Research in Engineering Design manuscript No.
(will be inserted by the editor)

A linguistic approach to assess the dynamics of design team preference in concept selection

Andy Dong · Somwrita Sarkar · Maria C. Yang · Tomonori Honda

Received: date / Accepted: date

Abstract This paper addresses the problem of describing the decision-making process of a committee of engineers based upon their verbalized linguistic appraisals of alternatives. First, we show a way to model an individual's evaluation of an alternative through natural language based on the Systemic-Functional Linguistics system of APPRAISAL. The linguistic model accounts for both the degree of intensity and the uncertainty of expressed evaluations. Second, this multi-dimensional linguistic model is converted into a scalar to represent the degree of intensity and a probability distribution function for the stated evaluation. Finally, we present a Markovian model to calculate the time-varying change in preferential probability, the probability that an alternative is the most preferred alternative. We further demonstrate how preferential probability toward attributes of alternatives correspond to preferential probability toward alternatives. We illustrate the method on two case studies to highlight the time-variant dynamics of preferences toward alternatives and attributes. This research contributes to process tracing in descriptive decision science to understand how engineers actually take decisions.

Keywords Decision based design · Ranking Alternatives · Social choice

1 Introduction

One of the classic decision-making problems in engineering design is concept selection, the analysis and evaluation of alternative concepts, leading to the selection or consolidation of one or more concepts for further development. A range of normative decision-making tools and methods for concept selection exist, including concept screening [33], pair-wise comparison charts [5], concept scoring matrices [6, 24], and multi attribute utility analysis [26, 31]. Further along the engineering design process, utility-theory based methods have been developed to support the selection of design parameters so as to optimize or trade-off the objectives in a manner that adheres to the axioms of expected utility theory [32].

Yet, at the fuzzy front end of engineering design, information is the most lacking, objectives are still being ascertained, and outcomes are uncertain. At the

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.

A. Dong ✉
Faculty of Engineering and Information Technologies
University of Sydney
Sydney, NSW 2006 Australia
E-mail: andy.dong@sydney.edu.au

S. Sarkar
Design Lab
Faculty of Architecture, Design, and Planning
University of Sydney
Sydney, NSW 2006 Australia
E-mail: somwrita.sarkar@sydney.edu.au

M. C. Yang
Department of Mechanical Engineering and Engineering Systems Division
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
E-mail: mcyang@mit.edu

T. Honda
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
E-mail: tomonori@mit.edu

stage when a company is in between ideation and making a strong commitment to take a project forward, preferences toward alternatives can dynamically change as a result of discussion between stakeholders, interaction with the alternatives, and the introduction of more information, all of which could update preferences and uncertainties [18]. There is also the problem of equivocality, wherein stakeholders may have differing and competing interpretations of likelihoods of success, which requires negotiation and discussion between stakeholders to 'make sense' of the situation [36]. In short, during this phase, preferences are mutable because uncertainties are high and equivocality is low [8]. Under such circumstances, supporting the decision-makers to resolve uncertainties by pushing them toward accuracy through the axioms of expected utility theory downplays the dynamics of preference change. Instead, it may be more useful to provide descriptions of the time-varying dynamics of preferences toward alternatives so that the design team can scrutinize their decision-making process.

The purpose of this research is to model the decision that a committee will take based upon what each committee member has stated about preferences toward an alternative or a set of alternatives being assessed. The method's linguistic model and numerical data for appraisal values are based on the way that English speakers express evaluations and judgments. Hence, the method could be described as a model for how an English speaker would interpret a committee's preference based on the committee's linguistic appraisals toward alternatives. In engineering, as with many fields, decisions are not always formally modelled but only spoken or written about. There is a need for decision support tools to provide process tracing, thereby increasing the accountability and transparency of decisions. Our model will adopt a natural language analysis based approach to describe the preferences toward alternatives. This is particularly important because despite the presence of analytical, mathematical, or logically based systems of decision making (such as expected utility maximization), natural language continues to be the most commonly used form of communication for the purposes of exchanging subjective preferences.

In this study, we constrain ourselves to the situation in which designers may have a set of preferences for a discrete set of mutually exclusive alternatives. They may also have preferences over attributes associated with the alternatives (i.e., decision criteria), which influence their preference toward the alternatives. This situation is similar to concept scoring. In concept scoring, the committee prepares a matrix of design alternatives, generally with a reference alternative, and a set of selection criteria or attributes associated with each of the alternatives. The committee can either equally weight the selection criteria or provide differential weights. Through discussion, the committee assigns a rating to each alternative; the rating can be relative to the reference or ordinal. If $\mathbf{D} = [d_1, d_2, \ldots, d_N]$ is the set of design alternatives, $\mathbf{A} = [a_1, a_2, \ldots, a_M]$ is the set of selection criteria, and $\mathbf{R} = [r_{11}, r_{12}, \ldots, r_{1n}; \ddots ; r_{M1}, r_{M2}, \ldots, r_{MN}]$ is the matrix of ratings, then the total score S_i for each alternative i is given by $S_i = a_1 r_{1i} + a_2 r_{2i} + \ldots + a_M r_{Mi}$. In our model, we use linguistic evidence to update the values of the weights on the selection criteria and the ratings to obtain the score as a *preferential probability*: the probability that an alternative or attribute is the most preferred (important) at a given time.

This method builds upon our prior research in developing probabilistic approaches for estimating a design team's preference toward alternatives that are described by a set of attributes [15]. One crucial assumption in our prior method was that individuals speak more often (and in a positive light) about alternatives for which they have a stronger preference and less often about those they prefer less. In other words, the frequency of occurrence of an utterance about an alternative was considered in building a probabilistic model to describe the relationship between preferences over consecutive time intervals. From the time-variant model, we estimated how likely a choice is to be most preferred by a design team over a given period of time. The preference model based on linguistic data was compared to the preference model based on survey data containing the participants' preference ratings (between 0 and 1) for each design alternative at periodic intervals. We showed a correspondence between the utterance data and the survey data, which confirmed our assumption. Nonetheless, we recognize that this assumption may not hold true in all situations, even though frequency of occurrence turned out to be a reasonable, manifest indicator of preference.

In this article, we recognize that there is preference information embedded in the syntax and semantics of linguistic appraisals. As such, we expand upon our prior method to take into account the grammar of the linguistic expression of appraisal toward an alternative as an indication of an engineer's preference toward an alternative. Team members can express their preferences as attitudes toward attributes using finely calibrated language, and these nuances in attitude are evaluated by applying an established model of language. Second, we note that individuals consider not only alternatives but also specific attributes of these alternatives. A team may prefer one automobile design over another, but this preference presumably grows out of a combination of preferences for particular attributes such as fuel efficiency, acceleration, and handling. This paper offers investigations of both of these circumstances.

2 Mathematical Model of Preference

Consider the following general empirical scenario that we use to formalize the model. Suppose we have an engineer or a group of engineers discussing their preferences toward alternatives on the basis of how the alternatives perform over a set of attributes. In general, these attributes can be either explicitly set out as in a design brief or implicitly constructed by the engineers during an evolving design meeting [19].

Let N be the total number of design alternatives, and M be the total number of design attributes. In the model development, N and M are assumed fixed for the duration of the design session. This may be an incomplete modelling assumption for the ideation process when the number of alternatives or attributes are still emerging but is not an impractical modelling assumption for the later stages of conceptual design when the number of alternatives and attributes are more fixed. Nonetheless, it is not a restriction on the model. The model is valid even when N and M are not fixed. The elimination of an alternative or attribute is trivial as its preferential probability and transition probabilities can be set to 0. If they increase, then the number of rows and columns in the transition probabilities matrix is increased by a corresponding amount, and the length of the vector of preferential probabilities is also increased. For clarity of presentation and no loss of generality, we keep them fixed in the model development.

Let $\mathbf{D} = [d_1, d_2, \ldots, d_N]$ describe the vector of design alternatives, and $\mathbf{A} = [a_1, a_2, \ldots, a_M]$ describe the vector of design attributes over which the alternatives are being assessed. Time $t = 0$ signifies the start of the design session. Thereafter, each utterance by each member of the team is considered a discrete time step. Figure 1 shows an illustrative example of appraisals of alternatives based on a set of attributes, with alternatives and attributes *emphasized.*

d_1 *Glass coffee carafe* seems to have the most *capacity.*
d_2 So there is a drawback for *stainless-steel* because it is *heavy.*
d_3 *Plastic carafe* No, it is not *easy to clean.* It's not *attractive.*

Fig. 1 Sample linguistic appraisals of alternatives with alternatives and attributes typeset in italics

We now develop the preference model. We will compute the probability that a certain design alternative d is the most preferred (dominant) design alternative at time step t_i. Obviously, there will be some preference value (positive or negative) for the other non-dominant design alternatives. We represent these preference values by υ. Then the probability that d is the most preferred (strictly dominating) alternative at time step i, with the preference value $\upsilon_d > \upsilon_j, j = 1, \ldots, N, j \neq d$. If the strict dominance is replaced with a relaxed definition of dominance, then $\upsilon_d \geq \upsilon_j, j = 1, \ldots, N, j \neq d$, with no loss of generality. Thus, if there are two or more design alternatives that are equally preferred, this does not alter the model in any way.

