Int. J. Man–Machine Studies (1988) **29,** 149–157

The evaluation of verbal models

RAMI ZWICK

Department of Marketing, 707 Business Administration Building, The Pennsylvania State University, University Park, PA 16802–9976, USA

(Received 24 July 1987)

This paper proposes an operational method for evaluating verbal models. The method is based on a statistical technique in which the performance of the verbal model is compared to the performance of an alternative simple random choice model. The method is demonstrated by using experimental data to evaluate Yager's model (1978; 1984) of fuzzy probabilities.

Introduction

A verbal model is a model that allows for linguistic rather than numerical variables, and for causal relationships between the variables to be formulated verbally rather than mathematically (Wenstøp, 1976). Examples for such include Kickert's model (1979) in which Mulder's theory of power is reformalized verbally, Wenstøp's (1976) deductive verbal models of organizations, and Yager's (1979; 1984), Kwakernaak's (1978; 1979), and Zadeh's (1968; 1981) models of fuzzy probabilities (Zwick & Wallsten, 1989).

A fundamental issue in the theory of modelling is validation of a model. In the conventional modelling exercise this rarely consists of more than the calculation of some accuracy measure between the model and the data. However, with verbal models this approach encounters several difficulties. Several authors (Wenstøp, 1975; Yager, 1978; Tong, 1980) have discussed methods for comparing linguistic values—empirical and predicted—for validation purposes. However, these techniques suffer from several problems that will be discussed next.

Tong's (1980) evaluation of fuzzy (verbal) models derived from experimental data

Tong evaluates a verbal model in terms of its complexity, accuracy and uncertainty. Complexity is measured by the number of linguistic relations that make up the model. Accuracy is measured by a distance function (squared error, or an absolute difference) between the *de-fuzzified* output of the verbal model and the non-fuzzy measured data. Finally, uncertainty is measured by either the non-probabilistic entropy of the membership function of the model's output set (A), or by a function of the cardinality of A. As was emphasized by Tong (1980), these indices are appropriate in some contexts, but they are not necessarily a general solution. Especially problematic is the de-fuzzification, which may be achieved in several ways, with no published evidence to suggest that any method is superior. Secondly, it may not be a trivial task to impose an ordering on the models on the basis of such

0020-7373/88/070149 + 09$03.00/0

measures. Finally, his technique lacks any statistical foundation. It is not clear, using these indices, whether one model is significantly better than another, or whether the observations are reasonable under the hypothesis that the model is valid.

Yager's (1978) linguistic models and fuzzy truths

Yager (1978) introduced a technique for validating fuzzy set models that is based upon the concept of compatibility between two fuzzy subsets of the same universe. The application of his methodology results in a linguistic truth value (i.e. *true, almost true, more or less true, false,* etc.), which measures the validity of the model for a piece of data.

Assume that we have a proposition (statement or equation) $p \triangleq X$ is F (where F is a fuzzy subset of U). The truth value of this proposition is defined to be the degree of consistency of p with some observed data that is expressed as a reference proposition r (Bellman & Zadeh, 1976). Thus:

$$V(p) = C(p, r)$$

where $V(p)$ is the truth value of p with respect to r, and C is a consistency function which maps ordered pairs of propositions and reference statements into truth values. Assume now that r, our reference proposition, is itself a verbal proposition of the form:

$$r \triangleq X \text{ is } G$$

where G is a fuzzy subset of U. In this case the truth value of p becomes a linguistic truth value defined as:

$$£ = F(G)$$

where $F(G)$ is a fuzzy subset over the unit interval defined (Zadeh, 1977) as:

$$F(G) = \left[\frac{G(u)}{F(u)}\right] \text{ for all } u \varepsilon U.$$

For example, let $p \triangleq$ John is *old*, and let $r \triangleq$ John is *close to 70*. Then if we define *old* as:

$$F = old = \left\{\frac{0{\cdot}4}{30}\ \frac{0{\cdot}5}{40}\ \frac{0{\cdot}6}{50}\ \frac{0{\cdot}7}{60}\ \frac{0{\cdot}8}{70}\ \frac{0{\cdot}9}{80}\ \frac{1}{90}\ \frac{1}{100}\right\},$$

