hope is" (page 18), leaving the reader quite puzzled concerning the probabilities the authors intend to imply.

ADDITIONAL REFERENCE

WOLF, C. (1987). Scoring the economic forecasters. *The Public Interest* No. 88 Summer.

Rejoinder

Frederick Mosteller and Cleo Youtz

Because our work on codification is still developing, we are most grateful to Carl Morris for arranging discussion by experts from various fields so that we can benefit from their suggestions and criticisms before we firm up our plans. (In what follows, "we" refers to Mosteller and Youtz.) We much appreciate the efforts of the discussants to help us avoid pitfalls and guide us to further material relevant to this work. In this paper we do not propose a codification but organize some material that may be useful in producing one. To treat all the issues raised by the discussants would require much more research than has been done in this field so far. We will, however, respond to the central issues.

We have no problem with the idea of including variability as part of the codification. For example, the data suggest that *even chance* has very little variability associated with it, whereas *possible* has huge variability. Some measures of properties of the acceptability function or of variability built up from variation within individuals, between individuals, and contextual sources offer options for presenting variability along with location (average value). How that information can be profitably communicated poses a question we have not yet settled. We illustrate one option when we discuss Kadane's comments.

Most of the discussants wish to emphasize the effects of context on meaning, and they do this in various ways. We will take up these matters as we go through the comments individually.

Wallsten and Budescu make four main points about the difficulties in trying to produce a codification, and they make some suggestions about such a program. "First, individual differences in the use and understanding of linguistic probability expressions are large, reliable, and probably very resistant to change." For us this offers a reason for including information about variability in a codification, and it explains why it may be important to know what sort of variability people trying to communicate need to face. Part of the effort in codification then should be to inform people about variation. "Second, probability phrases have vague meanings to individuals. Any attempt to render them precise will of necessity overlook the important semantic role of this vagueness." This additional variability again is something that has to be included in a plan to create a codification. "Third, context effects on the meanings of probability phrases are substantial and probably cannot be eliminated." As these authors suggest later in their discussion, "it is possible that a subset of phrases can be selected whose meanings are more or less agreed upon." Possibly some phrases are relatively resistant to context or perhaps we can get people to learn to standardize them. The procedure would still have to face the variability associated with individuals, and so it may well be that we have to learn how little can be communicated with probabilistic phrases because they often have broad ranges (considerable variability). We have not decided on a way to communicate these distributions, though we have used the interquartile range here as one method. "Finally, there is often a need to communicate not only a best probability estimate, but also information about the amount and nature of supporting evidence." Although this is true, it may represent a need that goes beyond the notion of a codification, just as a whole theory of probability and risk assessment may be required. Wallsten and Budescu suggest that a program that uncovers the various communication roles for probability phrases together with numerical techniques could make it possible for people to express information about the state of evidence and precision of their opinion. Such a substantial program goes beyond what we have in mind.

Winkler also mentions such an idea in his discussion. We like such a program, indeed, statisticians and psychologists and others have this program in mind in their teaching and research. After 100 years or so of work on it, we all have a long way to go. Our idea is much less ambitious.

Cliff, like Wallsten and Budescu, encourages us to pay more attention to the variability of the meanings of expressions and less to measures of location. As we mentioned above, we do not find this inconsistent with codification.

Cliff suggests that, first, isolation of communicators and, second, specificity of referent will be needed if we are to be successful. We discuss the first point because the second falls under the general heading of context. Because nearly everyone is a statistician some

of the time, it is hard to imagine bringing everyone into line with any proposed codification. Possibly subgroups though would find some standardization attractive just as they have found some mathematical and statistical language congenial. Discussions such as these make it possible to think about what features of a codification might add to its value and thus tend to make it catch on. We can explore further Cliff's suggestion about the effects of modifying stems. The modifiers we have used usually push the average for the modified expression nearer an extreme (zero or one) than was the average for the original expression. To the extent that this happens we would expect then a tightening of the distribution of values people assign just because of the anchoring effect of the extremes.

If a codification were to be based on one or a few stems, the set of modifiers chosen would be easy to remember and easy to interpret as an ordering, and so the idea of modifiers has attraction (we return to this in discussing Kadane's contribution). In closing, Cliff emphasizes the possibility that what we are doing is to make people more careful in reporting their probability statements, and that some of what we do may be helpful for this purpose. That goal includes a large part of what we have in mind.

Clark illustrates the importance of context in a variety of circumstances and puts his finger on a difficulty, distinguishing between word meaning and word use. Whereas we might like to believe that we can inquire about the numerical meaning of probability phrases in No Context situations, Clark warns that instead of No Context we may be dealing with Unknown Context and thus the truth may be more variable than we appreciate. Furthermore, he regards questions put in the No Context situation as questions the respondent treats as relative to a particular context C which of course may vary from time to time and from individual to individual. He suggests as a practical matter that we might be able to phrase questions that would go a ways toward standardizing the context C and then relating the meaning of the phrase to the percentiles of the distribution of possibilities in the context C.

Winkler gives us the benefit of one of his own instructional studies for his students, showing them the sort of variability associated with their interpretations of probability meanings. He too emphasizes the variability of the responses, and how surprised students are about this variability. Again, this encourages us to want to deliver more information about variability in a codification.

