Journal of the Operational Research Society (2011) 62, 1638–1654

www.palgrave-journals.com/jors/

Multicriteria evidential reasoning decision modelling and analysis—prioritizing voices of customer

J-B Yang[1*], D-L Xu[1], X Xie[1,2] and AK Maddulapalli[3]

[1]*The University of Manchester, Manchester, UK;* [2]*Dalian Maritime University, Dalian, China; and* [3]*India Science Lab, GM R&D, Bangalore, India*

In this paper, a new methodology is investigated to support the prioritization of the voices of customers through various customer satisfaction surveys. This new methodology consists of two key components: an innovative evidence-driven decision modelling framework for representing and transforming large amounts of data sets and a generic reasoning-based decision support process for aggregating evidence to prioritize the voices of customer on the basis of the Evidential Reasoning (ER) approach. Methods and frameworks for data collection and representation via multiple customer satisfaction surveys were examined first and the distinctive features of quantitative and qualitative survey data are analysed. Several novel yet natural and pragmatic rule-based functions are then proposed to transform survey data systematically and consistently from different measurement scales to a common scale, with the original features and profiles of the data preserved in the transformation process. These new transformation functions are proposed to mimic expert judgement processes and designed to be sufficiently flexible and rigorous so that expert judgements and domain specific knowledge can be taken into account naturally, systematically and consistently in the transformation process. The ER approach is used for synthesizing quantitative and qualitative data under uncertainty that can be caused due to missing data and ambiguous survey questions. A new generic method is also proposed for ranking the voices of customer based on qualitative measurement scales without having to quantify assessment grades to fixed numerical values. A case study is examined using an Intelligent Decision System (IDS) to illustrate the application of the decision modelling framework and decision support process for prioritizing the voices of customer for a world-leading car manufacturer.

Journal of the Operational Research Society (2011) **62,** 1638–1654. doi:10.1057/jors.2010.118
Published online 1 September 2010

Keywords: evidential reasoning; multiple criteria decision analysis; voices of customer; ranking

Introduction

World markets are becoming increasingly competitive and dynamic, driven by rapid changes in customer demands. Rapid and accurate identification of customer demands, systematic assessment of product quality and genuine customer-driven new product development are essential to success. Many firms worldwide have focused on total quality management (Griffin and Hauser, 1993; Chan and Wu, 2002, 2005). One aspect of the focus on total quality management has been the widespread adoption of Quality Function Deployment (QFD). QFD is a cross-functional planning methodology, commonly used to ensure that customer needs (referred to as the voices of customer) are deployed through product planning, part development, process planning and production planning. Through the appropriate implementation of QFD, a firm can improve engineering knowledge, productivity and quality and reduce costs, product development time and engineering changes (Besterfield *et al*, 2003). QFD has now become a standard practice by many leading firms and has been successfully implemented worldwide (Chan and Wu, 2002).

The successful implementation of QFD depends on the identification and prioritization of the voices of customer. A customer voice is a description, stated in the customer's own words, of the benefit to be fulfilled by a product or service. Identifying the voices of customer is primarily a qualitative research process. Many methods such as one-to-one interviews with customers or customer focus groups can be used to find customer needs, which can be classified into strategic, tactical and operational needs. Griffin and Hauser (1993) reported the empirical studies on the identification of the voices of customers using various methods. In this paper it is assumed that a list of

*Correspondence: J-B Yang, Manchester Business School, The University of Manchester, Manchester M15 6PB, UK.
E-mail: jian-bo.yang@mbs.ac.uk

customer voices has already been created and it focuses on how to prioritize the voices of customers.

The prioritization of the voices of customer has attracted much attention in the research community and a number of prioritization methods have been suggested in the literature. For example, the analytic hierarchy process (AHP) (Saaty, 1994; Kwong and Bai, 2002), the entropy method (Chan *et al*, 1999; Chan and Wu, 2005), the weighted sum method (Chen and Weng, 2003), the fuzzy weighted average method (Chen and Weng, 2006), the reasoning based methods (Temponi *et al*, 1999; Chen *et al*, 2004), and the preference disaggregation method—MUSA method (Grigoroudis and Siskos, 2002). Since customer needs and preferences can be expressed in different formats, such as numerical or linguistic, research has been conducted to aggregate customer voices of different formats. For example, Büyüközkan and Feyzioğlu (2005) presented a fuzzy logic based group decision making approach, which allows multiple expression formats in QFD. Ho *et al* (1999) investigated group behaviours in QFD and reported an integrated group decision-making approach for aggregating team members' opinions to prioritize customer needs. In these studies, it is assumed that customer needs can always be expressed using the specific frameworks as designed in these methods. However, this may not be the case if multiple data sources are used for prioritizing the voices of customer, because the analysts may not always be able to control the design of data formats, in particular if data are generated from external sources. A question arises as to how to prioritize the voices of customer using data of various formats generated from different data sources without losing or distorting original information.

