Journal of Experimental Psychology: General
1979, Vol. 108, No. 4, 486–489

On Measuring Fuzziness: A Comment on "A Fuzzy Set Approach to Modifiers and Vagueness in Natural Language"

David C. Rubin
Duke University

SUMMARY

Hersh and Caramazza's application of fuzzy set theory to vagueness in natural language is criticized for including in their measures of fuzziness response variability due to experimental and statistical procedures.

Recently, Hersh and Caramazza (1976) applied fuzzy set theory to the meaning of phrases involving modifiers. By giving a yes or no response, subjects in their four experiments indicated whether a phrase was an appropriate description of a stimulus. In particular, phrases including *small, not small, very small,* and *not very large* were judged with respect to a set of 12 squares varying in size. The effects of the modifiers in the phrases could be understood in terms of fuzzy set theory both for group data as well as for most individuals.

No argument is made here with the hypothesis that natural language concepts are vague or with the claim that fuzzy set theory is a valuable way to view that vagueness (McCloskey & Glucksberg, 1978; Oden, 1977). What is questioned is the confounding of the measurement of fuzziness in natural language categories with response variability due to experimental and statistical procedures.

Consider the following thought experiment. A complete set of 12 squares in ascending size order is placed before you at one time, and you are to decide whether the word *small* applies to each square. You would probably say yes to the first n squares and no to the next $12 - n$ squares, producing a step function, not an ogive. In fact, you could show no evidence of conceptual fuzziness. Many other subjects would also be tested in this experiment and the results averaged. There would be some variation across people in the number of squares said to be small, but each individual subject would produce a step function. Hence, the average result would be an ogive indicating fuzziness, but this would be due to averaging and would not be typical of any individual. From such data, one could say that people have various ideas about concepts such as *small* and that this constitutes part of the fuzziness of their concepts. But this would only be a claim that words have different meanings to different people, not that words can best be described as having fuzzy meanings.

Repeating the same task several times with the same indivadual subject may be a way to solve this problem. As an indication that fuzziness exists, this procedure can at least show that individuals vary in their use of language over time (McCloskey & Glucksberg, 1978). Whether all such variation over time should be taken as evidence for fuzziness is a definitional problem.

Instead of seeing all of the squares at once in ascending order, consider a second thought experiment in which you will see them one at a time in random order. Each square will be

I wish to thank Gregory Lockhead for his comments. Support for this research was provided by a Biomedical Research Support Grant through Duke University.

Requests for reprints should be sent to David C. Rubin, Psychology Department, Duke University, Durham, North Carolina 27706.

presented for only 1 sec, with a 5-sec period between presentations of squares. Your judgment task will be the same: Does the word *small* apply to each square presented? You may do worse now, because the task is harder: You must keep your criterion in memory if you are to be as consistent as you were in the initial thought experiment. Since you probably could not be as consistent, you would show some response variability because of the difficulties introduced by the change in procedures. Thus, for example, if you were asked to judge which squares were smaller than a standard square shown once at the beginning of the judgment task, you might still have considerable response variability even though there is no fuzziness in the concept *smaller than a standard square*.

In Hersh and Caramazza's Experiment 2 (1976), individual subjects were studied using procedures similar to those of the second thought experiment. Thus, the demands of the task itself could have produced the observed response variability. In Hersh and Caramazza's Experiment 4, in which all squares were presented simultaneously, the experimental procedure was unlikely to produce response variability; however, averaging over subjects could have been responsible for the reported response variability. In Hersh and Caramazza's Experiments 1 and 3, both the procedure and the averaging could have contributed to the observed response variability. Thus, based on their data there is no need to postulate fuzziness in natural language to account for the response variability observed.

People can judge some things extremely well (Stevens, 1975) and others quite poorly (Nisbett & Wilson, 1977). If people are as good at judging their use of linguistic terms as psychologists and linguists usually assume, then there is a simple way out of this methodological dilemma. If you want to know what people do, just ask them.

Consider a third thought experiment. You are presented with all 12 squares in ascending order, just as in the first thought experiment, and are asked: "Relative to the 12 squares shown, what percentage of the time would you consider the term *small* to be appropriate for each square?" You respond in percentages, not in zeros and ones. You could give a step function, but more likely you would give a smooth ogive curve. This experiment is not perfect, since you could be highly influenced by the wording of the instructions, but at least the response measure used gives you the freedom to provide an indication of the fuzziness of your concepts.

The three thought experiments just described were actually carried out. In the first experiment, 17 undergraduate volunteers were asked: "Relative to the 12 squares shown, for which of the squares would you consider the term *small* to be appropriate? Indicate your judgments by writing a 'yes' or a 'no' on the lines beside the squares." The squares were all displayed in order on an 8½ by 11 in. sheet of paper. Although the squares were smaller in absolute size than those used by Hersh and Caramazza (1976), their relative sizes were the same. All subjects responded with step functions. The average curve was an ogive. In order to quantify these results, Hempel's (1939) formula for vagueness presented by Hersh and Caramazza was applied. The formula produces a minimum value of zero when all responses in a set are either ones or zeros and a maximum value of one when all responses in a set are 50%. The vagueness of all individual responses had to be zero, because individuals could only respond with zeros or ones. The vagueness of the grouped data was .15.