To compute the time-varying preferential probabilities, we present a Markov chain model. The Markov model consists of a dynamic transition matrix of probabilities that contain a trace of how linguistic utterances may reflect actual preference value change. Suppose the team must select one choice from N alternatives, and assume that the alternatives are mutually exclusive and independent. Then, the vector of preferential probabilities at any time step t is:

$$\mathbf{P}_t = [\pi_{1,t}, \pi_{2,t}, \ldots, \pi_{N,t}] \tag{1}$$

and $\sum \mathbf{P}_t = 1$. In the absence of any prior knowledge, at the start $t = 0$, the probability that any of the alternatives is the most preferred alternative can be considered equivalent. That is, without any loss of generality, the initial probability that any alternative is the most preferred one is $\frac{1}{N}$:

$$\mathbf{P}_0 = [\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}] \tag{2}$$

For example, we consider a small example where the designers are choosing between $N = 3$ alternatives. Then, the initial probability that any of the alternatives 1, 2, or 3 is the most preferred is $\frac{1}{3}$. Now let's assume that a designer says, "I really like the first one." Between time step $t = 0$ and $t = 1$, we should be able to calculate a transition probability to reflect the linguistic evidence that tells us that the subjective preference toward alternative 1 may be increasing. As a consequence of obtaining the linguistic data, the preferential probability should transition from all alternatives having equal probability to alternative 1 having a higher preferential probability. Preferential probabilities for alternatives 2 and 3 should likewise decrease. To calculate the preference transition, we need the appraisal value toward alternative 1. This appraisal value is used to update the transition matrix and compute the preferential probability. Thus, we have to compute preferential probabilities for each pairwise transition: what is the probability that alternative i is the most preferred alternative at time step t, given that alternative j was the most preferred in time step $t-1$ and so on. The transition probability matrix $\mathbf{T}$ for N alternatives is represented by:

$$\mathbf{T} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix} \tag{3}$$

At the initial state $t = 0$, similar to the vector of preferential probabilities $\mathbf{P_0}$ in Eq. (2), $\mathbf{T}$ is initialized to:

$$\mathbf{T}_0 = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix} \tag{4}$$

If $\mathbf{P}_0$ is already biased because of prior knowledge, it does not make sense to have $\mathbf{T}_0$ be unbiased, because $\mathbf{P}_0\mathbf{T}_0$ would push the preferential probabilities toward a uniform distribution. Ideally, $\mathbf{P}_0$ is a stable equilibrium state of $\mathbf{T}_0$ such that $\mathbf{P}_0\mathbf{T}_0 = \mathbf{P}_0$. This initial distribution satisfies this condition. (Note that a $\mathbf{T}_0$ that satisfies this condition is not unique.)

To compute the transition matrix at each time step, we apply the appraisal values obtained from linguistic data. Suppose at time step t, linguistic data is obtained such that the appraisal value for alternative 1 is positive $+a$. If the linguistic data for alternative 1 is a negative appraisal, equivalently the appraisal value will be $-a$ with no loss of generality. Then, the transition matrix can be updated as:

$$\mathbf{T} = \begin{bmatrix} (p_{11} + a) & p_{12} - (\frac{a}{N-1}) & \cdots & p_{1N} - (\frac{a}{N-1}) \\ p_{21} + a & p_{22} & \cdots & p_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1} + a & p_{N2} & \cdots & p_{NN} \end{bmatrix} \tag{5}$$

The idea is that the transition probabilities for alternative 1, relative to the transition probabilities for all of the other alternatives in the previous time step, should increase by the appraisal value $+a$ when we receive a positive appraisal for alternative 1 (or equivalently, decrease by $-a$ when we receive a negative appraisal for alternative 1). In the language of Markov chains, this is stated as: the probability that alternative 1 is the preferred alternative in the current time step, given that another alternative was the preferred alternative in the prior time step, is increased by $+a$. This is reflected in the first column of Eq. (5). The transition probabilities for the other alternatives, relative to alternative 1, should then change in the opposite direction. However, we do not explicitly know by how much they should change. As such, we equally distribute the negative of the appraisal value to the remaining transition probabilities by a factor $-\frac{a}{N-1}$ under the assumption that the more that an individual prefers an alternative, then it is less likely that the individual prefers the other alternatives in equal proportion. Transition probabilities for alternatives for which we have not received any linguistic data should remain unchanged. To maintain transparency of the effect of the appraisal values on the transition probabilities, for the moment, we relax the rule requiring that each row in the transition probability matrix should sum to 1 as required by the axioms of probability.

An alternative formulation would be to modify the preference values v by the appraisal values a directly, increasing the value of v for an alternative for which we have received a positive appraisal and decreasing the values of v for the other alternatives proportionally. While this formulation may appear intuitively obvious, our experiments showed that this formulation has two problems. First, it makes a strong assumption that the appraisals are direct expressions of preference value. We believe that the linguistic appraisals are indicators of likely preference; that is, they affect the probability that an alternative is the most preferred alternative, which is why we have chosen to modify the transition probabilities with the appraisal values. Second, directly modifying the preference values with the appraisal values led to "runaway" preferences values, which did not agree with the surveys to identify the participants' actual preferences during the experiments. As such, we do not continue with this formulation.

In summary, transition probabilities for an alternative for which we receive appraisal data change relative to the other alternatives. Transition probabilities between any two alternatives for which we receive no appraisal data are kept the same. Mathematically stated, if an appraisal is received on alternative i, then we add the appraisal value a with the appropriate sign (positive or negative) to the i^{th} column and add its negative by a factor $\frac{a}{N-1}$ to the i^{th} row except for the ii^{th} entry.

Because of the presence of negative appraisals over time, $\mathbf{T}_t$ can have negative values. To correct this numerical artefact, we rescale the entries of $\mathbf{T}_t$ between 0 and 1, preserving the relative relationships between the transition probabilities but relaxing the requirement that the rows must sum to 1.[1]

$$\mathbf{T}_t^{scale} = \frac{\mathbf{T}_t - min(\mathbf{T}_t)}{max(\mathbf{T}_t - min(\mathbf{T}_t))}. \tag{6}$$

We note, however, that the unscaled transition matrix $\mathbf{T}$ is used to recalculate the new transition matrix in the next iteration. Using the unscaled version of $\mathbf{T}$ preserves the numerical effect of repeated positive or negative appraisals over time, leading to a more pronounced numerical distinction between the preferential probabilities over time. It also maintains visual transparency of the effect of repeated positive or negative appraisals of an alternative, which is important for human understanding of the method.

Now, given $\mathbf{P}$ at time $t-1$ and the rescaled $\mathbf{T}_t^{scale}$ at time t, $\mathbf{P}$ at time t is updated by vector-matrix multiplication:

$$\mathbf{P}_t = \mathbf{P}_{t-1}\mathbf{T}_t^{scale} \tag{7}$$

In a final step, we normalize the values in $\mathbf{P}_t$ such that all values range between 0 and 1 and the sum totals 1 by dividing $\mathbf{P}_t$ by the sum of its entries:

$$\mathbf{P}_t^{norm} = \frac{\mathbf{P}_t}{\sum\limits_{i=1}^{N} \pi_{i,t}} \tag{8}$$

where $\pi_{i,t}$ are the preferential probabilities for the $i = 1$ to N alternatives. Note that this simple normalization does not lose the relational information gained in the model and the transition probabilities matrix, but simply rescales the preferential probability values between 0 and 1 and ensures that $\sum \mathbf{P}_t^{norm} = 1$, as required by axioms of probability.

An outcome of the model is that the preferential probability for an alternative increases (decreases) with the number and intensity of positive (negative) appraisals toward the alternative. If after a series of positive linguistic appraisals of high appraisal value a designer were to select (as in actually choose) any other alternative as the most preferred alternative, then there is an inconsistency between the designer's linguistic appraisals and decision; the method captures this important inconsistency.

For now, we ask the reader to accept an arbitrary value a for an appraisal. The problem of how to compute an appraisal value from linguistic data is a separate part of the model, which is presented in Section 3.