In the manufacturing and service industries, it is a common practice for a firm to conduct customer satisfaction surveys as a means to identify what its customers want, what are the strengths of its products and services and where improvements should be made. Although customers may be invited to directly rate the importance of survey statements in a survey, they are more often asked to assess the performances of a firm's products or services as a stand-alone or in comparison with its competitors' products or services. As a matter of fact, data from different surveys are often used to identify those voices of customer that should be given higher priority (sometimes referred to as Key Voices of Customer) within the context and constraints of the overall product or service program.

Since most data generated from different customer satisfaction surveys are the performance assessments of a firm's and its competitors' products or services measured on qualitative scales, such as 'satisfied', 'neutral' and 'dissatisfied', they may not be in a format that can be directly used to prioritize the voices of customer. As such, they need be transformed to a common scale in an appropriate format before they can serve as evidence to support voice prioritization. Such data transformation is domain specific and task dependent, so it requires expert judgments and domain specific knowledge. Furthermore, due to the qualitative nature of performance assessments measured on qualitative scales, it is a significant challenge to preserve the original features and profiles of data during the transformation. This means that the transformed data should still be of a qualitative nature, maintain the profiles of both good and bad performances, have little or no loss of information and also keep any uncertainty intact.

In this paper, the qualitative features of different surveys, both internal and external, are investigated first. A novel belief structure is suggested to provide a generic framework for representing survey data measured on qualitative scales, which can preserve rich information generated from surveys and provides scope for designing survey questionnaires in flexible ways. Based on the general principles proposed for transformation of raw data to decision information (Yang, 2001) and the extensive discussions and consultation with a number of senior researchers, engineers and managers of a world leading car manufacturer, including several face-to-face meetings and many web-based group meetings, several innovative yet natural and pragmatic rule-based transformation methods and functions are proposed to capture domain specific knowledge and mimic expert judgement processes for mapping original survey data of a qualitative nature measured on various scales to equivalent distribution assessments measured on a common scale. Based on the extensive accumulated experiences and practices of the company as well as the previous research conducted by the authors, new criteria created from transformed survey data for assessing the voices of customer are established and investigated so that the problem of prioritizing the voices of customer can be formulated as a generic multiple criteria decision analysis (MCDA) problem, which, however, is characterized by a belief decision matrix (Yang and Xu, 2002a) rather than a conventional decision matrix (Keeney and Raiffa, 1993; Winston, 1994). The major difference between the two MCDA modelling frameworks is that the former employs a belief structure or a distributed assessment to assess the profile of a voice on a criterion so that rich evidence can be captured and preserved for further decision analysis, while the latter only uses a single average number that can only capture the mean or expected value but not the profile of an assessment, thereby losing useful information for further decision analysis. In this paper, the former is referred to as evidence-driven decision modelling to distinguish it from the conventional mean value-based decision modelling.

To prioritize the voices of customer, evidence generated by transforming survey data to a common scale needs to be aggregated. The Evidential Reasoning (ER) approach (Yang and Sen, 1994; Yang and Singh, 1994; Yang, 2001; Yang and Xu, 2002a, b; Yang *et al*, 2006; Xu *et al*, 2006)

