Proceedings of the 2011 ASME IDETC
International Design Engineering Technical Conferences & Information in Engineering Conference
August 28 – 31, 2011, Washington, D.C.

DETC2011-47074

TIME-BASED MODELING OF LINGUISTIC PREFERENCE TO PREFERENTIAL PROBABILITY

Andy Dong
Design Lab
University of Sydney
Sydney, NSW, Australia

Tomonori Honda
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA

Maria C. Yang
Department of Mechanical Engineering
and Engineering Systems Division
Massachusetts Institute of Technology
Cambridge, MA

ABSTRACT

In this paper, we present a method to estimate the likely concept a committee of designers will select given their verbalized preferences toward each alternative. In order to perform this estimation, we present a new method of preference elicitation based on natural language. First, we show a way to model preference in the natural language of appraisal, which describes the degree of intensity and the uncertainty of preference based upon gradable semantic resources to express appraisals. We then show a way to map linguistic appraisals into probability distribution functions. Finally, we present a Markov model that utilizes these probability distribution functions in state transition matrices to calculate in a time-varying manner the change of preference over time. We present a case study to illustrate the validity of the method.

INTRODUCTION

Committee-based decision-making for the purpose of concept selection is a prototypical decision-making context in engineering design [1]. For non-routine, creative design problems, engineers typically produce multiple alternative concepts. In order to proceed, the alternative that best satisfies the criteria must be chosen. This concept selection entails the analysis and evaluation of alternative concepts, leading to the selection or consolidation of one or more concepts for further development. This choice among mutually exclusive alternatives is a classic decision problem. This decision-making context has a particularly crucial role in engineering management because it most accurately reflects the commercial realities of the innovation-oriented decision-making processes of internal innovation management (or R&D investment) committees [2].

Engineering design research has provided an array of methods to assist designers in making this choice and for the aggregation of preferences when this decision is taken in a committee [e.g., 3, 4-7]. Whatever method is adopted or promoted, the fundamental problem is two-fold: eliciting subjective preferences from decision-makers and aggregating preferences. A direct elicitation of preferences is awkward in situations where preferences dynamically change as a result of discussion and negotiation within a committee or interaction with the alternatives, which could update preferences. Further, there is evidence suggesting that engineers may not know correctly how to apply methods requiring preference data due to the challenges of defining the utility of attributes early in the design process [8]. This lack of knowledge results in concept selection methods being misused. The approach we advocate is to apply natural language processing to determine a preference order among a set of alternatives. In design, as with many fields, decisions are not always formally modeled but only spoken or written about, and there is a need for decision models based on discussion and negotiation to provide guidance for decision makers, thereby increasing the accountability and transparency of decisions.

In our prior research [9], we showed that it was possible to predict the concept selection decision of the committee based on a qualitative analysis of language expressing preferences and a maximum likelihood estimation of preference based on the linguistic data. We build upon that work to produce a formal model describing the decision-making based upon what the committee has stated about preferences toward an alternative. The challenge is to model preference and uncertainty as realized in natural language and then convert them into a formal model describing the decision-making (choice) that is taking place. In this study, we restrict the space of decision-making to the selection of an alternative from a set of choices by a committee, such as a small design team. While even the scope of this area of research is itself the basis of an entire set of empirical economics literature [10], our interest is in modeling the decision that is likely to take place given what the committee has said.

More formally, we would like to estimate the probability that an alternative is the most preferred alternative given the way that the committee has expressed its 'preference' toward an alternative using the natural language of appraisal [11]. Linguistic data is a suitable data set for preference giving and preference elicitation in engineering design [9] and is increasingly used to mine customer opinions toward products [12]. Intuitively, when presented with a set of discrete alternatives, if a person says, "Alternative 1 is a really good idea," then it is reasonable to increase the probability that alternative 1 is the most preferred one while decreasing the probabilities that the other alternatives are the most preferred. Thus, the linguistic data could provide time-varying information about the probability that the preference for alternative 1 will change in line with the degree and direction of a positive or negative orientation appraisal of alternative 1. The linguistic data reflects an update in preference as a person interacts with the alternatives or with others in the committee.

We will show that it is possible to produce a time-varying model of preference toward alternatives starting from normative ways in which appraisals may be stated. We assume that the linguistic data describes individual decision-maker's subjective preferences and that an overall trajectory of strongly positive and certain appraisals of a particular alternative reflects the committee's aggregation of preferences toward the committee's most preferred alternative. There is no assumption that the committee is necessarily behaving rationally or that the outcome of the committee's decision is the correct one. We will address these assumptions later in the paper in the discussion on how this method would fit into an array of methodologies for design decision-making.

METHOD

Decision-making by teams is a necessity in engineering design [1]. Suppose there is the situation wherein an engineer or group of engineers is discussing their preferences for different design alternatives. As they appraise (make verbal assessments of) the alternatives, their preference toward or away from an alternative may change in line with their discussion, and this change in preference should be reflected in their linguistic appraisals. Therefore, we wish to model the probability that a specific alternative is currently the most preferred alternative. We will use the following notation.

N: total number of design alternatives
D: the vector of all design alternatives, $D = \{d_1, d_2, \dots, d_N\}$
T: total number of time intervals
i: time interval, i=1 to T
π_i = most preferred design alternative at time interval i
d_j: j-th design alternative in the design selection problem, j=1 to N
$P(\pi_i = d_j)$: the probability that d_j is the most preferred alternative in time interval i (preferential probability)
e_i: linguistic data in time interval i