2.1 Hierarchical grouping of alternatives

In some design cases and experiments that we studied, the alternatives themselves could be hierarchically grouped. For example, in the Laptop configuration problem (see Table 7), there are 8 principal design alternative classes such as External shell, Screen size, etc. Each of these principal 8 categories contains further alternatives: for example, the External shell can be a plastic alloy (d_1), a magnesium alloy (d_2), or a titanium alloy (d_3). Such a hierarchical composition of alternatives is quite common. In such a case, the total number of alternatives is higher (31 alternatives for the laptop configuration case) than the number of classes into which they are grouped (8 alternative classes in the laptop configuration case). According to our method above, in the absence of hierarchical grouping of alternatives, each time a positive or negative appraisal is made on one alternative, all the rows and columns of the entire matrix are altered. However, if the alternatives are hierarchically grouped, then it will be unreasonable to alter all the rows and columns when the designer is actually making an appraisal within a particular alternative class. For example, the preferential probability on the Screen size should not change if the designer is making an appraisal on titanium, magnesium, or plastic as a choice for the External shell, assuming independence of categories. Only the preferential probabilities of the alternatives in that particular class should change. To account for this, whenever we study a design case with hierarchical grouping of alternatives, we perform a block matrix computation. To incorporate a new appraisal on an alternative, only the rows and columns of the sub-matrix comprising one alternative class are altered, leaving the other parts of the matrix untouched. In formal terms, suppose there are two alternative classes, with 1 to N alternatives in one class and $N+1$ to K alternatives in the second class. Then, Eq. (5) from the previous section will now become:

$$\mathbf{T} = \begin{bmatrix} (p_{11}+a) & p_{12} - (\frac{a}{N-1}) & \dots & p_{1N} - (\frac{a}{N-1}) & p_{1(N+1)} & \dots & p_{1K} \\ p_{21}+a & p_{22} & \dots & p_{2N} & p_{2(N+1)} & \dots & p_{2K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ p_{N1}+a & p_{N2} & \dots & p_{NN} & p_{N(N+1)} & \dots & p_{NK} \\ p_{N(N+1)} & \dots & \dots & \dots & \dots & \dots & p_{(N+1)K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ p_{K1} & \dots & \dots & \dots & \dots & \dots & p_{KK} \end{bmatrix}. \tag{9}$$

We note here that any number of hierarchical levels can be accounted for by the computation in this way. The normalization and preferential probability computation steps described in the previous section remain unchanged.

[1] An alternative approach would be to normalise the transition matrix at this step such that the axioms of probability are satisfied. We have experimented with this approach, and found that it results in the loss of numerical distinction between the preferential probabilities over time. It is thus numerically advantageous not to scale the transition matrix at each iteration, but rather to rescale the preferential probabilities in the final step, Eq. 8.

3 Language Model

In this section, we discuss how to compute a single scalar number for the subjective utility of an alternative from linguistic data. We use these values to estimate the transition probabilities, from which we calculate the change in preferential probability at each time step. Our perspective is that subjective preference about alternatives is developed through an engineer's positive or negative evaluations of alternatives. These evaluations can be explicitly revealed through the language of appraisal [3,4], semantic and grammatical forms of language for expressing judgments. For brevity, we refer to these as linguistic appraisals. The idea behind the use of linguistic appraisals as the basis for the elicitation of preference and uncertainty is that, intuitively, if a person expresses a linguistic appraisal such as, "Alternative 1 is a really good idea," then it is reasonable to predict that there is now a higher likelihood that alternative 1 is becoming more preferred relative to the others. Thus, as the committee discusses the alternatives, the linguistic data could provide time-varying information about how the preference for alternative 1 changes in terms of the degree and direction of the change, depending upon whether a positive or negative appraisal of alternative 1 is provided. As a person interacts with the alternatives and with others in the committee and obtains more information, preferences toward those alternatives would continually change. The analysis of linguistic appraisals would provide us a way to estimate the state transition probability for a given alternative.

First, we identify the semantic resources that can express a linguistic appraisal. Semantic resources are ways of expressing meaning through language. In Systemic-Functional Linguistics, there are five semantic resources for expressing a linguistic appraisal in the system of APPRAISAL [22]: Attitude; Engagement; Graduation; Polarity; and Orientation. The resource of Attitude has to do with making evaluations, such as whether something is 'good' or 'bad', 'right' or 'wrong'. Engagement is a scale for the speaker's commitment to the evaluation, such as 'sort of believe' to 'truly believe'. Graduation deals with the strength of evaluation, such as 'very good product' or 'extremely bad product'. Both Polarity and Orientation relate to whether the appraisal is positive or negative. The resources of Attitude, Engagement and Graduation are gradable resources (i.e., have a nominal scale) for evaluating alternatives, and these three will figure into calculating the appraisal value a. Orientation and Polarity are accounted for by the positive or negative sign of the appraisal value a. The process for identifying these semantic resources in linguistic appraisals in design are detailed by Dong et al., and has a high degree of objectivity and reliability since the coding strictly follows rules of grammar [4].

We group the words utilized in each linguistic appraisal into appraisal groups [37]. Whitelaw [37] applies 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 formed by categorizing the words utilised in each linguistic appraisal by the semantic resource of APPRAISAL. Each semantic resource could have gradable values of low, medium, and high, as shown in Figure 2.

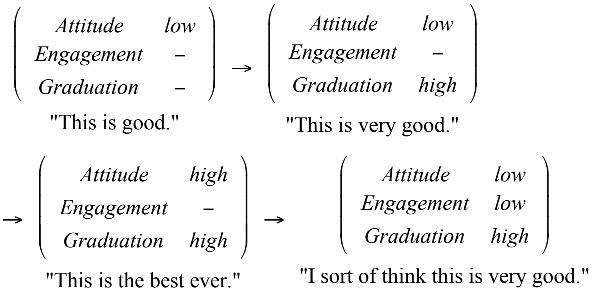

Fig. 2 Use of semantic resources in various linguistic appraisals and their gradable values

In other words, each semantic resource, Attitude (A), Engagement (E), and Graduation (G) could have variable intensities. Given the flexibility of natural language, the intensities could have a continuous, ordinal scale (Attitude: good, better, best, best ever; Engagement: sort of think, really think, truly believe; Graduation: kind of, quite, very, extremely). For the purposes of this research, we discretize this continuous scale and choose three nominal scales: high, medium, and low. Thus, we have the following model. The engineers' discussion is recorded in a transcript. We extract all the utterances that directly evaluate alternatives over a set of attributes. These utterances are ordered in time (dimensionless time steps). Each utterance has an Attitude and may have Engagement and Graduation semantic resources. Each semantic resource has one of 3 possible gradation values: high, medium, or low. Note that the language model for the appraisals and semantic resources is typical to all of natural language and not just design appraisals. The language model is based on the linguistic realization of appraisals of design alternatives in the most general form, independent of the specificities of a particular design domain or problem.

Given the three gradable values for the semantic resources of Engagement and Graduation, with the possibility that neither of these semantic resources is in use, and 3 gradable values for the semantic resource of Attitude, which must always be used, there are a total of 27 (= 3 × 3 × 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". The canonical appraisals are shown in Table 1. Each of these statements has a different level of intensity of judgment as well as the uncertainty toward the judgment. This concept of intensity of judgment is similar to Subasic's concept of intensity [29], but we do not attempt to assign a numerical intensity to each semantic resource. Rather, we map the gradable values for the semantic resources of appraisal into a scalar for the intensity of the entire appraisal and the uncertainty associated with the linguistic appraisal. Based on this mapping, it becomes possible to estimate the appraisal value for any arbitrary linguistic appraisal since a linguistic appraisal can be broken down into its constituent semantic resources and the gradable values per semantic resource.

The next step is to obtain broad consensus on the intensity and equivocality of the canonical linguistic appraisals listed in Table 1. We obtained this data through crowdsourcing using the online service Amazon Mechanical Turk. Amazon Mechanical Turk connects 'requesters' to 'workers' to complete simple tasks for pay, such as translating a short and ambiguous sentence or distinguishing between the colors ochre and rust. Mechanical Turk is increasingly being used for social science research [23] and for labelling data for machine learning [28]. 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, as described in Table 1. 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 degree intensity, and rejected results from workers who reversed the order of intensity 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. In total, we asked each worker to rate one set of 13 appraisals, which includes 9 from the canonical set, 1 crossover, and 3 quality control appraisals, from 1-very weak to 5-very strong with the midpoint being 3-neither weak nor strong. The work-

Table 1 Canonical Forms of Linguistic Appraisals 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.

Table 2 Amazon Mechanical Turk workers data

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

ers were neither providing a preference for a given set of alternatives nor giving market feedback on a particular product; they were simply asked to consider the given linguistic appraisals in the context of the common use of language, as if they were commenting to a friend about a movie they had recently seen. In the instruction, they were asked to consider if a statement such as "I really think that this is much better" reflects a stronger or weaker judgment than "I sort of think that this is sort of good." We 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' on 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.

