relevant. The average reduction for the "high" group is 0.4%, and the average increase for the "low" group is 0.5%. Rough calculations suggest that 0.4% is about one standard deviation for the average of either the upper or lower half of the expressions, and so the changes observed are comparable to those that sampling fluctuations would suggest. The absolute size does not seem substantial, and so the hope is that non-response is not highly related to the probabilities science writers attach to expressions.

SPECIAL TOPICS

Effect of Modifiers and Prefixes

Cliff [24] in a study of the meanings of expressions in a larger context found evidence that ratings of expressions were modified in an approximately multiplicative fashion when adverbs were introduced to modify adjectives. The ratings fell on a scale having a neutral zero point and running from negative (unfavorable) numbers through positive (favorable). For example, *very wicked* had a rating 1.25 times that of *wicked.* Thus the adverbs provided multipliers obtained by averaging the effects over the expressions where the adverbs were used.

In a similar spirit, we want to provide summary values of the effects of modifiers used in the 52 expressions, though we have to deal with a probability scale running from 0 to 100%. As a simple method of summarizing we use multipliers that relate to the distance to the nearer extreme, though we could rephrase the method so as to treat 50% like the zero point of the Cliff scale.

The modifiers form three classes of expressions: (a) those where "very" reduces a probability that is originally less than 50%, (b) those where "very" increases a probability that is originally greater than 50%, and (c) those where "not" (in-, im-, un-) changes an expression from originally greater than 50% to less than 50%. We find it convenient to work with distances from the extremes and to relate the initial distance from the nearer extreme (0% or 100%) to the final distance from its nearer extreme. When we do this, a single multiplicative parameter summarizes the effect of all the modifiers. If x is the initial value and x_{mod} is the modified value, and k is the multiplier, then the estimated relations for the three classes are:

(a) $$x_{\text{very}} = kx$$

(expressions where "very" reduces the estimate),

(b) $$100 - x_{\text{very}} = k(100 - x)$$

(expressions where "very" increases the estimate),

(c) $$x_{\text{not}} = k(100 - x)$$

(expressions of negation: not, in-, un, im-).

With these conventions, the single value $k = 1/2$, works well for all three groups using the unweighted averages (or weighted or the science writers' data) of all studies from Table 1. (We used the average log of the ratio of the second column to the first in Tables 4, 5 and 6

TABLE 4
The effects of modifiers on the average probability for various expressions. Class (a): *"Very" reduces the estimate.*

Class (a)	x	x_{very}	$\frac{1}{2}x$
Infrequent	17	7	8.5
Low probability	16	6	8
Unlikely	17	11	8.5
Improbable	16	7	8
Seldom	12	7	6
Rarely	9	4	4.5

TABLE 5
The effects of modifiers on the average probability for various expressions. Class (b): *"Very" increases the estimate.*

Class (b)	$100 - x$	$100 - x_{\text{very}}$	$\frac{1}{2}(100 - x)$
Frequent	45	19	22.5
High probability	16	9	8
Likely	31	18	15.5
Probable	30	18	15
Often	35	15	17.5

TABLE 6
The effects of modifiers on the average probability for various expressions. Class (c): *Negation sends the estimate toward the opposite end of the scale.*

Class (c)	$100 - x$	x_{not}	$\frac{1}{2}(100 - x)$
Frequent	45	17	22.5
Probable	30	16	15
Likely	31	17	15.5
Often	35	15	17.5
Very often	15	13	7.5[a]
Usually	23	19	11.5[a]

[a] Substantial deviation.

to get the 0.5.) A multiplier of ½ for the change in distance to the nearer extreme works well for nearly all expressions, as comparison of the last two columns for each class shows.

In class (c) *very often* (converting to *not very often*) and *usually* (converting to *unusually*) have large residuals for reasons each reader will be able to speculate about. Outlier analysis does not suggest them to be extreme, nevertheless some of us will feel that *unusually* is not a satisfactory negation of the probabilistic meaning of *usually* but has an extra different slant. Similarly, the *very* of *not very often* seems to soften the *not* rather than to strengthen the *often*. At any rate, whether these are included or not, a value of $k = \frac{1}{2}$ is a useful rule of thumb.