Let's say that the team must select one choice from 3 alternatives, and assume that the alternatives are mutually exclusive. In the absence of any prior knowledge, at the start, i=0, the probability that any of the alternatives is the most preferred alternative can be considered equivalent. That is, without any loss of generality, that initial probability that any alternative is the most preferred one is 1/3, or more generally 1/N. We call this probability the *preferential probability*, the probability that a choice is the most preferred choice over all others at any given moment in time. Now, let's suppose that a person says, "I really like the first one." This means that between time interval i=0 and i=1, we should be able to calculate a transition probability to reflect linguistic evidence that tells us that the preferential probability will transition from $P(\pi_0 = d_1)$ with probability p_{11}, from $P(\pi_0 = d_2)$ with probability p_{21} and from $P(\pi_0 = d_3)$ with probability p_{31} such that in the next state alternative d_1 will be the preferred alternative with probability $P(\pi_1 = d_1)$. The value of $P(\pi_1 = d_1)$ is the sum of the product of the prior probabilities (that d_1 was the most preferred alternative) and the respective state transition probabilities. Further, given the absence of any other linguistic data, we could say that the transition probabilities p_{12} and p_{13} will be less than p_{11}. To perform this modeling, we need to model preference giving as a time-varying activity where the probability for the preference toward (or away from) an alternative depends upon what is being said. We perform this modeling using a Markov chain and a formal model for the semantic resources in expressing a preference as an appraisal of an alternative.

Markov Model

The first step is to model the probability that a design alternative d_j is the most preferred one in time interval i as new linguistic data is provided, that is, as the discussion takes place. We can model this using a Markov chain. The Markov chain states that the probability that in the current state the most preferred alternative d_j is based on the previous state only. Further, we know that $\sum P(\pi_i = d_j) = 1$ (preferential probabilities at each interval sum up to 1) since each alternative is discretely defined and we assume that they are conditionally independent of each other. In a dialogue, we consider that at each interval, there is a probability that an alternative is currently the preferred alternative. An interval is defined by a linguistic appraisal. The state transition matrix gives us the

probability that when alternative d_1 is the most preferred alternative, it is followed by d_1 with probability p_{11} in the next interval, by d_2 with probability p_{12} and by d_3 by p_{13}. Likewise, when alternative d_2 is the most preferred alternative, it is followed by d_1 with probability p_{21} in the next interval, by d_2 with probability p_{22} and by d_3 with probability p_{23}. Depending upon the number of alternatives, we have an N-dimensional state transition matrix. The constraint on the state transition matrix in this formulation is that the transition probabilities from one alternative (*j*) to the all other alternatives must sum up to 1 at each interval *i*. That is, $\sum_{u=j,v=1}^{N} p_{uv}^{i} = 1$ (each row in the state transition matrix must add up to 1 at each interval *i*).

We now require an equation that tells us the value of the new probability, that is, the probability that a design alternative d_j is the most preferred one in time interval *i* based on the transition probability. We can calculate this probability value in a recursive way. The probability that alternative d_j is the most preferred one is the sum over each alternative's preferential probability in time interval *i*-1 multiplied by its state transition probability. In other words, the probability at interval *i* that the most preferred alternative is d_j is a function of the probability that any of the design alternatives in time interval *i*-1 was the most preferred alternative and the state transition probability. By doing so, we recursively take into account the trajectory of preference for d_j, from *i*=0 to *i*-1, that will affect its preference in time interval *i*.

For example, suppose a designer must select among 3 alternatives. Then suppose that the designer is given three utterances to state the preferences, without using ordinal values. Eq. 1 expresses the probability that the most preferred option is alternative 1 at interval 3:

$$P(\pi_3 = d_1) = P(\pi_2 = d_1)\cdot p_{11} + P(\pi_2 = d_2)\cdot p_{21} + P(\pi_2 = d_3)\cdot p_{31} \quad (1)$$

The Markov chain and the calculation of $P(\pi_i = d_j)$ is illustrated in Figure 1 for the 3 alternative problem between interval 2 and 3 with the state transition probabilities labeled. In general, the preferential probability at interval *i* is given by Eq. 2:

$$P(\pi_i = d_j) = \sum_{u=1}^{N} P(\pi_{i-1} = d_u)\cdot p_{uj} \quad (2)$$

Using this equation, it is possible to calculate the time-varying preferential probabilities in a recursive manner. Suppose that at interval *i*=2, the preferential probabilities are $P(\pi_2 = d_1) = 0.6$, $P(\pi_2 = d_2) = 0.3$ and $P(\pi_2 = d_3) = 0.1$. Let us now suppose that we obtain linguistic data. (We will discuss how we obtain this matrix in the next section.)

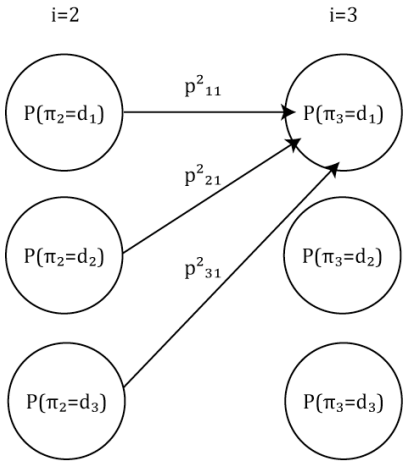

Figure 1 State transition between intervals 2 and 3

The state transition matrix is given by:

$$\begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \end{bmatrix}$$

We can calculate the new preferential probabilities by the matrix multiplication:

$$\begin{bmatrix} 0.6 & 0.3 & 0.1 \end{bmatrix} \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.25 & 0.25 \end{bmatrix}$$

What is needed is a way to estimate the state transition probabilities based on linguistic data. In the next section, we describe a method to do this.

Language Model

Using the formal modeling of the language of appraisal in design [13, 14], we can analyze natural language to identify the way that people express the degree of intensity and the uncertainty of a preference. The analysis will provide us a way to estimate the state transition probability for a given appraisal. We posit that the transition probability for appraisal clauses of varying degree and orientation would order the transition probabilities from highest to lowest. Qualitatively, we are suggesting that the transition probability is higher for the appraisal clause "This is a really good concept" than the appraisal clause "This is a so-so concept".

First, we identify the semantic resources that can express an appraisal. Semantic resources are ways of expressing meaning through language. In functional linguistics, there are five semantic resources for appraisal [15]: Attitude; Engagement; Graduation; Polarity; and Orientation. The resources of Attitude, Engagement and Graduation are gradable resources for evaluating alternatives.