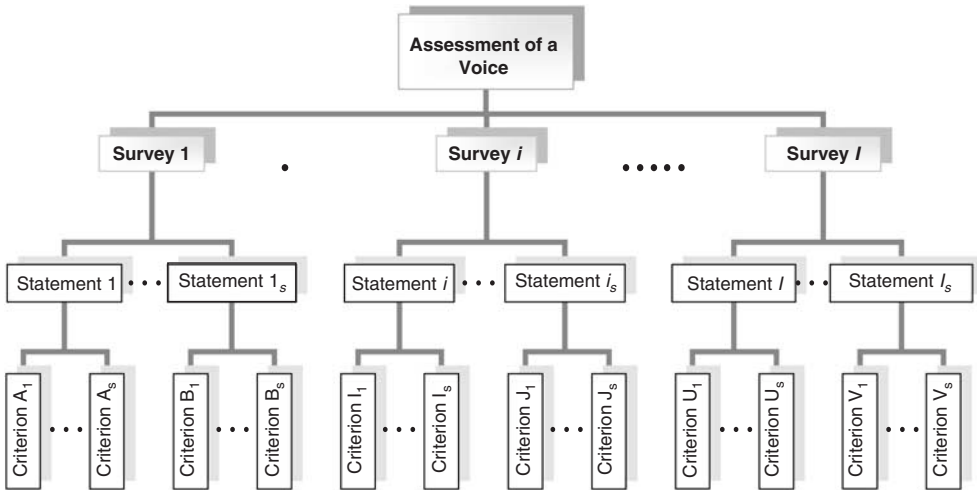

Figure 1 A generic framework for assessing the voices of customer.

provides a unique reasoning-based process for combining distributed assessments. It has been applied to many areas such as design and product assessment (Yang and Xu, 1998; Chin *et al*, 2009). In this paper, the ER approach is employed for prioritizing the voices of customer. Prioritizing voices requires eventually ranking them on individual criteria or on the overall criterion. A simple approach for ranking voices assessed on a qualitative scale is to quantify each assessment grade of the scale to a certain fixed value, calculate the mean value of a qualitative assessment and then rank voices based on their mean values. However, such an approach can only generate a rank order in a narrow sense of mean value but cannot consider the richer information of an assessment. In this paper, a generic ranking method is proposed as part of the reasoning-based decision support process for prioritizing the voices of customer. This method does not require the assessment grades to be quantified to fixed values but can allow them to take any values that suit their qualitative definitions and meanings.

The evidence-driven decision modelling framework and the reasoning-based decision support process have been applied to the prioritization of the voices of customer for a world-leading car manufacturer. Although it is not possible to report the application in great detail due to its commercial confidentiality and large scale, a smaller yet relevant case study is reported in this paper to illustrate the principle and procedure of implementing the proposed methodology and its potential for a wider range of applications. The intelligent decision support system (IDS) developed on the basis of the ER approach (Yang and Xu, 2005) is used to support the case study. Many of the analysis results reported in this paper are generated using IDS.

In the rest of this paper, the process and features for collecting evidence through internal and external customer satisfaction surveys are first examined. Several typical rule-based transformation functions are then analysed. In the following section, the ER approach is briefly outlined first and a new generic ranking method is then proposed. Finally a case study is reported and the paper is concluded with discussions.

Evidence collection through customer satisfaction surveys

In Figure 1, a generic hierarchy is illustrated for voice assessment. In general, a voice may be assessed through one or multiple surveys but maybe not through all surveys. It is possible that a voice is assessed by more than one statement in a survey. From the survey data, multiple criteria can be created to generate appropriate evidence for assessing and prioritizing a voice. So, the assessment of a voice is in principle a multi-level multi-criteria decision analysis (MCDA) problem.

Internal surveys

In manufacturing and service industries, it is common for a firm to conduct costumer satisfactory surveys, either by inviting customers to evaluate new products in a firm clinic or by sending questionnaires to customers. In the car manufacturing industry, for example, a firm's market research team may select a group of its own vehicles and its competitors' vehicles for evaluation according to its needs. An internal survey is normally supported by an internally designed questionnaire, consisting of many statements, such as whether 'a vehicle's fuel economy is satisfactory'.

One advantage of conducting an internal survey is that a firm can design a questionnaire according to its needs and control the process and scope of the survey. This means that in its internal survey a firm can define a comprehensive

set of statements (or voices of customer). This way, evidence collected from an internal survey can be directly used to support the prioritization of the voices of customer for future new product development programmes.