Reyna [14] also studied the effects of negation on probabilities of things happening. Like ours, many of her subjects did not assign the extreme probabilities to expressions implying certainty. For example, 59% did not assign 0 to *impossible*.

Stability over Time

Simpson [18] studied the variability of quantitative meanings for 20 qualitative terms over time by comparing responses to two questionnaires, one completed in 1942 by 335 students and the other in 1962 by 395 students. He used the midpoint of the range given by the student for each expression as its location value. For the 20 terms, the average absolute deviation between the means of the midpoints for the 1942 and 1962 responses was only 2.0%. Thus the results are strikingly similar.

Translation

Grigoriu and Mihaescu [25] translated 30 probability expressions into Romanian equivalents and found "the average numerical values were similar for different professional groups [physicians, medical students, and medical related professionals] and very close to the values reported in the English literature" (page 364). Among 16 expressions that appear both in their study and among the 52 entries in Table 1, the unweighted averages given in Table 1 differed by 10 percentage points or more for 5 expressions.

Context and Translation

Beyth-Marom [22] reports an elaborate set of experiments carried out in a forecasting organization in Israel with 30 expressions in Hebrew. These were translated into English, and 8 are among our 52. With respect to context, she found more variation in the numerical evaluations of expressions when they were given in a context of the likelihood of future specific events than when merely given as expressions to be evaluated (she speaks of presenting expressions in isolation). The expressions in common with our study produced numerical probabilities in isolation close to ours, in spite of the translation problem.

Order

Order of presentation is one kind of context, and our Form A versus Form B results offer some information on this topic. For five high-probability expressions with averages between 90 and 100, the gain was small for Form A over Form B, averaging just 0.3%. For the 17 expressions with averages between 50 and 90%, the gain for Form A over Form B averaged 2.5%. For 8 expressions with averages between 20 and 50%, the gain averaged 1.3%. For 22 expressions with averages between 0 and 20%, the average gain was negative, −0.3%. Thus, for the most part, Form A (listing from high to low) led to slightly higher averages than Form B (low to high), except for the very-low-probability expressions, where Form B tended to produce slightly higher values. We were pleased that order had not mattered much, though we feared it might and that fear motivated the two forms.

Scale

Similarly Kong, Barnett, Mosteller and Youtz [8] found that the scale offered to the respondent made some difference in the average estimate. For example, for *almost certain* subjects who were offered a free choice averaged 78%; those offered choices on an equally spaced scale (by tens) averaged 82%, not much different; and those offered a spread-out scale emphasizing the very high percentages averaged 90%, substantially higher.

DISCUSSION

People often say that they would prefer to use the actual numbers if they were available rather than use qualitative expressions. Those who write about quantitative things will usually find that somewhere in the course of the exposition it is convenient and communicative to escape from the numerical mode and give a collective idea rather than a specific number—for example, to say of a collection of events that they often happen or that they rarely happen. In addition, it is not at all verified that lay readers acquire better information from a number than from a qualitative expression. That issue is open for research. It probably depends on the reader and the message being delivered.

We now turn back to the proposal for codification. Some progress suggested by the information produced by other studies and from the science writers suggests that several expressions among the 52 not be used for codification, especially those in the bottom panel of Table 3: *liable to happen* (with its overtone of risk), *sometimes, not infrequent* (double negative), *not unreasonable* (logical rather than probabilistic), *might happen,* and *possible* (because of many distinct but definite meanings). All have large IQRs and low acceptability modes.

For extreme probabilities such as 98% or 99%, *always* and *certain,* and for 2% and 1%, *impossible* and *never* are available.

The stem *probability* offers a good spread with its modifiers from about 5 to about 90%. We did not include *very* with the stem *chance,* but using our $k = 1/2$, we could impute about 90% to *very high chance* and about 7% to *very low chance* and the stem *chance* then would offer a range similar to that of the stem *probability.* The stems *frequent, likely, probable,* and *often* do not cover as wide a range, though they offer nearly equivalent sets of four expressions each.