and *close to 70* as:

$$G = close\ to\ 70 = \left\{\frac{0}{30}\ \frac{0{\cdot}2}{40}\ \frac{0{\cdot}4}{50}\ \frac{0{\cdot}8}{60}\ \frac{1}{70}\ \frac{0{\cdot}8}{80}\ \frac{0{\cdot}4}{90}\ \frac{0{\cdot}2}{100}\right\},$$

then the truth of p with respect to r is given by

$$£ = \left\{\frac{0}{0{\cdot}4}\ \frac{0{\cdot}2}{0{\cdot}5}\ \frac{0{\cdot}4}{0{\cdot}6}\ \frac{0{\cdot}8}{0{\cdot}7}\ \frac{1}{0{\cdot}8}\ \frac{0{\cdot}8}{0{\cdot}9}\ \frac{0{\cdot}4}{1}\right\}.$$

£ could be expressed using linguistic approximation as "near 0.8".

This technique suffers from several problems. It is *ad hoc,* lacks a firm foundation, and lacks a statistical theory to back it up. The compatibility measure itself that is being used in this technique is problematic. Usually $F(G) \neq G(F)$, which is an undesirable property in the models testing context. For example, why should the truth value of a model that predicts that "John is old" given that in reality John is "close to 70" be different from the truth value of a model that predicts that "John is close to 70" given that in reality "John is old". A second and more severe problem is the fact that $F(F)$ cannot be interpreted, in most cases, as "absolutely true" as should be expected if the reference proposition is exactly what the model predicts.

In what follows I will present a model testing technique that is based on both statistical and logical grounds, and that is especially suited to verbal models.

Model testing technique

The problem of evaluating models by testing their empirical consequences has both statistical and logical aspects. The statistical problems are those of determining how well the model fits the data. However, in the social sciences "scientific laws" do not assert deterministic invariance, and seldom will a model demonstrate a perfect fit to the data. Consequently the emphasis shifts from "factually true" to "usefulness" and "efficiency" (Zimmerman, 1985). To determine the usefulness and efficiency of a model we need to compare it to an alternative simpler model, and to show that the addition in complexity is compensated by significant improvement in predictive power. In the numeric-response models, established techniques such as linear regression or correlation can be used. However, these techniques are not easily transferrable to linguistic models. The current technique is an adaptation of the classical techniques to the realm of verbal models.

Let VM be a verbal model and let O be the set of all possible *verbal* states of nature. A verbal state of nature is a fuzzy subset of an appropriate universe of discourse. For example, let VM be Yager's probability model (the probability of a fuzzy event given a crisp random crisp variable). In this case O is the vocabulary set of a specific individual the model is trying to simulate. In this work, O will be considered to be finite and known. The verbal model predicts one of the n possible verbal states of nature in O. The proposed technique compares the model to an alternative simple *baseline m*odel, which uses a uniform random process to predict a verbal state of nature in O.

An important issue with regard to the baseline model concerns the unit of analysis to which the equal likelihood assumption should be applied. Namely, should each phrase (in O) be considered equally likely, or should the uniformity assumption be applied to equivalent classes (synonyms)? I adopt the second approach. Thus, the unrestricted baseline model randomly selects an equivalence class, rather than a phrase, as the predicted response. As a result, the phrases themselves are not equally likely, but each equivalence class of synonyms is.

The following procedure can be used to determine equivalence classes. Using the membership function representation of each verbal state of nature in O, compute the pairwise distances between all words using the best similarity index for the specific context at hand (Zwick, Carlstein & Budescu, 1987). Then use the distance matrix to cluster the verbal states of nature (using any one of several cluster analysis techniques). Unfortunately, there are no satisfactory methods for determining the number of population clusters for any type of cluster analysis (Everitt, 1979). Consequently, I recommend that the analysis be repeated with different levels of clustering and that the effect on the model testing results be observed.