We can group the semantic resources into appraisal groups, "groups and phrases in a text giving what kind and intensity of appraisal is expressed" [16]. Whitelaw uses a strict grammatical definition wherein an appraisal group "comprises of a head adjective with defined attitude type, with an optional preceding list of appraisal modifiers, each denoting a transformation of one or more appraisal attributes of the head". In our

formulation, an appraisal group is the set of semantic resources applied in the realization of an appraisal. Each resource could have gradable values of low, medium, and high, as shown in Figure 2.

Given the three gradable values for the semantic resources of Engagement and Graduation, with the possibility that these semantic resources are not always in use, and 3 gradable values for the semantic resource of Attitude, which must always be used in an appraisal, there are in total 48 (=4 × 4 × 3) canonical ways to express an appraisal, ranging from "This is good" to "I sort of think that this is sort of good" to "I really think that this is the very best." Each of these statements has a different level of intensity of judgment. This concept of intensity of judgment is similar to Subasic's concept of intensity [17], but we do not attempt to assign a numerical intensity to each semantic resource. Rather, we map the use of a semantic resource of appraisal and its gradable value into a measure of the 'intensity' of the entire appraisal. Based on this mapping, it becomes possible to estimate the state transition probability for any arbitrary appraisal since an appraisal can be broken down into its constituent semantic resources and the gradable values per resource.

$$\begin{pmatrix} Attitude & low \\ Engagement & - \\ Graduation & - \end{pmatrix} \rightarrow \begin{pmatrix} Attitude & low \\ Engagement & - \\ Graduation & high \end{pmatrix}$$

"This is good." "This is very good."

$$\rightarrow \begin{pmatrix} Attitude & high \\ Engagement & - \\ Graduation & high \end{pmatrix} \rightarrow \begin{pmatrix} Attitude & low \\ Engagement & low \\ Graduation & high \end{pmatrix}$$

"This is the best ever." "I sort of think this is very good."

Figure 2 Use of semantic resources in various appraisals and their gradable values

We obtained data on the intensity of canonical ways to express an appraisal through crowd sourcing using the online service Amazon Mechanical Turk. Mechanical Turk is increasingly being used for social science research [18] and for labelling data for machine learning [19]. Excluding all instances of appraisals with no appearance of the semantic resources of Engagement and Graduation, there are 27 canonical statements of appraisal that could be rated. (Annex A). We asked respondents or "workers" in Mechanical Turk parlance, to rate a total of 13 canonical appraisals from 1 (very weak) to 5 (very strong) with the midpoint 3 being neither weak nor strong. The workers were not providing a preference for given set of alternatives; they were asked to consider the appraisals in the abstract, as if they were judging a movie they had recently seen. In the instructions, they were asked to consider if a statement such as "I really think that this is much better." reflects a stronger or weaker appraisal of a movie than "I sort of think that this is sort of good." We randomly divided the 27 sentences into three sets of 9. Each set included one crossover sentence from another set so that we could check if the responses by workers from each set were statistically similar. Additionally, we used three sentences "This is so-so", "This is good" and "This is excellent" as controls to ensure the quality of their work. We expected workers to rate these three sentences in ascending order of intensity of judgment, and rejected results from workers who reversed the order of intensity of judgment for these three control sentences or who placed 2 or more of them at the same level of intensity. The workers were allowed to place both "This is so-so" and "This is good" in the neither weak nor strong category, however. We also rejected results from the workers if there were any empty responses, if all the responses were of the same intensity, or if there appeared to be a systematic clicking of responses. Workers were paid USD0.50 per set of 13 sentences, and were paid on average about USD12.55 per hour, which is approximately the living wage for a single adult on East and West Coast metropolitan cities of the US.

RESULTS

Data Set

To illustrate the method, we constructed a scenario in which designers may have a set of preferences for prescribed alternatives. They may also have preferences over attributes for each of the alternatives, which influence their preference for the alternatives. We set up the experimental condition wherein they are discussing the reasons for their preference for an alternative.

We have described this dataset in a prior paper [9], and continue to use this dataset for continuity. The team's task (below) was to choose a carafe (of glass, plastic, or steel) and filter (of gold, paper, or titanium) for a coffeemaker, each with three possible design alternatives. Note that in this paper, only the transcript statements regarding the carafe were analyzed:

Imagine you are a retired person who is a coffee connoisseur. Your day cannot begin until you make coffee each morning for you and your spouse. You are in good health but are not as strong or mobile as you were when you were younger. As a connoisseur, you prefer fresh ground coffee to instant coffee like Folger's, and you are well informed about the various types of gourmet coffee available, as well as the tools and equipment to prepare it. However, you are now on a fixed income and are conscious about how you spend your money, which is why you make coffee at home rather than visit Peet's every morning.

The team was told that the total cost for the carafe and filter could not exceed $35. Prior to the experiment, each participant was trained using a think-aloud exercise to practice saying each alternative using its proper name ("glass carafe" or "glass pot") rather than an ambiguous pronoun ("this" or "that") in order to facilitate the tracking of design alternatives in the transcript. During the experiment, they discussed their preferences and rationale with each other until a consensus was reached. This discussion was audio- and video-recorded and then transcribed.

During the same exercise, participants were asked to fill out surveys approximately every 10 minutes with their preference ratings for the alternatives. The experiment lasted 50

minutes, including 10 minutes for instruction and training, and 8 minutes for filling out 5 surveys during the session. Paper-based surveys were completed individually. Individuals were asked to provide an optional, brief rationale for their rating and ranking to decrease the possibility of arbitrary ratings.

Research on how groups engage in discussion suggests that members begin a discussion with only partial, independent knowledge of a topic. Group discussion can then play a role in eliciting this incomplete knowledge so that better decisions may be made [20]. In order to encourage discussion among the group members as well as better simulate a more realistic team experience, information about the design choices was provided in the following ways. First, team members were individually provided with detailed information about one of the three alternatives (for example, only the glass carafe), thus simulating a partial knowledge scenario. Team members would then discuss product features as a group in order to uncover additional information about the other alternatives.