Three batches of statements were run through Amazon Mechanical Turk, with 100 valid responses taken from each batch. More than 100 workers participated in each batch, but about 33% provided faulty data, mostly by not completing the task. The batch statistics are shown in Table 2.

For each combination of $[A, E, G]$ we have one dependent variable, appraisal intensity score S. This is the format of the data provided by the Mechanical Turk experiments. Thus, from the Mechanical Turk experiments, we have a matrix of size 100×27, where each row shows how a single representative worker graded any of the 27 statements, and each column shows the variation in response for a particular appraisal. Summing each column and taking the average thus tells us a single scalar appraisal intensity score for each 3 tuple:

$$A_j = \sum_{i=1}^{100} S_{ij}/100 \tag{10}$$

where S_{ij} are the individual scores by each respondent.

Descriptive statistics for the 3 control statements are shown in Table 3. The appraisal intensity scores increased in line with the expected direction. We note also that the standard deviation for the appraisal intensity scores decreases as the appraisal intensity scores increases. This implies that there is a higher level of uncertainty for weaker appraisals, a result that we will also find in the results for the 27 canonical statements.

Table 3 Appraisal intensity scores 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 4. From our data, the lowest appraisal came out to be 2.22, and this corresponded to the "lowest" tuple value $[A = L, E = L, G = L]$ and the highest appraisal came out to be 4.81, and this corresponded to the "highest" tuple value $[A = H, E = H, G = H]$. This result verifies the nominal scaling of the appraisals, since the respondents were unaware of our semantically based linguistic coding scheme and only saw some of the 27 statements. The statements were generated in order of predicted intensity of appraisal within a set of 9 statements (Q1-Q9, Q10-Q18, and Q19-Q27), and are shown in this order in Table 4. 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 appraisal intensity score within each of these sets. There is a recurrent pattern of a drop in score 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 appraisal intensity scores 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 appraisal intensity scores.

To check the consistency of the Mechanical Turk workers across the various batches, we conducted non-parametric tests, due to the non-normality of the distribution and ordinal values for the appraisal values collected, to determine if the mean values reported by the Mechanical Turk workers were similar. Independent samples Mann-Whitney U Tests were calculated for the crossover statements to determine if there is a statistically significant difference in the mean appraisal intensity scores. Crossover statements were tested across two independent batches. There was no statistically significant difference for Q16 ($U = 4757.5, p = 0.520$), but a statistically significant difference for Q9 ($U =

Table 4 Appraisal intensity scores for canonical appraisals

	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

Table 5 Kolmogorov-Smirnov Test for similarity of distributions

Batch	Statistic	So-so	Good	Excellent
1 and 2	Kolmogorov-Smirnov Z	1.061	0.566	0.495
	Asymp. Sig (2-tailed)	0.211	0.906	0.967
1 and 3	Kolmogorov-Smirnov Z	0.919	0.141	1.131
	Asymp. Sig (2-tailed)	0.367	1.000	0.155
2 and 3	Kolmogorov-Smirnov Z	0.141	0.707	0.636
	Asymp. Sig (2-tailed)	1.000	0.699	0.813

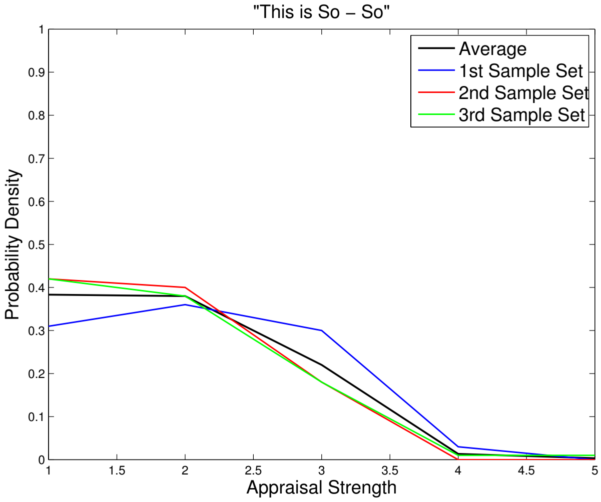

Fig. 3 Probability Distribution Function for "This is so-so."

$3662.0, p = 0.000$) and Q14 ($U = 4179.5, p = 0.032$). Similarly, we performed a Kruskal-Wallis H test to determine if there is a statistically significant difference in the means for the control statements. There was no statistically significant difference between the batches for the control statement "This is good" ($H(2) = 1.135, p = 0.568$), but there was a statistically significant difference in the means between the batches for the control statements "This is so-so" ($H(2) = 6.807, p = 0.033$) and "This is excellent" ($H(2) = 16.918, p = 0.000$).

Due to the statistically significant difference in the means, we compared the distributions of the results using the two-sample Kolmogovor-Smirnov test. For Q9, there is a statistically significant difference in the distributions (Kolmogorov-Smirnov Z = 1.980, $p = 0.001$), but no statistically significant difference in the distributions for Q14 (Kolmogorov-Smirnov Z = 0.849, $p = 0.468$) and Q16 (Kolmogorov-Smirnov Z = 0.495, $p = 0.967$). When comparing the distributions for the control statements, we find no statistically significant difference between batches as shown in Table 5.

Visually, the similarity of the distributions of the control statements is evident as is their skewness. As shown by the probability distribution functions for each of the control statements in Figures 3, 4, and 5, the distribution is left-skewed or right-skewed for the weakest and strongest appraisals, respectively, and more normally distributed for "This is good." This pattern is similarly reflected in the probability distribution functions for the canonical appraisals. The weakest appraisal, Q1, has a left-skewed distribution (Figure 6), a medium-strength appraisal, Q7, has a normally-distributed distribution (Figure 7), and a strong appraisal, Q27, has a right-skewed distribution (Figure 8). In summary, the data collected from the Mechanical Turk workers is sufficiently valid to calculate the appraisal values because the null hypothesis is satisfied for all of the control statements and for 2 of the 3 crossover statements to conclude that the samples were drawn from the same distribution.

To create an equally graded scale for the appraisal values $a = (0, 1]$, we normalized the 1×27 vector of raw appraisal intensity scores from Eq. (10) as per Eq. (11):

$$A_{normalized} = \frac{A_i - min(A)}{max((A - min(A))} \qquad (11)$$

For a positive appraisal, the calculated normalised appraisal values ranged from $0.135[L, L, L]$ to $0.991[H, H, H]$. For a negative appraisal, with no loss of generality, the corresponding values are the negative of these. To use these values in the calculation the transition proba-

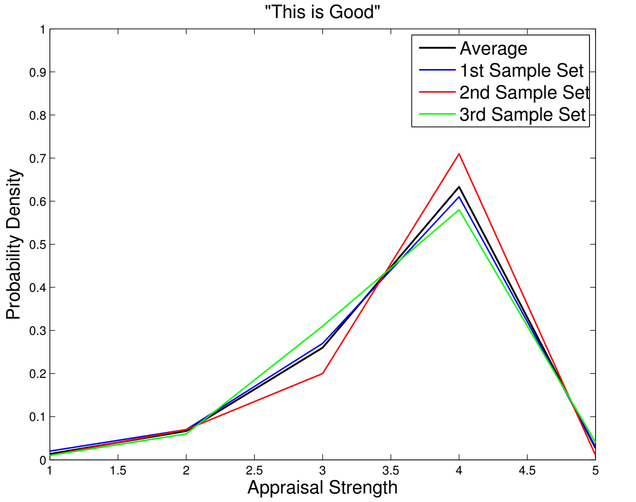

Fig. 4 Probability Distribution Function for "This is good."

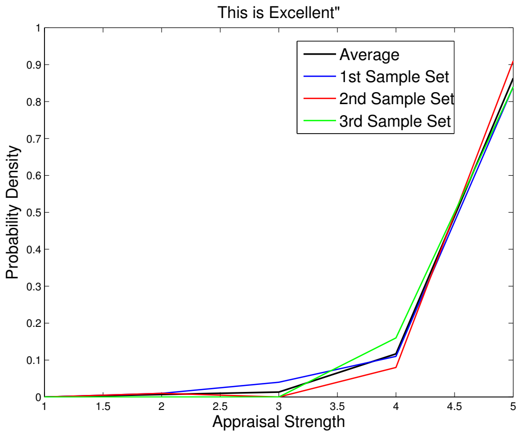

Fig. 5 Probability Distribution Function for "This is excellent."