To determine the quality of the verbal model, observe its behaviour along several trials (say m) and compare the predicted outcomes with the observed ones. This can be accomplished by computing the distances between the predicted and the observed response clusters using the appropriate similarity index, where distance is defined to be the average distance between the observed response and the members of the predicted cluster. In addition, for each trial, find the discrete sampling distribution of the distances between a randomly selected cluster and the observed response, which is simply the set of all possible distances from the observed response to the set of all clusters in O (call this random variable X_i). The unrestricted baseline model assumes that responses are independent. Hence, under the unrestricted baseline model assumption we have a sequence of m independent random variables, $\{X_i\}_{i=1}^{m}$, the mean and standard deviation of which are known. According to the Liapunov version of the central limit theorem (Rao, 1975, p. 107):

$$\gamma_m = \frac{\sum_{i=1}^{m} (X_i - \mu_i)}{C_m}$$

tends to the standard normal. Where

$$E(X_i) = \mu_i, \quad \text{and} \quad C_m = \left(\sum_{i=1}^{m} \sigma_i^2\right)^{1/2}.$$

Based on this approximate sampling distribution we can compute the probability of the standardized observed mean distance (or a smaller value) under the unrestricted baseline model assumptions. A small probability value indicates that the tested model is out-performing the unrestricted baseline model, while a sizeable probability value indicates that the tested model does not improve prediction beyond the performance of a random unrestricted baseline model. Any verbal model should pass the initial test to deserve further consideration.

To investigate further the predictive power of the tested models, the number of clusters from which the baseline model is allowed to randomly choose a predicted response is successively restricted. The restriction mechanism is analogous to an *a priori* piece of information regarding the rough location of the observed data. Clearly the predictive power of the unrestricted baseline model depends on the number of clusters in O. Few clusters increase the random baseline hit rate, making it harder for the tested model to distinguish itself. I recommend that the analysis be carried out sequentially, eliminating at each stage the cluster that is the farthest away from the observed one in the previous stage. At each stage, the approximate sampling distribution of the standardized mean distance under the restricted baseline model can be computed and the probability of the observed mean distance can be determined. The analysis can be reported by the number of clusters that are eliminated before the model tested ceases to out-perform the restricted baseline model ($p > 0{\cdot}05$).