The appraisals in the data set were analyzed by AD and MCY. Twenty-seven statements containing appraisals were identified in the transcript. Seven statements were used for training and arbitration purposes to ensure that the two coders could code the transcript consistently and reliably. A total of 72 semantic resources of appraisal were coded. AD checked each coder's work for consistency and made corrections where needed. A Krippendorf's alpha of 0.8188 [21] for intercoder reliability was achieved, which is considered acceptable.

Language Data

Three batches of statements were run, with a remuneration of USD0.50 per batch of responses, known as a "HIT" in Mechanical Turk jargon. 100 valid responses were taken from each batch. The batch statistics are shown in Table 1.

Table 1 Mechanical Turk work

HIT	Workers	Average Time to Complete (minutes)	Effective Hourly Rate (USD)
1	150	2	13.24
2	149	2	13.24
3	138	2	11.18

Descriptive statistics for the 3 control statements are shown in Table 2. The intensity of the judgment increased in line with the expected direction. We note also that the standard deviation for the intensity of the judgment decreases with the strength of the judgment. This implies that there is a higher level of uncertainty in weaker appraisals, a result that we find in the intensity judgments of the 27 statements.

Table 2 Intensity judgments for control statements

	N	Mean	Std. Deviation
This is so-so	300	1.87	.820
This is good	300	3.59	.714
This is excellent	300	4.84	.452

Descriptive statistics for the 27 statements rated by the Mechanical Turk workers are shown in Table 3. The statements were generated in order of predicted intensity of judgment within a set of 9 statements (Q1-Q9, Q10-Q18, and Q19-Q27), and are shown in this order. However, the statements were presented in random order to the workers, and workers received statements from across the sets. Generally, the trend is of increasing intensity of judgment within each of these sets. There is a recurrent pattern of a drop in intensity judgment between statements Q6 and Q7, Q15 and Q16, and Q24 and Q25. Each of the lower value statements combined a high engagement with a low graduation, such as "I really think that this is sort of good" and "I really think that this is sort of the best." In general, statements with a low value for the semantic resource of Graduation (Q1, Q4 and Q7; Q10, Q13, and Q16; Q19, Q22 and Q25) received the lowest rating of intensity within their respective sets. The consistency of these results across the sets further confirms the validity of the data and that the use of the semantic resource of Graduation with a low gradable value will produce the weakest judgments.

Table 3 Intensity judgments by Mechanical Turk workers

	N	Mean	Std. Deviation
Q1	100	2.22	.811
Q2	100	3.08	.761
Q3	100	2.92	.884
Q4	100	2.74	.747
Q5	100	3.43	.728
Q6	100	3.61	.803
Q7	100	2.94	.763
Q8	100	4.11	.665
Q9	100	4.41	.570
Q10	100	2.36	.871
Q11	100	3.18	.821
Q12	100	2.83	.911
Q13	100	2.75	.903
Q14	100	3.47	.958
Q15	100	3.62	.801
Q16	100	3.11	.680
Q17	100	4.05	.687
Q18	100	4.09	.793
Q19	100	2.98	1.155
Q20	100	3.49	.980
Q21	100	3.83	.911
Q22	100	3.35	1.029
Q23	100	3.94	.983
Q24	100	4.31	.849
Q25	100	3.64	.927
Q26	100	4.62	.599
Q27	100	4.81	.443

Independent samples t-test statistics were calculated for the crossover statements, that is, for statements repeated across the sets of statements that the Mechanical Turk workers rated. The difference in mean values were not statistically significant at the α=0.01 level for one question, Q14 [$t(198) = -2.136$, $p=0.022$], and at the α=0.05 level for two questions, Q9 [$t(198) = 0.412$, $p=0.053$] and Q16 [$t(198) = 0.793$, $p=0.933$]. Similarly, a one-way between group ANOVA test was conducted to compare the effect of the different workers on the control questions. There was no significant effect at the α=0.01 level for "This is so-so" [$F(2,297) = 3.635$, $p=0.028$], and no significant effect at the α=0.05 level for "This is good"

[F(2,297) = 0.339, p=0.713] and "This is excellent" [F(2,297) = 1.492, p=0.227]. We note that the weaker the judgment, the slighter stronger the statistical significance that the workers differ in the way that they rank the intensity of the judgment. This suggests that weaker appraisals are accompanied with more uncertainty, and, more specifically, that the semantic resource of Engagement increases the 'spread' of the results. In summary, this data is sufficiently valid for the purpose of calculating the state transition probabilities.

Time-based preferential probabilities

The final element of this work is to quantify the preference information for each alternative based on captured linguistic information. The model given in the section *Markov Model* describes a time dependent preferential probability. A key component of this Markov model is a state transition probability that depends on the linguistic appraisal. This transition probability depends on two distinct factors of appraisal. The first factor is the appraisal strength, obtained from the Mechanical Turk data. The second factor is whether the orientation of the appraisal increases or decreases the preferential probability. A positive appraisal increases the preferential probability, whereas a negative appraisal decreases it. This factor is non-trivial to determine because we need to map the appraisal of an alternative by itself or an alternative based on one or more of its attributes into changes in its preferential probability. This factor should depend on the importance of a particular attribute in the overall design. However, for this paper, we assume that we can obtain an average over all attributes. Thus, the transition probability for increasing (positive) and decreasing (negative) appraisals will be given in the form below.