Precisely 50% is well nailed down by *even chance* and *as often as not.* A similar value expressing more uncertainty would be delivered by *moderate probability.*

Statisticians may wish to participate in the next stage of assessment. Although popular views of the probabilities associated with these expressions should help guide final choices to avoid straining the language, the decisions to choose specific codifications of either kind must still be ones that scientists are comfortable with and ones that do not call for unnecessarily fine or too many distinctions.

APPENDIX

Choice of Expressions

The probability expressions included in this study emerged from searches and from papers sent to us by scholars. Augustine Kong in 1983 made a computer search for papers dealing with quantitative meanings of probabilistic expressions. He found 18 articles, and their references led to others. After publication of Kong, Barnett, Mosteller and Youtz [8], several readers kindly wrote us about additional papers. From about 40 articles we found more than 300 expressions, many being evaluated by a sample of respondents. Our colleague Timothy Reagan reviewed and sorted them into three categories: probability, frequency, and other. After reviewing Reagan's analysis, Lincoln Moses and F. Mosteller chose 52 expressions for this study. One aim was to choose a set of expressions whose associated numerical averages would cover well the range of probabilities (expressed in percentages) from 0 to 100%.

Respondents

Most respondents in the studies that we summarize in Table 1 were students, physicians, and health workers. New data in the current study come from members of an association of science writers. Their views are important because science writers communicate information from scientists to the public, including other scientists.

With the cooperation of the governing body of the Council for the Advancement of Science Writing, Inc., we informed members of the National Association of Science Writers (NASW) through their newsletter of the proposed survey of science writers. Barbara J. Culliton, President of the Council, wrote a letter to accompany our questionnaire sent to the NASW members.

The Questionnaire

Respondents were asked, first, to give the probability (as a percentage expressed to the nearest whole number from 0 to 100) that they personally would attach to each of the expressions. (Respondents in the studies in Table 1 often answered such a question.) Second, they were asked to give the range of probabilities that they thought their readers would associate with that expression. The range gives some idea of the variability associated with an expression; substantial disagreements about meaning could raise questions about the value for codification. Although one does not expect delicate distinctions from such an inquiry, it seems very different to offer 25–30 as a range as opposed to 5–85, for instance.

The questionnaire grouped the expressions by stems, as in the set based on the stem *likely: very likely, likely, unlikely,* and *very unlikely.* The stems were *always, certain, chance, frequent, happen, likely, never, often, possible, probability, probable, rarely, seldom,* and *usually.* Based on their grammatical meanings, the expressions were ordered for each stem. If the order had been haphazard, or the items spread through the questionnaire, conscientious respondents would have required a long time to complete the task, because to be consistent they would have to hunt up their estimates for other expressions with the same stem.

We used two forms of the questionnaire. In Form A expressions for each stem were ordered from high probability to low, and in Form B from low to high. Arbitrary choices of ordering might make a difference, and a design with a balancing approach using also the reverse order offered some protection against bias. We analyzed the forms separately and then combined the results.

The Mailings and Response Rate

On May 6, 1987, we mailed 637 questionnaires to members of the NASW in the United States and Canada. On June 4 we sent a follow-up mailing to the 475 who had not yet responded. From the two mailings we received 238 replies with mainly usable responses; about 5% were either returned undelivered or returned blank by the science writers.

Editing the Responses

Under the most carefully controlled conditions, respondents do as they please, as experimenters and survey scientists well know. Although we asked the respondents to give estimates in whole numbers, some gave answers such as 0.001, 0.02, <50, >70, 1+, 60–, 10–20, which required editing. Some changes were simple. For example, we changed 0.0001 to 0 and 10–20 to the midpoint, 15. When we could not make a reasonable adjustment, we changed the response to a blank.

In work with scales in questionnaires, it is a familiar finding that respondents occasionally get turned around and give the complement of the number they intend. For example, in our study *always* might have produced a personal estimate of 2% and a range of 0–5%, whereas the respondent intended 98% and 95–100%. When such errors occurred on the more extreme expressions, we changed the estimates to their complements.

ACKNOWLEDGMENTS

We thank Constance Brown, Barbara Culliton, John D. Emerson, David C. Hoaglin, Lincoln E. Moses, Marjorie Olson, Timothy R. Reagan, Christopher Schmid, Katherine Taylor-Halvorsen and Wallace Waterfall for their many thoughtful comments and suggestions. This work was facilitated in part by Grant SES-84-01422 from the National Science Foundation.