In an internal survey, customers are usually asked to evaluate products on each statement. Evaluations provided by customers are subjective in nature and may be given on the basis of an ordinal scale having several assessment grades. For instance, the following five grade ordinal scale may be used:

$$H_1 = \{\text{'}H_{1,1}\text{—}\textit{Dissatisfied Completely}\text{'}, \text{'}H_{1,2}\text{—}\textit{Dissatisfied}\text{'}, \text{'}H_{1,3}\text{—}\textit{Neutral}\text{'}, \text{'}H_{1,4}\text{—}\textit{Satisfied}\text{'}, \text{'}H_{1,5}\text{—}\textit{Satisfied Completely}\text{'}\} \quad (1)$$

For assessing a product, a customer may choose to tick one of the grades on a statement (voice). If a group of K customers participate in a survey and $k_{l,n,}$ of them tick a grade $H_{1,n}$ for assessing a product on a voice A_l, then the frequency or the degree of belief $\beta^l_{1,n}$ to which a product is assessed by the whole group of customers to the grade $H_{1,n}$ on the voice A_l is given as follows:

$$\beta^l_{1,n} = \frac{k_{l,n}}{K} \quad (2)$$

The evaluation rating of a product on the voice A_l by the whole group of customers can then be represented by the following distribution:

$$S(A_l) = \{(H_{1,1}, \beta^l_{1,1}), (H_{1,2}, \beta^l_{1,2}), (H_{1,3}, \beta^l_{1,3}), (H_{1,4}, \beta^l_{1,4}), (H_{1,5}, \beta^l_{1,5}), (H_1, \beta^l_{H_1})\} \quad (3)$$

In (3), $0 \leqslant \beta^l_{1,n} \leqslant 1$; $\sum_{n=1}^{5} \beta^l_{1,n} \leqslant 1$ and $\beta^l_{H_1} = 1 - \sum_{n=1}^{5} \beta^l_{1,n}$ which measures the percentage of customers who do not provide any assessment of a product on the voice A_l. In other words, $\beta^l_{H_1}$ represents the amount of missing information or the degree of ignorance of the customer group in assessing the voice A_l. This distributed evaluation (3) adequately records the collected assessment information in its original format, preserves the diversity of customer views and thus provides rich evidence for further decision analysis.

Following a survey, it is common to quantify assessment grades and then generate a mean evaluation rating as an indicator for a distributed assessment. Let $u(H_{1,n})$ be the score (value or utility) assigned to $H_{1,n}$. If $\beta^l_{H1} = 0$, the mean value of a distribution shown in (3) is then given by

$$u(S(A_l)) = \sum_{n=1}^{5} \beta^l_{1,n} u(H_{1,n}) \quad (4)$$

A mean evaluation rating provides some evidence for decision making. For example, if a product is given a high mean rating on a voice, it means that this voice should be paid high attention in future product development so that the strength of the company product in the area can be maintained. On the other hand, if a company product is given a low mean rating on a voice, it means that this voice should also be paid high attention so that the company weakness in the area can be improved. So, the mean evaluation rating of a product can be used as a criterion for assessing whether a voice should be given high priority. Another possible criterion is to use mean ratings for assessing the competitive position of a company product with respect to its competitors' products. For example, if the mean rating of a company product is significantly different from its competitors' products on a voice, then the voice should be given high priority.

While quantifying a grade $H_{1,n}$ to a fixed score $u(H_{1,n})$ is in itself problematic and may result in the partial or even wrong interpretation of data collected (as explained later in Sections 3.2 and 3.3), replacing a distributed assessment by its mean value will certainly lead to loss (and maybe distortion) of useful information. Take the following two distributed assessments for example.

$$S(A_j) = \{(H_{1,1}, 0.5), (H_{1,2}, 0), (H_{1,3}, 0), (H_{1,4}, 0), (H_{1,5}, 0.5)\} \quad (5)$$

$$S(A_k) = \{(H_{1,1}, 0), (H_{1,2}, 0), (H_{1,3}, 1), (H_{1,4}, 0), (H_{1,5}, 0)\} \quad (6)$$

If the grades are quantified evenly, or $u(H_{1,n}) = u(H_{1,n-1}) + \delta$ with δ a positive number for $n = 2, \ldots, 5$, for example $u(H_{1,1}) = 1$, $u(H_{1,2}) = 2$, $u(H_{1,3}) = 3$, $u(H_{1,4}) = 4$ and $u(H_{1,5}) = 5$, we would have $u(S(A_j)) = u(S(A_k)) = u(H_{1,3})$. Although the two assessments have the same mean value, they carry rather different types of evidence. $S(A_j)$ shows that the voice A_j should be given high priority in future product development as the product performance on this voice is either very strong or very weak, either of which is of a great concern, while $S(A_k)$ shows that A_k can be given low or no priority as there is little concern with the product performance on this voice.