Without any loss of generality, let us assume that the linguistic information at time i (e_i) is about alternative m. If an appraisal is positive in orientation with strength S_i, then Eq. 3 gives each element p_{uv} in the state transition matrix where u=1 to N and v=1 to N:

$$p_{uv}(e_i) = \begin{cases} c_+ S_i & \text{if } u = m \text{ and } u \neq v \\ 1 - c_+ S_i & \text{if } u = v \text{ and } i \neq m \\ 1 & u = v = m \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where c_+ is a positive factor that determines how much the average positive appraisal for any attribute changes the preferential probability. This transition probability essentially increases the preferential probability for alternative m by transferring part of the preferential probability from other alternatives. There is a bound for these two factors, $0 \le c_+ \le 1$ and $0 \le S_i \le 1$. An example of S_i=1 is when a designer states, "This is the cheapest design ever", which corresponds to a 5 rating in the Mechanical Turk data. An example of c_+=1 is when design teams behave such that *any* appraisal of strength S_i=1 for alternative m causes alternative m to be the most preferred alternative independent of preferences from an earlier time. When c_+=0, *any* appraisal does not make any difference in a design team's decision, such as when a design team already has its mind set on a particular alternative.

Similarly, Eq. 4 gives the state transition probability matrix for a negative appraisal with strength S_i (negative appraisal) about alternative m:

$$p_{uv}(e_i) = \begin{cases} 1 + c_- S_i & u = u = m \\ -1 \cdot c_- S_i / (N-1) & \text{if } u \neq v \text{ and } v = m \\ 1 & \text{if } u = v \text{ and } v \neq m \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

This state transition matrix transfers the preferential probability from alternative m to all other alternatives equally. Note that c- is bounded by $-1 \le c_- \le 0$. Similar to the c+ case, when c- is equal to 0, then any appraisal does not influence the preferential probability. When c- is -1, then the strongest negative appraisal about any attribute about any alternative will reduce the corresponding preferential probability to 0. This represents a design team that discards an alternative for any bad attributes regardless of how the alternative performs on other attributes. Determining the value of c+ and c- is critical for extending this work, but it is outside the scope of this paper. Instead, we will show how to understand the result of the proposed method given parametric uncertainty in c+ and c- values.

We have implemented this Markov model on the coffee carafe data set using the Mechanical Turk appraisal data. Only a few possible combinations of Attitude, Engagement, and Graduation occurred in the transcript (see Figure 3). This suggests that there may be a few combinations of appraisals that we need to understand carefully. Each of these appraisals has an underlying uncertainty distribution that has been captured by Mechanical Turk. Refer to Figure 4 for the distribution for an appraisal with low gradable values for each semantic resource. Some appraisals have more uncertainty in strength than others, as shown in Figure 5, which compares the means and standard deviations for all 27 appraisals and 3 control statements sampled from Mechanical Turk. This result demonstrates the necessity of quantifying the impact of these uncertainties on the calculated preferential probabilities.

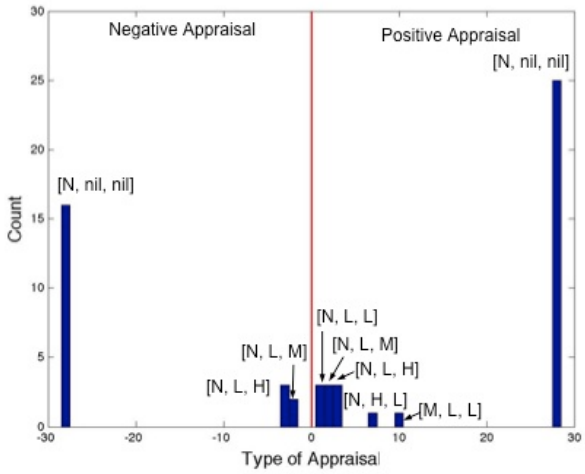

Figure 3 Appraisals in coffee carafe data set

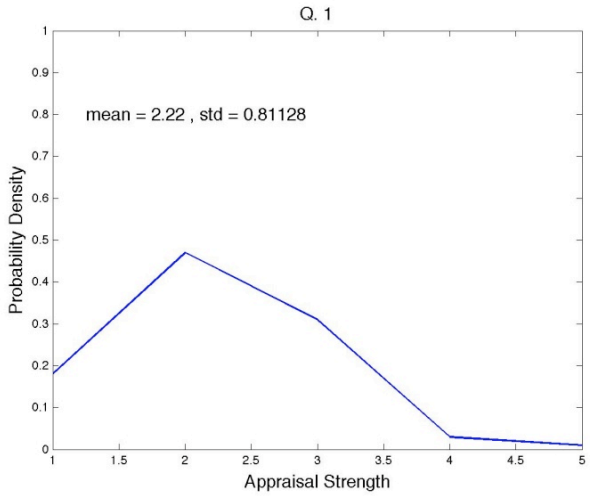

Figure 4 Probability density for appraisal [L,L,L]

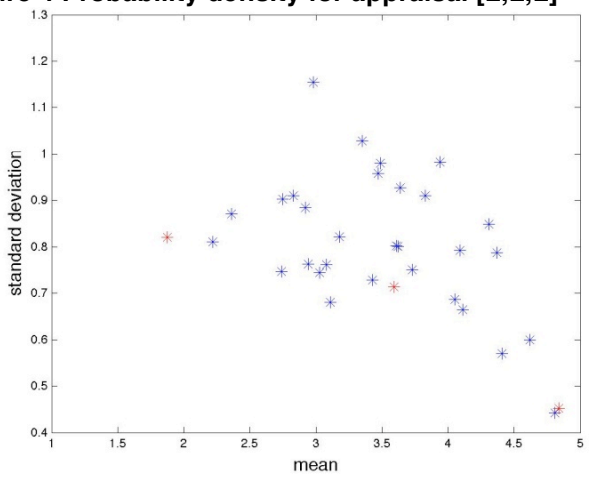

Figure 5 Mean vs Standard Deviation for 27 Questions (blue) and 3 control statements (red)

In this initial analysis, we fix c_+ and c_- and focus on the quantifiable appraisal uncertainty. We can obtain the probability density for the strength of an appraisal (e.g., Figure 4) by assuming that 100 samples are sufficient to approximate the underlying distribution. We verified this by comparing 3 different samples for the control statement "This is good" (Figure 6). Given these probability distributions and assumed c_+ and c_-, we can run a Monte Carlo simulation to determine the distribution of preferential probabilities (Figure 7).