REFERENCES AND NOTES

[1] Bryant, G. D. and Norman, G. R. (1980). Expressions of probability: Words and numbers. *New England J. Med.* **302** 411.
n = 32: 16 physicians responding twice each; averages read from chart.

[2] Budescu, D. V. and Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes* **36** 391–405.
n = 32: faculty and graduate students in psychology, University of North Carolina, Chapel Hill.

[3] Hakel, M. D. (1968). How often is often? *Amer. Psychol.* **23** 533–534.
n = 100: students in an introductory psychology course, University of Minnesota.

[4] Hartley, J., Trueman, M. and Rodgers, A. (1984). The effects of verbal and numerical quantifiers on questionnaire responses. *Appl. Ergonomics* **15.2** 149–155.
Undergraduate students, University of Keele: n_a = 20 students given one set of expressions. n_b = 20 different students given another set. n_c = 20 students given still another set. a: *always, often, occasionally, seldom, never.* b: *always, often, fairly many times, occasionally, never.* c: *always, quite often, sometimes, infrequently, none of the time.*

[5] Hartsough, W. R. (1977). Assignment of subjective probabilities to verbal probability phrases as a function of locus of control and set conditions. *J. Psychology* **95** 87–97.
n = 60 (10 each in 6 groups): students in an introductory psychology course.

[6] Johnson, E. M. (1973). Numerical encoding of qualitative expressions of uncertainty. Technical Paper 250, U.S. Army Research Institute for the Behavioral and Social Sciences, AD 780 814.
n = 28: 14 U.S. Army enlisted men and 14 extension college students.

[7] Kenney, R. M. (1981). Between never and always. *New England J. Med.* **305** 1097–1098.
n = 24: members of Pathology Department, Massachusetts General Hospital.

[8] Kong, A., Barnett, G. O., Mosteller, F. and Youtz, C. (1986). How medical professionals evaluate expressions of probability. *New England J. Med.* **315** 740–744.
$n_a \approx 140$. High probability scale. $n_b \approx 134$. Uniform scale. $n_c \approx 170$. Free choice. $n_d \approx 144$. Low probability scale. Physicians, medical, non-medical, and auxiliary students and nurses, nationwide.

[9] Lichtenstein, S. and Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Sciences* **9** 563–564.
$n \approx 186$: System Development Corporation employees.

[10] LEVINE, J. M. and ELDREDGE, D. (December 1970). The effects of ancillary information upon photo interpreter performance. American Institutes for Research, Institute for Research in Psychobiology, Washington Office, AIR-20131-12/70-FR.
$n = 20$: enlisted U.S. Army image interpreters.

[11] MAPES, R. E. A. (1979). Verbal and numerical estimates of probability in therapeutic contexts. *Social Science and Medicine* **13A** 277–282.
$n_a = 29$: physicians given expression "Side effects with chloramphenicol are *frequent*." $n_b = 33$: physicians given expression "Side effects with neomycin sulphate are *frequent*."

[12] NAKAO, M. A. and AXELROD, S. (1983). Numbers are better than words: Verbal specifications of frequency have no place in medicine. *Amer. J. Med.* **74** 1061–1065.
$n_a = 103$ physicians. $n_b = 106$ physicians. Means read from chart.

[13] REAGAN, R. T., MOSTELLER, F. and YOUTZ, C. (1989). The quantitative meanings of verbal probability expressions. *J. Appl. Psychology* **74** 433–442.
$n = 115$: undergraduates in a psychology course, Stanford University.

[14] REYNA, V. F. (1981). The language of possibility and probability: Effects of negation on meaning. *Memory and Cognition* **9** 642–650.
$n = 41$ adult volunteers.

[15] ROBERTS, D. E. and GUPTA, G. (1987). To the editor. *New England J. Med.* **316** 550.
$n_a = 45$ house staff. $n_b = 24$ attending physicians.

[16] ROBERTSON, W. O. (1983). Quantifying the meanings of words. *J. Amer. Med. Assoc.* **249** 2631–2632.
$n_a = 53$: Seattle physicians. $n_b = 80$: graduate students at the University of Washington's School of Business Administration. $n_c = 40$: Board of Trustees at the Children's Orthopedic Hospital and Medical Center, Seattle.