The above analysis reveals that belief degrees assigned to assessment grades could themselves be used as separate evidence to derive criteria on which to assess voices. For example, if a company product achieves a significantly higher belief degree to the top grade (ie $H_{1,5}$ in (1)) than its competitors on a voice, then the voice should be given high priority so that the company strengths could be maintained. On the other hand, if a company product achieves a significantly higher belief degree to the bottom grade (ie $H_{1,1}$ in (1)) than its competitors on a voice, then that voice should be given high priority so that the company weaknesses could be improved. However, if a company product achieves a higher belief degree to a middle grade (eg, $H_{1,3}$ in (1)) than its competitors on a voice, then that voice should be given low or no priority.

Based on evidence collected from an internal survey, several criteria could be formulated on which to assess customer voices, including the distributed evaluation rating of a company product, its mean rating, and its competitive position in comparison with its competitors' products based on either mean evaluation ratings or belief degrees assessed to top and/or bottom grades. In general, a voice is regarded to be of high priority if on this voice a company product has high or low evaluation ratings and performs significantly different from its competitors' products. As investigated in next sections, such evidence can be systematically extracted from data gathered from internal surveys using rule-based data transformation techniques.

External surveys

While internal surveys are important means for gathering evidence to assess the voices of customer, external surveys can generate large scale and arguably less biased data sets. In the car manufacturing industry, for example, the well-known J.D. Power and Associates Automotive Performance Execution And Layout Study (APEAL, www.jdpa.com) measures customer ratings on retail purchasers and lessees from randomly drawn samples after about 90 days of ownership on over 100 vehicle attributes covering a number of performance areas. It is a mail-in survey that asks the customers what they like and don't like on various vehicle attributes. The evaluations in APEAL are based on an ordinal ten grade scale, interpreted as follows:

$$\begin{aligned} H_2 = \{&\text{`}H_{2,1}\text{—Unacceptable'}, \text{`}H_{2,2}\text{—Very Bad'},\\ &\text{`}H_{2,3}\text{—Bad'}, \text{`}H_{2,4}\text{—Below Average'},\\ &\text{`}H_{2,5}\text{—Average'}, \text{`}H_{2,6}\text{—Above Average'},\\ &\text{`}H_{2,7}\text{—Satisfactory'}, \text{`}H_{2,8}\text{—Good'},\\ &\text{`}H_{2,9}\text{—Very Good'}, \text{`}H_{2,10}\text{—Outstanding'}\} \end{aligned} \tag{7}$$

The evaluation rating of a product on a voice A_l from an external survey can also be represented by a distribution as follows:

$$\begin{aligned} S(A_l) = \{&(H_{2,1}, \beta^l_{2,1}), (H_{2,2}, \beta^l_{2,2}), (H_{2,3}, \beta^l_{2,3}),\\ &(H_{2,4}, \beta^l_{2,4}), (H_{2,5}, \beta^l_{2,5}), (H_{2,6}, \beta^l_{2,6}),\\ &(H_{2,7}, \beta^l_{2,7}), (H_{2,8}, \beta^l_{2,8}), (H_{2,9}, \beta^l_{2,9}),\\ &(H_{2,10}, \beta^l_{2,10}), (H_2, \beta^l_{H_2})\} \end{aligned} \tag{8}$$

where $\beta^l_{2,n}$, $n = 1, \ldots, 10$ can be interpreted and calculated as in (2) and $\beta^l_{H_2}$ measures ignorance, which can be explained in the same way as for $\beta^l_{H_1}$. Let $u(\beta^l_{2,n})$ be the score assigned to $H_{2,n}$. If $\beta^l_{H_2} = 0$, the mean value of a distribution shown in (8) is given by

$$u(S(A_l)) = \sum_{n=1}^{10} \beta^l_{2,n} u(H_{2,n}) \tag{9}$$

The mean evaluation ratings of a company product on the scale shown in (7) can be used as a criterion on which to assess whether a statement should be given high priority. Another criterion based on the mean rating is to assess the competitive position of a company product with respect to its competitors.