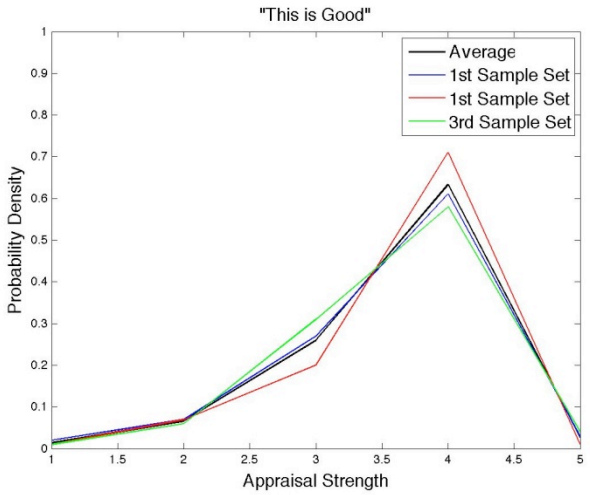

Figure 6 Sanity check for utilizing experimental probability distribution

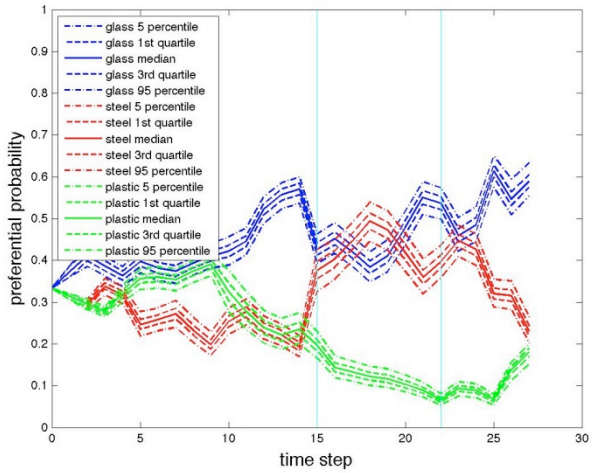

Figure 7 Uncertainty distribution for the preferential probability represented using percentiles when c+ = 0.1 and c-= 0.2

This Markovian based model is sensitive to the occurrence of an appraisal during the team discussion. An appraisal that occurs later in time has a much higher impact than in our previous work on Appraisal PPT [9]. This creates a situation where the uncertainty distribution for a preferential probability for the glass carafe overlaps with the one for steel (see Figure 8). This overlap represents the fact that for some design teams with a particular set of appraisals, steel may be more likely to be preferred over glass at some periods of time. Given this information shown in Figure 8 with the additional covariance information, we can calculate the probability of the team preferring steel the most. In this case, it is 15% at when time is 15, 0% when time is 22, and 0% when time is 27.

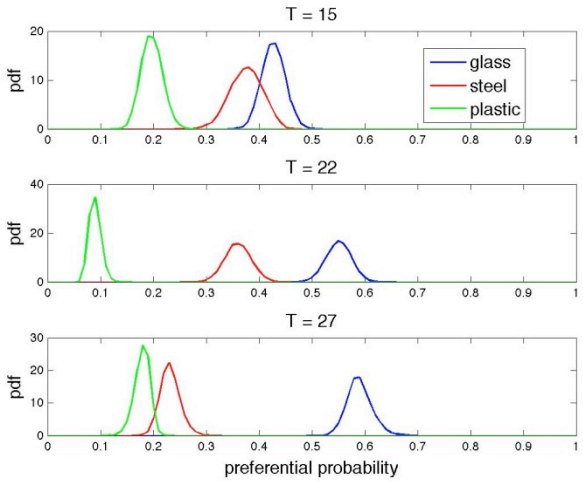

Figure 8 Uncertainty distribution for preferential probability right before survey is given

Finally, we need to address the issue of c_+ and c_-, which are hidden parameters. Figure 9 and Figure 10 show the sensitivity analysis for c_+ and c_-, respectively. It illustrates that for some time steps, the result is highly sensitive to c_+ and c_- values, but not for other time steps. Furthermore, if there is a particular time step that a designer cares about, we can create a boundary in (c_+ and c_-) space for the most likely preferred alternatives (Figure 11). This sensitivity map provides a mechanism for a designer to determine his or her confidence that the result is robust to all uncertainties. As the designer or design team converges to a particular alternative, this sensitivity map should show more robustness to the c_+ and c_- values, as shown in Figure 12.