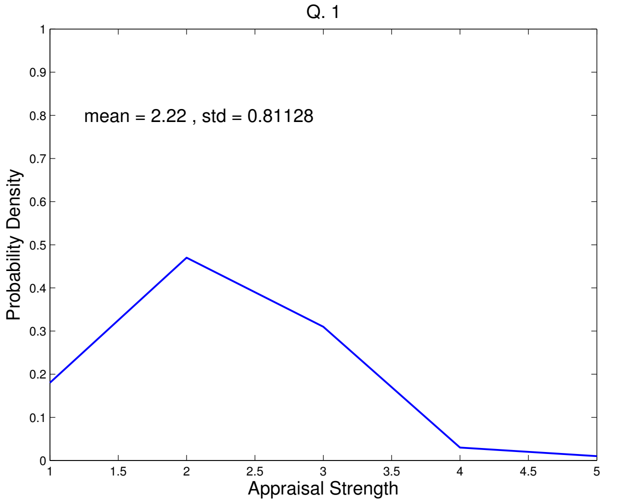

Fig. 6 Probability Distribution Function for "I sort of think that this is sort of good."

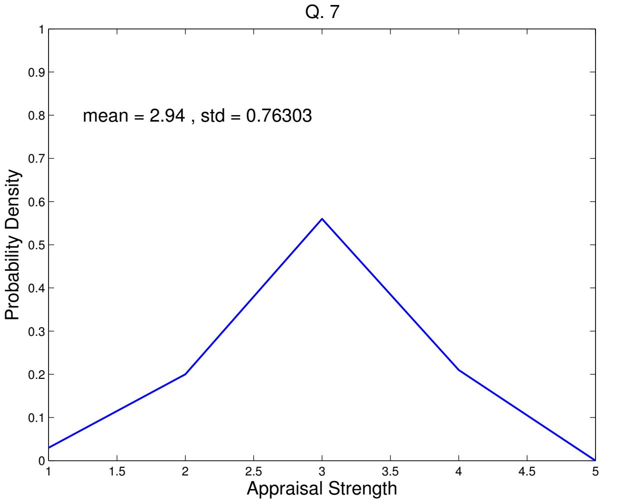

Fig. 7 Probability Distribution Function for "I really think that this is sort of good."

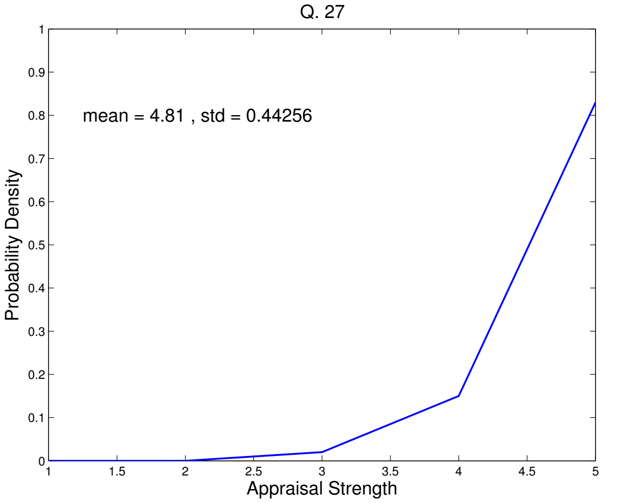

Fig. 8 Probability Distribution Function for "I really think that this is the very best."

bilities, we simply look up the appraisal values from $A_{normalized}$ for a given tuple of $[A, E, G]$.

We note here that in live design sessions, we observed that not all utterances involved appraisals using all three semantic resources. For example, a statement such as "So you're going to want to go with the titanium." leads to a code of $[A = L, -, -]$, since there is no Engagement or Graduation involved in the utterance. To cover the entire set of all combinations, including the cases where only one of the semantic resources A, E, or G are present or cases where two of these are present and one is absent, we perform a simple linear interpolation, and average over all possible values of the missing element(s). For example, consider the $[L, M, -]$ case: We take the combinations $[L, M, L] = 0.306931$, $[L, M, M] = 0.534653$, and $[L, M, H] = 0.594059$, and perform a simple average to give us the appraisal score for $[L, M, -] = (0.306931 + 0.534653 + 0.594059)/3 = 0.478548$.

4 Experimental Results

4.1 Time-Variant Preferential Probability for Alternatives

We first describe the results from an experiment in selecting a single alternative from a mutually exclusive set of alternatives. We have previously described this experiment in another paper [15], and, thus utilize this data set to compare approaches. The team's task was to choose a carafe of glass, plastic, or steel and filter of gold, paper, or titanium for a coffeemaker. 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 [10]. In order to encourage discussion among the group members and 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 linguistic appraisals in the data set were analyzed by AD and MCY. We developed the following notation to code the appraisals: $\mu_{\#}(d_{\#})(\alpha_{\#})(+, -)$ where $\mu_{\#}$ indicates the time step at which the appraisal was expressed, $d_{\#}$ indicates the alternative toward which the appraisal is directed, if any, $\alpha_{\#}$ indicates the attribute toward which the appraisal is directed, if any, and + or - indicate the direction (orientation) of the appraisal as positive or negative, respectively. Thus, the appraisal, "Glass coffee carafe seems to have the most capacity" would be coded as $\mu_{41}(d_1)(\alpha_7)+$ and utilizing the following semantic resources at the given level: Attitude = capacity (L), Engagement = seems (L), and Graduation = most (H). A total of 54 appraisals were coded. Seven statements of linguistic appraisals were used for training and arbitration purposes to ensure that the two coders could code the transcript consistently and reliably for the appearance of a linguistic appraisal and the correct categorization of a word by semantic resource for appraisal. AD checked each coder's work for consistency (e.g., not consistently coding the same word in as the same and correct type of semantic resource for appraisal) and made corrections where needed. A Krippendorf's alpha of 0.8188 for inter-coder reliability was achieved, which is considered acceptable [12].

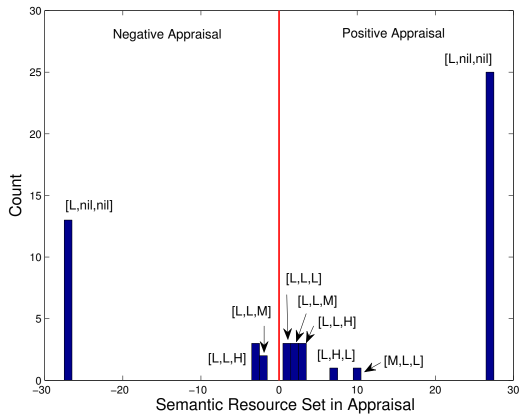

Fig. 9 Distribution of Semantic Resources in Positive and Negative Coffee Carafe Appraisals.

Only a few possible combinations of Attitude, Engagement, and Graduation occurred in the transcript as shown in Figure 9. Each of these appraisals has an underlying uncertainty distribution that has been captured by Mechanical Turk. Refer to Figure 3 for the distribution for an appraisal with low gradable values for each semantic resource. Some appraisals have more uncertainty than others, as shown in Tables 3 and 4, which compares the means and standard deviations for all 27 appraisals and 3 control statements sampled from Mechanical Turk. These results demonstrate the importance of considering the expressed appraisals in the

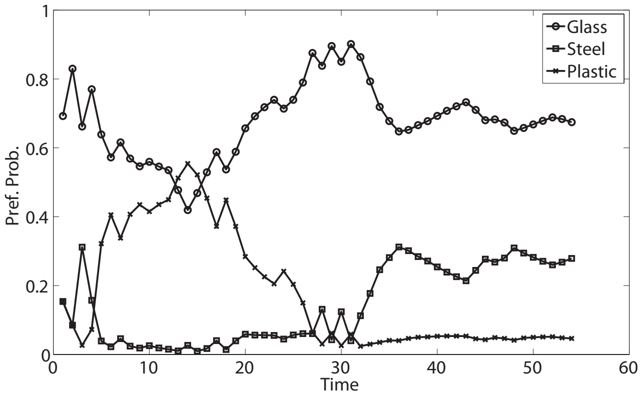

Fig. 10 Coffee Carafe Preferential Probability Time Variation

Table 6 Number of positive and negative utterances for the coffee carafe alternatives from design session transcript

	Positive	Negative
Glass	19	3
Steel	13	8
Plastic	4	7

calculation of the preferential probabilities and not just the occurrence of a positive or negative appraisal as we have done previously [15].