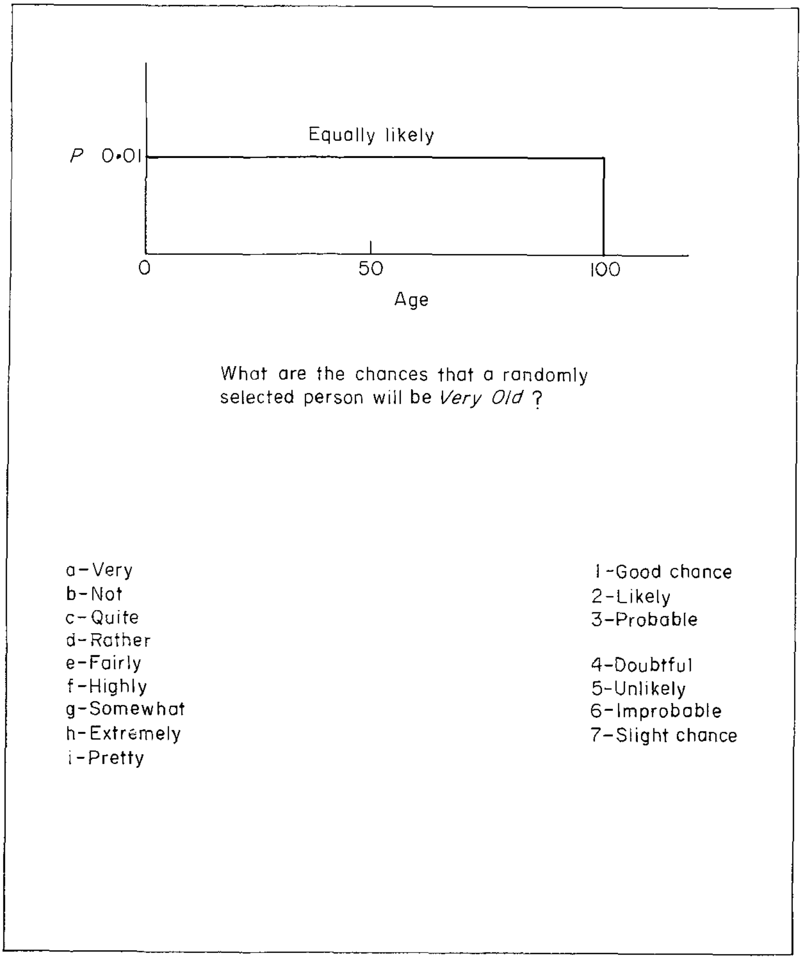
EXAMPLE

This example uses the data of a single subject to validate Yager's model (1979, 1984) of fuzzy probability. Full details, as well as the application to a much larger body of data, can be found in Zwick & Wallsten (1989).

Yager's model (1984)

Consider the data in the top part of Fig. 1. What are the chances that a randomly selected person from this population will be *very old*?

Yager (1979) noted that intuitively the probability of a fuzzy event (A) should be



What are the chances that a randomly selected person will be *Very Old* ?

a-Very
b-Not
c-Quite
d-Rather
e-Fairly
f-Highly
g-Somewhat
h-Extremely
i-Pretty

1-Good chance
2-Likely
3-Probable

4-Doubtful
5-Unlikely
6-Improbable
7-Slight chance

FIG. 1. Probability estimation trial (from Zwick & Wallsten, 1989).

a fuzzy subset itself. Hence, he assumed that for each possible numeric response, subjects evaluate the truth values of the propositions: "the probability of A is at least p", and "the probability of A is at most p", and combine these evaluations through the min rule. The evaluations are accomplished by imagining or mentally simulating different possible crisp events (A_α's) associated with the linguistic term defining event A. Formally, let A be a fuzzy subset of Ω, and let P be a crisp probability measure defined on Ω, then using the extension principle, and following Zadeh's (1981) work on fuzzy cardinality, Yager defined the following three probabilities:

$$(1)\ \mu_{FGP(A)}(p) = \sup\{\alpha | P(A_\alpha) \geq p\}, \qquad p \varepsilon [0, 1]$$

$\mu_{FGP(A)}(p)$ should be interpreted as the truth value of the proposition: "the probability of A is at least p".

$$(2)\ \mu_{FLP(A)}(p) = 1 - \mu_{FGP(\bar{A})}(p) = \sup_{\alpha}\{\alpha | P(\bar{A}_\alpha) \geq 1 - p\}$$

$\mu_{FGP(\bar{A})}(p)$ should be interpreted as the truth value of the proposition: "the probability of not-A is at least p", and $\mu_{FLP(A)}(p)$ as the truth value of the proposition: "the probability of A is at most p".

$$(3)\ FEP(A) = FGP(A) \cap FLP(A)$$

where $\mu_{FEP(A)}(p)$ is the truth value of the proposition: "the probability of A is p".

Model validation

Twenty subjects were tested for five sessions of approximately 1 to 1·5 hours each. Sessions 2 and 3 were one integral unit broken into two parts due to the lengthy nature of this unit. Similarly, Sessions 4 and 5 were an integral unit. In what follows, Sessions 2 and 3 will be referred to as Part 1, and Sessions 4 and 5 as Part 2 of the experiment. Part 2 was a replication of Part 1. Session 1 was for practice, and Parts 1 and 2 were for data collection. The practice and four data sessions were scheduled generally two days apart. The experiment was controlled by an IBM-PC with stimuli presented on a colour monitor and responses made on the keyboard.

During all sessions, subjects worked through five types of tasks comprising: (1) a probability estimation task (PE); (2) a linguistic probabilities scaling task (LPS); (3) a linguistic ages scaling task (LAS); (4) a linguistic probabilities similarity judgment task (SIM); and (5) a probability estimation–scaling task. Trials from all tasks were presented in all sessions in an intermixed random order.