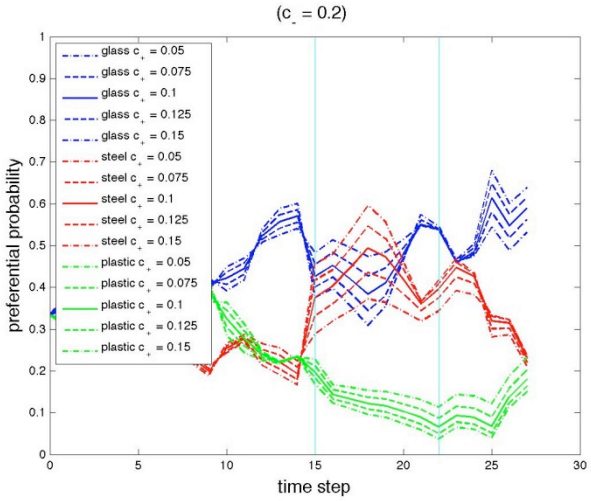

Figure 9 Sensitivity of median values with respect to c_+ values

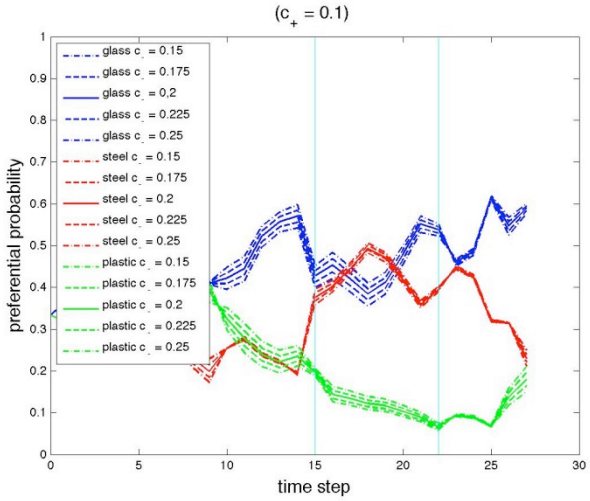

Figure 10 Sensitivity of median values with respect to c_- values

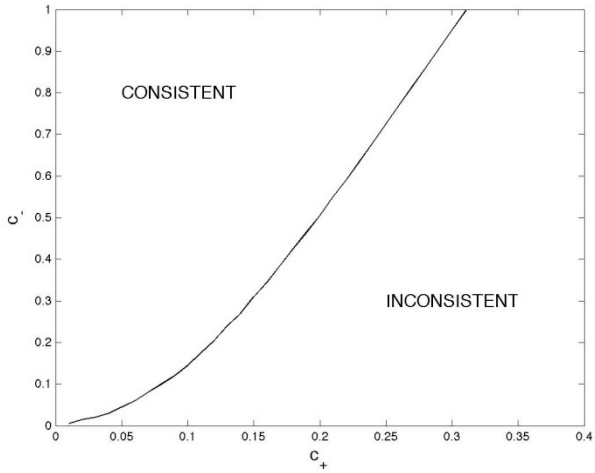

Figure 11 Boundary in (c_+,c_-) space for most likely preferred alternatives

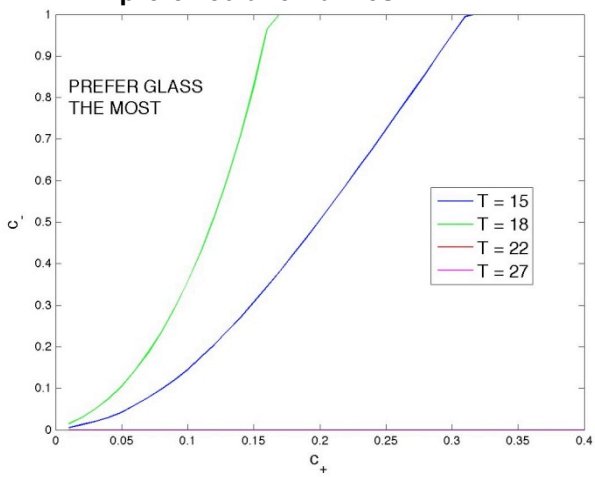

Figure 12 Time Dependence on the Sensitivity c_+ and c_- values

CONCLUSIONS

We have presented a method to estimate preferential probabilities for the selection of an alternative from a mutually

exclusive set by a committee such as a small design team or a design review panel. While the case study is limited to a synchronous discussion, we believe that the reported methodology can be extended to asynchronous discussions over a longer time scale wherein alternative concepts are continually being developed and refined. Further, we believe that the method would apply toward the elicitation of preferences from customers, who would describe their 'likes and dislikes' for each alternative and then choose the most preferred alternative.

This research develops a seed of systematization for the study of choice and subjective report without the need for direct elicitation. It also provides a more natural way to gather information about preferences in a more realistic way to elicit preferences, since preferences are not 'fixed' in the mind of the decision-maker but are subject to change from discussion, negotiation, further knowledge, and interaction with each alternative.

It is important to emphasize that the method provides a descriptive model of decision-making, not a normative one. Further, it neither directs the committee to make a specific choice nor assists the committee to make a utility-maximizing decision. Instead, we believe that this type of approach provides a quality control tool for decisions [22], especially when decision makers do not formally model their decisions. In situations wherein decisions are only talked about but not modeled, perhaps due to the complexity of the decision, there is nonetheless the expectation that the individuals used rational thought to guide their formation of subjective preferences and that the committee deliberated rigorously. We believe that identifying discrepancies between what a committee decides and what a committee says, literally, they will choose is perhaps the most valuable contribution that this work could make to decision-based design. In such a situation, a quality control question that could be asked is whether the decision that was taken is consistent with the degree of positive appraisal of an alternative (or negative appraisals of the alternatives) and the certainty of those appraisals. Did the committee choose the alternative that they were most positive about or did they choose some other alternative? In other words, this descriptive model can be compared to the outcome of the decision process, since the outcome is known with certainty. If there is a discrepancy between the descriptive model and the actual outcome, then the committee can be directed to review the decision. Other possibilities for quality control exist. The committee might ask if they were overly optimistic about a particular alternative, based on the existence of very strong positive appraisals for a particular alternative. Perhaps there was a "halo effect" in which once a very strong positive appraisal for an alternative was given, all other attributes for that alternative were deemed exemplary even if there is no correlation between the qualities of those attributes. The committee might ask how certain they are about the decision, and match up their level of perceived certainty with the level of uncertainty as expressed in their linguistic appraisals and calculated by the probability distribution of the preferential probabilities. In short, descriptive models of decision-making based on natural language provide a tool to inspect decisions and could form the basis of quality control mechanisms for decisions.

While this paper makes no claim to the cognition of decision-making, the possibility of *formally* detailing decision-making through language could give researchers both in engineering design and in other fields a new way to understand the cognitive processes behind decision-making. The calculation of preferential probabilities assumes mutually exclusive alternatives. While it is possible for a team to state a joint preference ("I think option A and option B are good"), we have not yet encountered sufficient linguistic evidence to collect data to form a joint distribution. In the future, we will continue to validate and develop these methods using a new data set that we are transcribing and analyzing of committees selecting from one of 7 innovative projects.