Figure 10 shows the results of calculating the time-variant preferential probabilities for the coffee carafe. It can be seen that the preferential probabilities fluctuate in time before they reach their final configuration. Specifically, note that the steel carafe alternative never has the highest preferential probability, but the plastic carafe is briefly likely to be the most preferred alternative. The glass carafe has the highest preferential probability at the end of the session. The calculations and ordering of preferential probabilities correspond with the actual choice indicated by the team in the survey. In terms of appraisal distributions, Table 6 shows the raw number of positive and negative utterances for each of the three alternatives. Again, the results on preferential probabilities correspond well with these, but note that merely using the distribution of the orientation (positive or negative) of utterances, it is not possible to analyze the fluctuations in preferences changing throughout the design session, something that is immediately obvious by observing the time-variation of the preferential probabilities.

4.2 Time-Variant Preferential Probability for Attributes

Section 4.1 described the results of analyzing linguistic appraisals of alternatives uttered by team members. However, we recognize that individuals could express preferences toward attributes rather than toward an alternative. For example, individuals may express a preference for lower cost over higher performance. Given the expression of this preference, rational behavior would require the decision maker to select the alternative having the best quality in relation to the preferred attribute, for example, the alternative having the lowest cost. To examine this phenomenon, we conducted the following experiment. In this exercise, three participants were asked to choose the configuration for a "high-end" and "low-end" bundle of laptops. Each configuration option is considered an alternative, as shown in Table 7, and the performance data are attributes, as shown in Table 8.

Table 7 Design alternatives for laptop configuration

Category	Alternative	Description
External Shell	d_1	Plastic alloy
	d_2	Magnesium alloy
	d_3	Titanium alloy
Screen size (inches)	d_4	11
	d_5	13
	d_6	15
	d_7	17
Display resolution	d_8	800x600
	d_9	1024x768
	d_{10}	1440x900
	d_{11}	1600x1200
	d_{12}	1400x1050
	d_{13}	1900x1200
CPU	d_{14}	i3
	d_{15}	i5
	d_{16}	i7
Disk drive	d_{17}	240GB
	d_{18}	320GB
	d_{19}	500GB
	d_{20}	720GB
	d_{21}	128SSD
	d_{22}	256SSD
Memory	d_{23}	2GB
	d_{24}	4GB
	d_{25}	6GB
	d_{26}	8GB
Battery	d_{27}	6Cell
	d_{28}	9Cell
	d_{29}	12Cell
Graphics board	d_{30}	Graphics3000
	d_{31}	FirePro

The configurations needed to satisfy the requirements for two different fictitious users. The first fic-

Table 8 Performance attributes for laptop configuration alternatives

Attribute	Description	Scale
α_1	cost (price)	Cost/benefit ratio
α_2	performance / speed	Any unit of time
α_3	display size	Diagonal dimension of display
α_4	display quality	Display resolution in pixels
α_5	battery life	Any unit of time
α_6	weight / portability	Any unit of weight
α_7	durability / strength / reliability	Any unit of strength
α_8	value	Any personal value such as convenience
α_9	disk storage space	Gigabytes
α_{10}	memory storage space	Megabytes

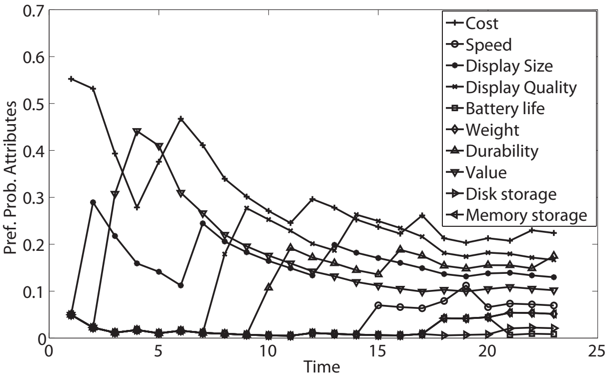

Fig. 11 Laptop Attributes Preferential Probability Time Variation. All the preferential probabilities at each individual time step sum to 1 (see Table 8).

titious user is a 7th grader is getting ready to start his last year of middle school and will be expected to bring a laptop every day and use it throughout the day in each of his classes. The second is a 30 year old photojournalist whose job requires that she keeps her laptop with her at all times, even while travelling to locations with limited access to electricity or trekking across rough terrain. To prevent any anchoring bias, the participants were not given any indicative bundle. Instead, they were provided a chart containing all of the configuration options, and the prices and performance data for each configuration option. This is a design configuration problem, with the participants having to configure two bundles, appraise each of the configuration options and the configured bundles, and set the importance of each of the attributes to the target users. The participants were given 60 minutes to complete the task, which was video-recorded and transcribed following the same procedure as the coffee carafe experiment. At the end of the experiment, the participants were asked to fill out a form individually in which they noted the individual options chosen for their final configuration.

Figure 11 shows the evolution of preferential probabilities for the attributes, using utterances from the transcript that focussed specifically on the attributes only. This graph describes changes in preferential probability toward attributes, reflecting statement such as, “So it looks like the user is going to be having really toward a low-end computer. His parents do not have money so low cost.” Clearly, in this utterance, the importance of the cost attribute is discussed in general terms without reference to any of the alternatives.

Figure 12 shows the evolution of preferential probabilities for the “low-end” bundle for the student. We do not show the results for the photojournalist as they are similar. For all 8 categories of options, the attribute with the highest preferential probability at the end of the session matched perfectly with the options chosen by the designers based upon the options reported in the survey taken at the end of the session. In this plot, the preferential probabilities at each individual time step for all 8 figures are normalized with respect to all 31 alternatives, and the initial preferential probability for all alternatives is initialized at 1/31 (≈ 0.032). A numerical artefact of this normalization is that preferential probabilities within a category may shift slightly even though no linguistic data was received for that category at a given time step. The shift, however, does not change any relative differences in preferential probability between options within a category or across categories. An alternative formulation could have each category described by its own transition matrix and set of preferential probabilities. However, such a formulation would not permit a comparison of the preferential probabilities across categories and relative to the preferential probabilities for attributes. For instance, since there are only 3 options for the Battery but 6 options for Disk drive, the initial preferential probabilities would be 1/3 and 1/6, respectively, and it is likely that the preferential probability for a Disk drive option would always be less than the preferential probability for a Battery option. Yet, as the results show in Figure 12, the preferential probability for the 500GB disk drive is higher than the preferential probability for the 9 cell battery, reflecting the cost differential and the importance of the cost attribute as shown in Figure 11. It costs an additional \$50 for the 500GB disk drive, but an additional \$100 for the 9 cell battery.

We note here that this transcript contained explicit discussions of preferences toward attributes and alternatives, whereas the analysis shown in Figure 12 is based upon linguistic appraisals toward alternatives only. One way to incorporate the explicit discussion of preference toward attributes into the calculation of the dynamically changing preferential probability for alternatives would be to increase the preferential probability for an alternative proportional to the (increasing or decreasing) preferential probability for an attribute, similar to the calculation performed in concept scoring when the weights on selection criteria change. However, this calculation would require prior knowledge about the quality of an alternative on all attributes. In other words, we would need to know *a priori* a rank ordering of the alternatives according to an attribute, which may not always be possible. Second, performing this calculation would mask potentially inconsistent behavior by the decision makers. We expect that when the committee expresses a preference toward an attribute, logically, they should also prefer the alternative that performs best on that attribute. That is, any attribute that is considered more important (or less important) by the committee should be reflected by a corresponding increase (or decrease) in the preferential probability for the associated alternatives. For example, in our analysis, cost and durability emerged as two of the top attributes, and the analysis showed that the preferential probability increased for alternatives performing well on these attributes. Therefore, in the preferential probability toward alternatives graph of Figure 12, the preferential probabilities are implicitly affected by attributes that emerged as being more important to the decision makers. If the situation is the opposite, that is, the decision makers express a higher preference for an attribute but then appraise an alternative higher when it actually performs worse (or worst) on that attribute, then we would have identified an inconsistency in their decision making. For these reasons, we do not explicitly incorporate the changing preferential probability for attributes directly into the calculation of the dynamically changing preferential probability for alternatives.

Instead, by reading the preferential probability graphs of the attributes and alternatives together, the influence of preferential probability for an attribute on the preferential probability for an alternative becomes visible, permitting a check of consistency or inconsistency in decision making. As explained previously, when individuals express an increased preferential probability for an attribute, they should, logically, choose the alternative having the best quality for this attribute. To demonstrate the influence, we discuss a part of the transcript. In the beginning of the design session, the transcript contains discussions on cost and display size. Figure 11 clearly shows this – in the first part of the preferential probability for attributes graph, it can be seen that cost and display size are the two dominant attributes discussed. Based on this, the first decisions finalized by the designers are the alternatives for the options for display resolution and the graphics card. In Figure 12, the preferential probability for alternatives graphs for display resolution and graphics card show that as a result of the increased preferential probability on the cost and display attributes, the 1024 × 768 resolution and the Graphics3000 card have clear dominance. As a second example, the designers had a concentrated discussion on durability. The point in Figure 11 when the preferential probability for the durability attribute starts to rise corresponds to the point in Figure 12 (Pref. Prob. External Shell) when the preferential probability for the titanium alternative becomes the most dominant. Later in the design session, since cost continues to dominate over all other attributes, with durability remaining the second most important attribute, but never dominating cost, the preferential probability for magnesium dominates over titanium, because it is seen to perform better than titanium on the cost attribute.