The probability estimation task was the core task of the experiment in which subjects were instructed to estimate the probabilities of certain events given the information presented on the screen. Figure 1 is an example for such a trial that is relevant to Yager's model. Subjects were instructed to respond by choosing one of seven primary terms, or by combining one primary term with one or two of nine modifiers. The objective of the scaling tasks (tasks 2 and 3) was to establish the subject's membership function for various linguistic probabilities and age phrases. A

TABLE 1

Subject 8's vocabulary at the 99·9% clustering level

Cluster 1
- Likely

Cluster 2
- Probable
- Quite probable
- Rather good chance

Cluster 3
- Good chance
- Fairly good chance
- Pretty good chance
- Improbable
- Rather improbable
- Somewhat improbable

Cluster 4
- Fairly improbable
- Quite improbable
- Fairly slight chance

Cluster 5
- Slight chance
- Pretty slight chance
- Rather slight chance
- Somewhat slight chance

Cluster 6
- Doubtful
- Fairly doubtful
- Pretty doubtful
- Quite doubtful
- Quite quite doubtful
- Rather doubtful
- Somewhat doubtful
- Very doubtful
- Extremely improbable
- Very slight chance

Cluster 7
- Unlikely
- Extremely unlikely
- Fairly unlikely
- Highly unlikely
- Quite unlikely
- Rather unlikely
- Somewhat unlikely

TABLE 2

Subject 8's observed versus predicted responses in Task PE by part and problem

	Observed		Predicted	Distance between observed and predicted	
Problem	Part 1	Part 2	Cluster	Part 1	Part 2
1	Somewhat slight chance (5)	Pretty slight chance (5)	5	0·001	0·001
2	Fairly improbable (4)	Rather improbable (3)	3	0·093	0·000
3	Extremely unlikely (7)	Extremely unlikely (7)	7	0·000	0·000
4	Somewhat slight chance (5)	Somewhat slight chance (5)	4	0·085	0·085
5	Slight chance (5)	Extremely improbable (6)	2	0·707	1·000
6	Rather doubtful (6)	Somewhat doubtful (6)	6	0·020	0·023
7	Somewhat improbable (3)	Fairly doubtful (6)	6	0·513	0·016
8	Fairly doubtful (6)	Rather doubtful (6)	5	0·189	0·189
9	Fairly doubtful (6)	Fairly doubtful (6)	6	0·016	0·016
10	Fairly doubtful (6)	Rather doubtful (6)	6	0·016	0·020
11	Quite improbable (4)	Fairly improbable (4)	2	0·475	0·496
12	Somewhat improbable (4)	Fairly improbable (4)	4	0·088	0·000
14	Improbable (3)	Somewhat slight chance (5)	4	0·065	0·088
15	Somewhat slight chance (5)	Quite improbable (4)	3	0·346	0·083
16	Somewhat doubtful (6)	Fairly doubtful (6)	6	0·023	0·023

Numbers in parentheses show the clusters to which the phrases belong (see Table 1).

direct magnitude estimation technique was used. The objective of the similarity task was to choose the best similarity index between membership functions within a subject (see Zwick *et al.*, 1987). See Zwick & Wallsten (1989) for a full description of all tasks and all data analysis.

Table 1 presents the clustering structure of Subject 8's vocabulary, and Table 2 presents Subject 8's predicted versus observed responses (by parts) in the probability estimation trials that were relevant to testing Yager's model. (There were 15 different problems presented once in each part.) In parenthesis beside the observed response is the cluster to which the response belongs (see Table 1). Column 4 presents the model's predicted cluster. The right-most two columns of Table 2 presents the distances between the predicted and the observed response clusters. For Subject 8 the mean distance (across parts) between observed and predicted responses was found to be 0·156 (see Table 2). Under the unrestricted baseline model's assumptions the probability of finding such a distance (or a smaller one) is extremely low ($P < 0{\cdot}0001$). This indicates that Yager's model predicts Subject 8's responses much better than the unrestricted baseline model.

To further investigate the predictive power of the tested models, the number of clusters from which the baseline model was allowed to randomly choose a predicted response was successively restricted.