ACKNOWLEDGEMENTS

The work described in this paper was supported in part by the National Science Foundation under Award CMMI-0900255. This research was also supported in part under Australian Research Council's Discovery Projects funding scheme (project number DP1095601). The opinions, findings, conclusions and recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Yang, M. C., 2010, "Consensus and Single Leader Decision-Making in Teams Using Structured Design Methods," Design Studies, 31(4), pp. 345-362.
[2] Delbecq, A. L., and Mills, P. K., 1985, "Managerial Practices That Enhance Innovation," Organizational Dynamics, 14(1), pp. 24-34.
[3] Frey, D., Herder, P., Wijnia, Y., Subrahmanian, E., Katsikopoulos, K., and Clausing, D., 2009, "The Pugh Controlled Convergence Method: Model-Based Evaluation and Implications for Design Theory," Research in Engineering Design, 20(1), pp. 41-58.
[4] Hazelrigg, G., 1998, "A Framework for Decision-Based Design," ASME Journal of Mechanical Design, 120(4), pp. 653-658.
[5] Reich, Y., 2010, "My Method Is Better!," Research in Engineering Design, 21(3), pp. 137-142.
[6] Scott, M. J., and Antonsson, E. K., 1999, "Arrow's Theorem and Engineering Design Decision Making," Research in Engineering Design, 11(4), pp. 218-228.
[7] Thurston, D. L., 1991, "A Formal Method for Subjective Design Evaluation with Multiple Attributes," Research in Engineering Design, 3(pp. 105-122.
[8] López-Mesa, B., and Bylund, N., 2011, "A Study of the Use of Concept Selection Methods from inside a Company," Research in Engineering Design, 22(1), pp. 7-27.
[9] Honda, T., Yang, M. C., Dong, A., and Ji, H., 2010, "Understanding the Language of Preference and Uncertainty in Engineering Design," Proc. 22nd International Conference on Design Theory and Methodology (DTM), Montreal, 5, pp.DETC2010-29045.
[10] Sah, R. K., and Stiglitz, J. E., 1988, "Committees, Hierarchies and Polyarchies," The Economic Journal, 98(391), pp. 451-470.

[11] Dong, A., Kleinsmann, M., and Valkenburg, R., 2009, "Affect-in-Cognition through the Language of Appraisals," Design Studies, 30(2), pp. 138-153.
[12] Dave, K., Lawrence, S., and Pennock, D. M., 2003, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," Proc. Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, pp.519-528.
[13] Dong, A., 2006, "How Am I Doing? The Language of Appraisal in Design," Proc. Design Computing and Cognition '06 (DCC06), J. S. Gero, ed. Eindhoven, The Netherlands, pp.385-404.
[14] Dong, A., 2009, The Language of Design: Theory and Computation, Springer, London.
[15] Martin, J. R., and White, P. R. R., 2005, The Language of Evaluation: Appraisal in English, Palgrave Macmillan, New York.
[16] Whitelaw, C., Garg, N., and Argamon, S., 2005, "Using Appraisal Groups for Sentiment Analysis," Proc. CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, pp.625-631.
[17] Subasic, P., and Huettner, A., 2001, "Affect Analysis of Text Using Fuzzy Semantic Typing," IEEE Transactions on Fuzzy Systems, 9(4), pp. 483-496.
[18] Paolacci, G., Chandler, J., and Ipeirotis, P. G., 2010, "Running Experiments on Amazon Mechanical Turk," Judgment and Decision Making, 5(5), pp. 411-419.
[19] Sorokin, A., and Forsyth, D., 2008, "Utility Data Annotation with Amazon Mechanical Turk," Proc. Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on, pp. 1-8.
[20] Gigone, D., and Hastie, R., 1997, "The Impact of Information on Small Group Choice," Journal of Personality and Social Psychology, 72(1), pp. 132-140.
[21] Hayes, A. F., and Krippendorff, K., 2007, "Answering the Call for a Standard Reliability Measure for Coding Data," Communication Methods and Measures, 1(1), pp. 77 - 89.
[22] Kahneman, D., Lovallo, D., and Sibony, O., 2011, "Before You Make That Big Decision..." Harvard Business Review, 89(6).

ANNEX A

APPRAISALS RATED BY MECHANICAL TURK WORKERS

The following table presents the appraisals rated by the Mechanical Turk workers. Each of the statements uses all three semantic resources, Attitude, Engagement, and Graduation, to form an appraisal. The level of the gradable resource is indicated by the values (L)ow, (M)edium and (H)igh.

Table 4 Statements Rated by Mechanical Turk Workers. A=Attitude (L=good; M=better; H=best); E=Engagement (L=sort of; M=pretty much; H=really); G=Graduation (L=sort of; M=much/quite; H=very/so much)

Option	A	E	G	Statement
Q1	L	L	L	I sort of think that this is sort of good.
Q2	L	L	M	I sort of think that this is quite good.
Q3	L	L	H	I sort of think that this is very good.
Q4	L	M	L	I pretty much think that this is sort of good.
Q5	L	M	M	I pretty much think that this is quite good.
Q6	L	M	H	I pretty much think that this is very good.
Q7	L	H	L	I really think that this is sort of good.
Q8	L	H	M	I really think that this is quite good.
Q9	L	H	H	I really think that this is very good.
Q10	M	L	L	I sort of think that this is sort of better.
Q11	M	L	M	I sort of think that this is much better.
Q12	M	L	H	I sort of think that this is so much better.
Q13	M	M	L	I pretty much think that this is sort of better.
Q14	M	M	M	I pretty much think that this is much better.
Q15	M	M	H	I pretty much think that this is so much better.
Q16	M	H	L	I really think that this is sort of better.
Q17	M	H	M	I really think that this is much better.
Q18	M	H	H	I really think that this is so much better.
Q19	H	L	L	I sort of think that this is sort of the best.
Q20	H	L	M	I sort of think that this is pretty much the best.
Q21	H	L	H	I sort of think that this is the very best.
Q22	H	M	L	I pretty much think that this is sort of the best.
Q23	H	M	M	I pretty much think that this is quite the best.
Q24	H	M	H	I pretty much think that this is the very best.
Q25	H	H	L	I really think that this is sort of the best.
Q26	H	H	M	I really think that this is quite the best.
Q27	H	H	H	I really think that this is the very best.