5 Discussion and Conclusion

We have presented a method to estimate preferential probabilities in concept selection from a mutually exclusive set by a committee such as a small design team or a design review panel. We have shown that preferential probabilities are not stable during committee discussions of design alternatives, and the intensity and uncertainty in the linguistic appraisals changes the relative ordering of preferential probabilities. While the case studies were limited to synchronous discussion, we believe that the reported method can be extended to asynchronous discussions over a longer time scale wherein alternative concepts are continually being developed and refined.

The analysis illustrates the dynamics of preference change that no other techniques (such as pairwise comparisons or analytic hierarchy process) can reveal, because they only focus on the final decision and not the conversation and changing attitudes toward alternatives as reflected in the conversation. First, referring to the graphs in Figure 12, it is clear which of the alternatives attracted debate and discussion before reaching a final decision as opposed to which alternatives were presumably clear from the start of the design session with no changes in preferential probability occurring over the entire design session. For example, while there was significant discussion on the choice of magnesium, titanium, and plastic as the external shell, there was almost no debate on the choice between the Graphics3000 versus the FirePro graphics card. It can be observed in sub-figures (Pref. Prob. External Shell) and (Pref. Prob. External Disk Drive) of Figure 12 that some of

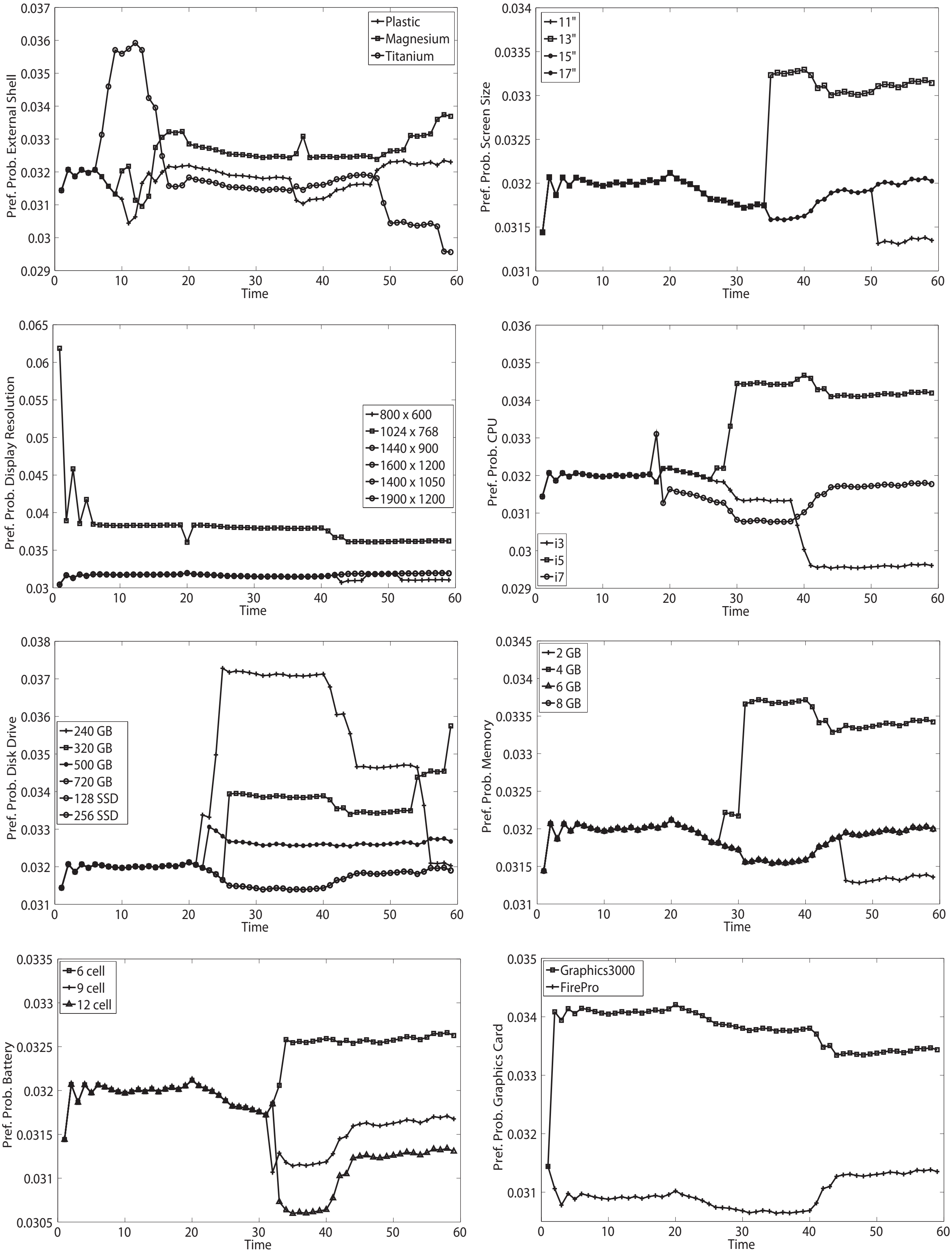

Fig. 12 Laptop Example Preferential Probability Time Variation. We show the preferential probability time-variant dynamics as 8 different plots to enable clarity of presentation, grouped by category type (see Table 7).

the alternatives (such as the titanium shell or the 240 GB disk drive) that came out as having the lowest preferential probability actually had, in interim parts of the session, the highest preferential probability. This has implications for management and decision making in complex engineering design. If we knew which components of an engineered product attracted more scrutiny through discussion, then one can devote more formal review resources to those components. Likewise, decision makers may wish to review those decisions apparently made "too" quickly to ascertain whether the choices are defensible.

A second interesting management and decision making implication for complex engineering design is that this analysis can help to uncover potential inconsistencies in decision making. In the experimental result presented above, we observed that for all 8 options, the final option with the highest preferential probability matched perfectly the options reported by the designers. In other words, the designers chose (actual decision making on alternatives) what they said they would (expressed preference toward alternatives). These experiments also served as a verification for the method; the method predicts preferential probabilities that are consistent with what was decided and linguistically stated in the design sessions. However, for larger, more complex design problems, the method can serve as a tool to detect inconsistencies in design decision making. Recall that the method is based on a linguistic model for how English speakers express evaluations and that the numerical data for appraisal values are based on a broad range of English speakers' interpretations of the strength of certain linguistic appraisals. Inconsistencies between the model's predicted decision and the committee's actual decisions is tantamount to inconsistencies between what the committee stated and how an English speaker would interpret their appraisals in the absence of any other contextual data. If the engineers' final preferences (as stated separately) do not match with the results of the final preferential probabilities, then it means that the engineers' expressed preferences evolving in the design session did not match their final actual preference. This can help to identify inconsistencies in design decision making, but not the cause of the inconsistency.

It is important to emphasize that the method provides a descriptive model of decision-making, not a normative one. The method neither directs the committee to make a specific choice nor assists the committee to make a utility-maximizing decision. We instead took a pragmatic stance toward methods for concept selection [25]. Engineers (engineering firms) may either have their own formal method(s) for concept selection or no formal method for concept selection other than discussion. If they have a formal method, the linguistic analysis provides additional support for their decision, as well as a means to interrogate the equivocality of the preference. If they have no formal method, then the linguistic analysis lends a degree of analysis to the decision and preceding discussion on preferences. Regardless of the situation, the method provides for process tracing on the dynamics of change in preferences.

We believe that identifying discrepancies between what a committee decides and what a committee says, literally, they will choose is the most valuable contribution that this work could make to decision-based design. This type of description of decision making provides a quality control tool for decisions [17], especially when decision makers do not formally model their decisions. In situations wherein decisions are only talked about but not modelled, perhaps due to the complexity of the decision, there is nonetheless the expectation that the individuals used disciplined thought to guide their formation of subjective preferences and that the committee deliberated vigorously. 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.