Table 2 reveals that in 15 cases Yager's model correctly predicts the response cluster. In nine cases Yager's model is off by only one level of probability expression (namely predicting a response cluster that is adjacent to the observed one), in three cases by two levels, in two cases by three levels and in one case by four levels. This indicates that, for Subject 8, Yager's model correctly captures the general location of the response if not the exact expression. I sequentially carried out the analysis demonstrated in Table 2, eliminating at each stage the cluster that is the farthest away from the observed one in the previous stage. For example, with regard to Subject 8's data (see Table 1), in the first stage cluster 1 (**likely**) was eliminated from problems for which the observed response cluster was 4, 5, 6, or 7. Cluster 7 was eliminated from problems for which the observed response cluster was 1, 2, or 3. In stage two, the next most distant cluster from the observed response cluster was eliminated and so on. In all cases, the predicted response cluster (by the tested model) was defined to be the cluster in the restricted vocabulary that is the closest to the predicted function. At each stage, the approximate sampling distribution of the standardized mean distance under the restricted baseline model was computed and the probability of the observed mean distance was determined. Even after eliminating all but three clusters (the observed response cluster plus two others) from Subject 8's vocabulary, Yager's model still out-performs the restricted baseline model ($P = 0{\cdot}03$). Only after eliminating all but two clusters does Yager's model cease to out-perform the restricted baseline model ($P = 0{\cdot}65$). Note that this is strong support for the ability of Yager's model to predict Subject 8's responses.

Conclusion

The purpose of this paper is to introduce a technique for validating verbal models. This technique can be used only in the case where the set of all possible outputs of the model is finite and known. The advantage of this technique is its statistical properties.

References

BELLMAN, R. E. & ZADEH, L. A. (1976) *Local and fuzzy logics*. ERL-M584. Electronics Research Laboratory, College of Engineering, University of California, Berkeley, California.

EVERITT, B. S. (1979). Unresolved problems in cluster analysis. *Biometrics,* **35,** 169–181.

KICKERT, W. J. M. (1979). An example of linguistic modeling, the case of Mulder's theory of power. In GUPTA, M. M., RAGADE, R. K. & YAGER, R. R. Eds, *Advances in Fuzzy Set Theory and Application.* Amsterdam: North Holland.

KWAKERNAAK, H. (1978). Fuzzy random variables, I: Definitions. *Information Sciences,* **15,** 1–29.

KWAKERNAAK, H. (1979). Fuzzy random variables, II: Algorithms and example for the discrete case. *Information Sciences,* **17,** 253–278.

RAO, C. R. (1965). *Linear Statistical Inference and Its Applications.* New York: John Wiley & Sons.

TONG, R. M. (1980). The evaluation of fuzzy models derived from experimental data. *Fuzzy Sets and Systems,* **4,** 1–12.

WENSTØP, F. (1976). Deductive verbal model or organization. *International Journal of Man–Machine Studies,* **8,** 293–311.

WENSTØP, F. (1980). Quantitative analysis with linguistic values. *Fuzzy Sets and Systems,* **4,** 99–115.

YAGER, R. R. (1978). Linguistic models and fuzzy truths. *International Journal of Man–Machine Studies,* **10,** 483–494.

YAGER, R. R. (1979). A note on probabilities of fuzzy events. *Information Sciences,* **18,** 113–129.

YAGER, R. R. (1984). A representation of the probability of a fuzzy subset. *Fuzzy Sets and Systems,* **13,** 273–284.

ZADEH, L. A. (1968). Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications,* **23,** 421–427.

ZADEH, L. A. (1977) *Theory of Fuzzy Sets.* Memo no. UCA/ERL M77-1. Electronics Research Laboratory, College of Engineering, University of California, Berkeley, California.

ZADEH, L. A. (1981). Fuzzy probabilities and their role in decision analysis. *Proceedings of the Fourth MIT/ONR Workshop on Command, Control and Communication,* pp. 159–179. Cambridge, MA: MIT.

ZIMMERMANN, H.-J. (1985). *Fuzzy Set Theory and Its Applications.* Boston: Kluwer-Nijhoff.

ZWICK, R., CARLSTEIN, E. & BUDESCU, D. V. (1987). Measures of similarity among fuzzy concepts: a comparative analysis. *International Journal of Approximate Reasoning,* **1,** 221–272.

ZWICK, R. & WALLSTEN, T. S. (1989). Combining stochastic uncertainty and linguistic inexactness: theory and experimental evaluation of four fuzzy probability models. *International Journal of Man–Machine Studies.* In press.