[17] SELVIDGE, J. (1972). Assigning probabilities to rare events. Ph.D. dissertation, Graduate School of Business Administration, George F. Baker Foundation, Harvard Univ. Subjects were Harvard Business School students in MBA program. $n_a = 59$: Estimates made on basis of a statement without context. $n_b = 127$: Contexts were provided.
Also in Mosteller, F. (1977). Assessing unknown numbers: Order of magnitude estimation. In *Statistics and Public Policy* (W. B. Fairley and F. Mosteller, eds.) 163–184. Addison-Wesley, Reading, Mass.

[18] SIMPSON, R. H. (1963). Stability in meanings for quantitative terms: A comparison over 20 years. *Q. J. Speech* **49** 146–151.
1942 study. $n_a = 335$: 86 high school and 249 college students. 1962 study. $n_b = 395$ university students.

[19] TOOGOOD, J. H. (1980). What do we mean by "usually"? *Lancet* **1** 1094.
$n = 51$: physicians, nurses, laboratory technologists, secondary school teachers, and engineers.

[20] Current study, estimates from science writers.
$n \approx 230$: science writers. Varies somewhat from expression to expression, 211–237.

[21] BARTHOLOMEW, D. J. (1961). A method of allowing for "not-at-home" bias in sample surveys. *Appl. Statist.* **10** 52–59.

[22] BEYTH-MAROM, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *J. Forecasting* **1** 257–269.

[23] BOFFEY, P. M. (1976). Anatomy of a decision how the nation declared war on swine flu. *Science* **192** 636–641.

[24] CLIFF, N. (1959). Adverbs as multipliers. *Psychol. Rev.* **66** 27–44. This paper stimulated a four-paper symposium on quantification of the effect of adverbs on the meaning of adjectives: *Chance* **1** (3) 32–51 (1988).

[25] GRIGORIU, B. D. and MIHAESCU, T. (1988). Evaluarea numerica a expresiilor de probabilitate folosite in limbajul medical. *Rev. Med. Chir. Soc. Med. Nat. Iasi* **92** 361–364.

[26] HANSEN, M. H. and HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.* **41** 517–529.

[27] KAHNEMAN, D. SLOVIC, P. and TVERSKY, A., eds. (1982). *Judgment under Uncertainty: Heuristics and Biases.* Cambridge Univ. Press, Cambridge.

[28] MOSTELLER, F. (1976). Swine flu: Quantifying the "possibility." *Science* **192** 1286, 1288.

[29] MOSTELLER, F. (1978). I. Non-sampling errors. In *International Encyclopedia of Statistics* (W. H. Kruskal and J. M. Tanur, eds.) **1** 208–229. The Free Press, New York.

[30] PEPPER, S. and PRYTULAK, L. S. (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitative expressions. *J. Res. in Personality* **8** 95–101.

[31] SLOVIC, P., FISCHOFF, B. and LICHTENSTEIN, S. (1982). Facts versus fears: Understanding perceived risk. In *Judgment under Uncertainty: Heuristics and Biases* (D. Kahneman, P. Slovic and A. Tversky, eds.) 463–489. Cambridge Univ. Press, Cambridge.

[32] TVERSKY, A. and KAHNEMAN, D. (1981). The framing of decisions and the psychology of choice. *Science* **211** 453–458.

[33] KENT, S. (1949). *Strategic Intelligence.* Princeton Univ. Press, Princeton, N.J.

Comment

Herbert H. Clark

In the last few years, Mosteller, Youtz and their colleagues have looked at probability and frequency expressions such as *usual*, *very likely*, *improbable*, *frequent* and *as often as not*. Their interest is in how these terms are used in communicating technical information, and their goal is to better that communication, to make it more precise. Their project has two phases. In the first, they will determine what these terms mean to the people who use them. In their own study they have found, for example, that *frequent* is judged to represent an average proportion of about 0.72 of the time with an interquartile range of about 0.15. If you say something is frequent, they claim, you are saying that it occurs about 72% of the time plus

Herbert H. Clark is Professor of Psychology, Stanford University, Stanford, California 94305.