It is important to note the limitations and assumptions associated with this linguistically-based method for describing decision making:

1. The calculation of preferential probabilities assumes mutually exclusive alternatives. While it is possible for a team to have and to state a joint preference ("I think option A and option B are equally good") and for a team to have a joint distribution across two or more alternatives in separate alternative classes, we have not yet encountered linguistic evidence of these possibilities. To handle this situation, we would recommend treating such an utterance as two separate time steps. The analyst would add $+a$ to option A in one time step and then add $+a$ to option B in the next time step, resulting in their having similar preferential probabilities. As results in Figure 12 show, our model reveals the situation when alternatives have similar preferential probabilities or when preferential probabilities for alternatives across two or more alternative classes have similar preferential probabilities. In such instances, we might conjecture that these alternatives have an implied joint distribution, but this distribution is not possible to calculate from the available linguistic data.
2. The model assumes that when an individual positively appraises an alternative, then the transition probability for the appraised alternative increases by the appraisal value and the transition probabilities for the other alternatives must necessarily decrease by a proportional amount (and vice-versa for a negative appraisal). If individuals actually have equal preference for alternatives, or a joint distribution for two more more alternatives, then it should be the case that they positively appraise the alternatives in roughly equal number and strength of evaluative stance over time. In this case, the model would show that two or more alternatives have similar preferential probabilities over time with the potential that the final set of preferential probabilities is simply $\frac{1}{N}$.
3. The cases presented were analyzed *post hoc*. The number of alternatives and attributes were known in advance and prescribed by the experimental conditions. The method itself does not require this *a priori* knowledge, though. The number of rows and columns in Eq. 5 can be dynamically modified based on the total number of alternatives N at any given moment in time. The subsequent normalisation of the transition probabilities should still enforce the rule that the final preferential probabilities in Eq. 7 must sum to 1.
4. At the start of the model, the set of preferential probabilities and transition probabilities is initialized at $\frac{1}{N}$ in the absence of any prior knowledge. This is not a requirement of the method. If decision makers have prior knowledge, the initial values can be initialized accordingly as long as the values satisfy the axioms of probability.

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 under the view that preferences are not 'fixed' in the mind of the decision-maker but are subject to change due to discussion, negotiation, further knowledge, and interaction with each alternative. The linguistic analysis of appraisals and their conversion into a preferential probability provides a complement to formal methods for concept selection, because this method provides an objective way to peer into the details of the conversation that led toward the selection of an alternative. By doing so, we may be able to move the debate [25] about methods for concept selection beyond which method definitively chooses the most socially optimal alternative toward equally important questions about behavior and framing effects in effect during the decision making process, both of which can negatively counteract formal methods. In our experience consulting with industry, decisions about alternatives are rarely clear and crisp; preferences are subject to engineering expertise and intuition. The quality and rigor around the discussion may be more important in making the right selection than proper application of the formal method [9].

While this paper makes no claim to the cognition of decision-making, the possibility of formally describing 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. Linguistic data should not be a complete substitute for preference data particularly when preference data for design trade-offs can be collected through techniques such as a lottery method. The nuances of expressing linguistic appraisals in English provide a window into the dynamics of preference formation and change in concept selection. While we have limited our analysis to in-person discussions, this research sets the foundation for the analysis of other forms of language-based communication between design teams as they take decisions. An analysis of these communication, as we have done for e-mail [35,34], could provide a systemic view on decision-making in large-scale, complex projects. Finally, we believe that the method could 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, and this would be a fruitful application of the method.

Acknowledgements The authors wish to thank the participation of the engineering students in the experiments.

References

1. Arrow, K.J.: Social choice and individual values, 2nd edn. Yale University Press, New Haven (1963)
2. Delbecq, A.L., Mills, P.K.: Managerial practices that enhance innovation. Organizational Dynamics **14**(1), 24–34 (1985)
3. Dong, A.: The language of design: theory and computation. Springer, London (2009)
4. Dong, A., Kleinsmann, M., Valkenburg, R.: Affect-in-cognition through the language of appraisals. Design Studies **30**(2), 138–153 (2009)
5. Dym, C.L., Wood, W.H., Scott, M.J.: Rank ordering engineering designs: pairwise comparison charts and borda counts. Research in Engineering Design **13**(4), 236–242 (2002)
6. Frey, D., Herder, P., Wijnia, Y., Subrahmanian, E., Katsikopoulos, K., Clausing, D.: The pugh controlled convergence method: model-based evaluation and implications for design theory. Research in Engineering Design **20**(1), 41–58 (2009)
7. Frey, D., Herder, P., Wijnia, Y., Subrahmanian, E., Katsikopoulos, K., de Neufville, R., Oye, K., Clausing, D.: Research in engineering design: the role of mathematical theory and empirical evidence. Research in Engineering Design **21**(3), 145–151 (2010)
8. Frishammar, J., Floren, H., Wincent, J.: Beyond managing uncertainty: Insights from studying equivocality in the fuzzy front end of product and process innovation projects. Engineering Management, IEEE Transactions on **58**(3), 551–563 (2011)
9. Garbuio, M., Lovallo, D.: The under-appreciated role of quality conversations in strategic decision-making. In: 71st Annual Meeting of the Academy of Management AoM 2011. Academy of Management (2011)
10. Gigone, D., Hastie, R.: The impact of information on small group choice. Journal of Personality and Social Psychology **72**(1), 132–140 (1997)
11. Gilovich, T., Griffin, D.W., Kahneman, D.: Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, New York (2002)
12. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures **1**(1), 77 – 89 (2007)
13. Hazelrigg, G.: A framework for decision-based design. ASME Journal of Mechanical Design **120**(4), 653–658 (1998)
14. Hazelrigg, G.: The pugh controlled convergence method: model-based evaluation and implications for design theory. Research in Engineering Design **21**(3), 143–144 (2010)
15. Ji, H., Yang, M., Honda, T.: An approach to the extraction of preference-related information from design team language. Research in Engineering Design **23**(2), 85–103 (2012)
16. Kahneman, D., Lovallo, D.: Timid choices and bold forecasts: A cognitive perspective on risk taking. Management Science **39**(1), 17–31 (1993)
17. Kahneman, D., Lovallo, D., Sibony, O.: Before you make that big decision... Harvard Business Review **89**(6), 50–60 (2011)
18. Kim, J., Wilemon, D.: Focusing the fuzzy frontend in new product development. R&D Management **32**(4), 269–279 (2002)
19. LeDantec, C.A., Do, E.Y.L.: The mechanisms of value transfer in design meetings. Design Studies **30**(2), 119 – 137 (2009)
20. Lewis, K.E., Chen, W., Schmidt, L.C.: Decision Making in Engineering Design. ASME Press, New York (2006)
21. López-Mesa, B., Bylund, N.: A study of the use of concept selection methods from inside a company. Research in Engineering Design **22**(1), 7–27 (2011)
22. Martin, J.R., White, P.R.R.: The Language of Evaluation: Appraisal in English. Palgrave Macmillan, New York (2005)
23. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on amazon mechanical turk. Judgment and Decision Making **5**(5), 411–419 (2010)
24. Pugh, S.: Total Design: Integrated Methods for Successful Product Engineering. Addison-Wesley (1991)
25. Reich, Y.: My method is better! Research in Engineering Design **21**(3), 137–142 (2010)
26. Scott, M.J., Antonsson, E.K.: Aggregation functions for engineering design trade-offs. Fuzzy Sets and Systems **99**(3), 253–264 (1998)
27. Scott, M.J., Antonsson, E.K.: Arrow's theorem and engineering design decision making. Research in Engineering Design **11**(4), 218–228 (1999)
28. Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on, pp. 1 –8 (2008). DOI 10.1109/CVPRW.2008.4562953
29. Subasic, P., Huettner, A.: Affect analysis of text using fuzzy semantic typing. IEEE Transactions on Fuzzy Systems **9**(4), 483–496 (2001)
30. Thaler, R.H., Sunstein, C.R.: Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven (2008)
31. Thurston, D.L.: A formal method for subjective design evaluation with multiple attributes. Research in Engineering Design **3**, 105–122 (1991)
32. Thurston, D.L.: Real and misconceived limitations to decision based design with utility analysis. Journal of Mechanical Design **123**(2), 176–182 (2001)
33. Ulrich, K.T., Eppinger, S.D.: Product Design and Development, 3 edn. McGraw-Hill/Irwin, New York (2004)
34. Wasiak, J., Hicks, B.J., Newnes, L., Dong, A: Understanding engineering email: the development of a taxonomy for identifying and classifying engineering work. Research in Engineering Design **21**(1), 43–64 (2010)
35. Wasiak, J., Hicks, B.J., Newnes, L., Loftus, C., Dong, A., Burrow, L.: Managing by e-mail: What e-mail can do for engineering project management. IEEE Transactions on Engineering Management **58**(3), 445–456 (2011)
36. Weick, K.E.: Sensemaking in organizations. Sage Publications, Thousand Oaks (1995)
37. Whitelaw, C., Garg, N., Argamon, S.: Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM '05, pp. 625–631. ACM, New York, NY, USA (2005). DOI 10.1145/1099554.1099714