The next thought experiment was divided into two groups. The first group was a replication of part of Hersh and Caramazza's Experiment 1. Nineteen subjects were presented with written instructions informing them that after viewing 12 squares in increasing order of size, they would be asked to judge whether the term *small* was appropriate to each of 24 squares shown in a random order. Following Hersh and Caramazza, the squares were individually projected on a screen for 1 sec with 5 sec between squares for the subjects to record their judgments. The second group of 19 subjects had a procedure identical to that of the first group except that their task was to judge whether the squares shown were smaller than a

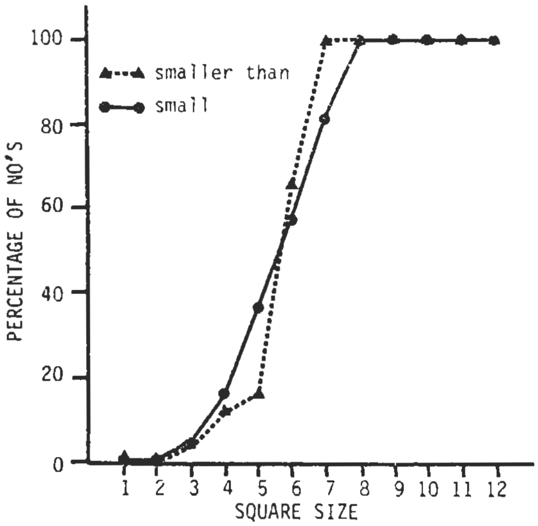

Figure 1. A comparison of judgments of small and smaller than a standard.

standard square: a 16-in. square that was the sixth smallest square in Hersh and Caramazza's series of 12 squares. The standard square was shown once, immediately before the judgment task began. Because there is no fuzziness in the concept *smaller than a 16-in. square*, all response variability observed in this *smaller than* group must be due to factors other than conceptual fuzziness, namely judgment, memory, and perceptual errors. The results are presented in Figure 1. Quantitatively, the vagueness for the *small* group was .27, whereas the vagueness for the *smaller than* group was .17. Thus, approximately 63% of the vagueness observed in judgments of *small* can be attributed to factors other than fuzziness in the concept *small*. Because the results of the *small* group are quite similar to Hersh and Caramazza's results (see their Figure 1), it can be assumed that about 63% of the response variability they present in their Experiments 1–3 is also due to the judgment task itself rather than to fuzziness of the concepts they studied.

In the final experiment, 19 subjects were presented with all 12 squares on a sheet of paper, as in the first thought experiment. They were asked to indicate the percentage of time they would consider the term *small* to be appropriate for each square rather than just whether the term *small* was appropriate or not. All subjects' responses showed a gradual increase: That is, there were no step functions. The vagueness calculated for individual subjects ranged from .21 to .83, with a mean of .51 ($SD = .18$). Thus, according to subjects' direct estimates, there is considerable fuzziness in the concept *small*—more than one would estimate from combining errors in a judgment task with individual differences.

The point of this reply is more than the methodological one of introducing a different measurement technique for fuzziness. Psychology cannot advance by ignoring its old knowledge when it embraces the new. The argument about individual versus group data for estimating fuzziness is a classic one in psychology; it is the same as the argument about all or none versus incremental learning (Millward, 1971). Averaging step functions yields an ogive curve. The argument about judgments comes from classical psychophysics. The task used by Hersh and Caramazza (1976) can be viewed as a modified method of comparison where the standard stimulus is internal or as a category classification task. From psychophysics, we understand a great deal about "errors" in this task, especially about how the range and magnitude of recently presented stimuli affect the classification of the current stimulus (Gravetter & Lockhead, 1973; Helson, 1964; Parducci, 1965). These "errors" are not random; they can be predicted with some accuracy. To lump these sources of response variability together and label them as fuzziness in a linguistic term masks what fuzziness might actually exist.

Response variability is not by itself evidence for fuzziness in linguistic terms. Response variability in the judgment of a term could occur for many reasons: variance within individuals due to fuzziness in the term itself, variance between individuals due to different linguistic usage, variance common to all experiments of the type performed whether they involve fuzzy concepts or not, random error variance, and so forth. These cannot be pooled into one large fuzziness term. In order to measure fuzziness, fuzziness must be operationally defined and the factors included in the definition must be isolated.

References

Gravetter, F., & Lockhead, G. R. Critical range as a frame of reference for stimulus judgment. *Psychological Review*, 1973, *80*, 203–216.

Helson, H. *Adaptation level theory*. New York: Harper & Row, 1964.

Hempel, C. G. Vagueness and logic. *Philosophy of Science*, 1939, *6*, 163–180.

Hersh, H. M., & Caramazza, A. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, 1976, *105*, 254–276.

McCloskey, M. E., & Glucksberg, S. Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 1978, *6*, 462–472.

Millward, Richard B. Theoretical and experimental approaches to human learning. In J. W. Kling & L. A. Riggs (Eds.), *Experimental psychology*. New York: Holt, Rinehart & Winston, 1971.

Nisbett, R. E., & Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 1977, *84*, 231–259.

Oden, G. C. Integration of fuzzy logical information. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 565–575.

Parducci, A. Category judgment: A range-frequency model. *Psychological Review*, 1965, *72*, 407–418.

Stevens, S. S. *Psychophysics*. New York: Wiley, 1975.

Received January 12, 1979 ■