Winkler emphasizes that the communications being made reflect the subjective degree of belief of the respondent or speaker rather than frequencies for the events. We think that the situation is rather mixed and that sometimes people are trying to communicate information about objective numbers using qualitative expressions and sometimes about their personal degrees of belief, just as Winkler suggests. We might benefit by giving more thought to this distinction in the work we are doing.

In thinking about developing a codification, we have been torn between using the kind of data that people generate and choosing expressions and their properties from among these and the alternative of forcing expressions into somewhat different molds so as to get suitable coverage. The weather investigation reported by Winkler suggests that sticking with what people already do would be more acceptable and therefore more productive.

Winkler offers a way of thinking about the distribution of the probability of an event and its Bayesian revision as a result of new information. He speculates that the posterior distribution following the introduction of information in the form of a numerical probability will be tighter than the posterior that occurs when the new information is qualitative in form. This seems reasonable, at first blush, though we suppose that looking into it would require an effort to create comparability of some sort between the two kinds of information.

As others do, Winkler worries that even if the emitters of probability information through qualitative expressions that have been codified have standardized their behavior, the receivers may not have, and that the latter would not appreciate the codification. We might hope that if codification became popular, both, emitters and receivers in small communities, as envisaged by Cliff, might know what the codification was intended to produce. The others would be no worse off than they are now if we do not introduce special constraints.

As Wallsten and Budescu did, Winkler wants to press the program of training in the meaning of numerical probability expressions, and of course having spent much of our lives on this, we cannot fault it as one goal.

Tanur wants more attention paid to context, and she points to evidence that context does matter. First, she discusses the point that Clark emphasized that queries without context may produce different responses from the same questions with a context. Second, she discusses the possibility that different contexts may make substantial differences. We and the discussants describe some possible aid for these troubles through selection of expressions for resistance to context and through education of the communicators.

Wolf illustrates problems where it would be helpful to be able to move from qualitative probability expressions to quantitative ones and others where the reverse trip would be constructive. He too notes that it is instructive to know about the variability implied by an expression as well as the level. He thinks of this

work as a possible way of raising the level of understanding of vague probabilistic expressions. Perhaps his appraisal of the benefits of such studies is similar to that of Cliff's—that it may help us think more carefully about what we say about probabilities.

Probably we are not well enough prepared in languages to pursue except at second- or third-hand the story of translations and their effects as Kruskal would like to see done. On the issue of logical relations, we do know that negation of an event does not always change the probability from p to $1-p$. Various authors have commented on this. (Lichtenstein and Newman, 1967, suggested this, and also Kong, Barnett, Mosteller and Youtz, 1986, discuss it.) We discuss the matter in our section on modifiers. We note there that the negation seems to follow a rule of thumb that halves the distance of the complement to zero, with all our examples starting with a stem whose probability exceeds ½. Kruskal seems to be calling attention also to *better than even chance* and *less than an even chance* as another pair that might logically sum to 100 percent. They do not quite: the means from Table 1 are 58 and 41, the medians 57.6 and 40.2.

We will pursue the references Kruskal offers in hope of learning better how to implement codification.

Kadane produced the codification exhibited in his Table 2. We have sometimes carried out a similar activity. As Kadane notices, and we agree, some expressions seem not to be ones we want to use. He finds that he can represent 11 intervals fairly well by choosing one expression from each (except for the interval 0.95 to 1.0).

When we have carried out such an exercise, we have been concerned that the list is hard to remember, and that is why we thought that for this purpose it might be well to have a list or several lists each based on a single stem. One such stem is *chance,* and it has been suggested by others, for example, Beyth-Marom (1982). Together with modifiers one can get a list of expressions based on single stems. For example, *very high chance, high chance, better than even chance, even chance, less than an even chance, low chance,* and *very low chance* (not all of which we have studied) would offer such a list, and we could try to associate intervals with each expression. If we want the actual interval to relate to the uncertainty people express, then it may be difficult to get proper coverage of the interval 0 to 100% using a single method of assessing variability such as the interquartile range. Kadane's choices behave fairly well from this point of view. Using the intervals from our Table 2 (rounded) and comparing them with his intervals, we get Table A. With a little petting and patting the agreement for an ultimate list might be improved though we must recognize that the quartiles from Mosteller and Youtz (Table 2) have considerable uncertainty themselves. In *even chance* we have more precision than the proposed interval, perhaps *about an even chance* would lengthen the interval satisfactorily. At any rate, this comparison suggests that both the level and the uncertainty could be approximately communicated in a codification.

Again we thank the discussants for taking time to write and document their comments. These will help us in our further work.

TABLE A
Comparison of expression intervals

Expression	Kadane interval	Science writers' interval (Mosteller and Youtz)
Almost never	0.00–0.05	0.01–0.05
Seldom	0.05–0.15	0.07–0.18
Infrequent	0.15–0.25	0.10–0.23
Sometimes	0.25–0.35	0.18–0.35
Less than an even chance	0.35–0.45	0.40–0.45
Even chance	0.45–0.55	0.50^{-}–0.50^{+}
More often than not	0.55–0.65	0.57–0.60
Often	0.65–0.75	0.65–0.75
High probability	0.75–0.85	0.77–0.87
Very high probability	0.85–0.95	0.90–0.95
(Virtually) certain	0.95–1.00	0.99–1.00