Survey information recorded in distributions shown in (8) can be used to acquire more evidence for formulating more criteria on which to assess whether a statement should be identified as a key customer voice. For example, if a company product achieves a higher accumulated belief degree to the top two grades (ie $H_{2,9}$ and $H_{2,10}$ in (8)) than its competitors on a voice, then the voice should be given high priority for maintaining the company strengths. Also, if a company product achieves a higher accumulated belief degree to the bottom two grades (ie $H_{2,1}$ and $H_{2,2}$ in (8)) than its competitors on a voice, then the voice should be given high priority for improving the company weaknesses. A question is how many grades of top (bottom) should be used for formulating such criteria. There is no definite answer to this question, which varies from application to application. A rule of thumb is that a criterion should be defined in such a way that meaningful evidence can be extracted from survey data. For example, if neither a company product nor its competitors perform badly, there is little point in using the bottom grade cumulative belief degree as a criterion.

A new survey framework

In the surveys discussed above, a respondent is allowed to tick only one of the assessment grades at a time and all statements in a survey are supposed to be measured on the same scale. However, respondents may occasionally wish to tick multiple grades each with a certain degree of belief. Ticking only one grade at a time is equivalent to assign 100% belief degree to the ticked grade and 0% belief degrees to all other grades. For evaluating a statement in a survey questionnaire as discussed in the section '*Internal surveys*', for example, an answer to a statement could take a form as shown in Table 1, where a respondent can choose to tick one or multiple grades and attach different belief degrees. From such a survey, the evaluation rating of a statement can be calculated as follows:

$$\beta_{l,n} = \frac{1}{K} \sum_{j=1}^{K} \beta_{l,n,j} \tag{10}$$

where $\beta_{l,n}$ is a mean belief degree assigned to the evaluation grade n on the lth statement, K the total number of respondents, and $\beta_{l,n,j}$ a belief degree assigned by respondent j to the evaluation grade n.

In the new framework, a statement can have a different set of grades; a respondent can provide complete or incomplete assessment with the total belief equal to or less

Table 1 New framework for assessing a statement

Assessment grade	*Dissatisfied completely*	*Dissatisfied*	*Neutral*	*Satisfied*	*Satisfied Completely*	*Unknown*
Survey statement l: 'Is the vehicle's fuel economy satisfactory?'						
Belief degree	$\beta_{l,1,j}$	$\beta_{l,2,j}$	$\beta_{l,3,j}$	$\beta_{l,4,j}$	$\beta_{l,5,j}$	$\beta_{l,H,j}$

Note: Tick one or multiple grades and provide degrees of belief $\beta_{l,n,j}$ with the total belief $\leqslant 1$.

than 1, or $\sum_{n=1}^{N}\beta_{l,n,j}=1$ or <1; moreover, the overall evaluation of a statement from a survey can be complete or incomplete with $\sum_{n=1}^{N}\beta_{l,n}=1$ or <1. Such flexibility is useful in cases where respondents may not be able to answer some questions. In next sections, the evidential reasoning approach will be introduced which can handle such information systematically without losing information or making assumptions unnecessarily.

In a more general framework, a respondent can be allowed to assign beliefs to subsets of grades. For example, a respondent may state that he is not dissatisfied with the statement, so his belief is completely assigned to the subset of the grades {Neutral, Satisfied, Satisfied Completely}. Such interval assessments can be aggregated using the extended evidential reasoning approach (Xu *et al*, 2006). On the other hand, interval belief degrees can also be handled by the extended evidential reasoning approach (Wang *et al*, 2006).

Rule-based data transformation for evidence-driven decision modelling

Necessity for data transformation to a common scale

The above discussion shows that different measurement scales may be used in different surveys. While a company can design its internal surveys, it does not always have the direct control of the design of external surveys. A question then arises as to how to use various types of survey information under a single framework to support the prioritization of the voices of customer. A simple approach would be to generate mean values from each survey and normalize them based on a single cardinal scale. As discussed in the previous section, however, this would cause the loss and even distortion of useful information, consequently leading to partial or biased assessment results. It has been found in our research that expert judgments are routinely used in industry for interpreting data generated from surveys. Such interpretation requires sturdy domain knowledge and substantial expertise. In this section, we investigate the principles and approaches that could be used to interpret survey data systematically using expert judgements and domain knowledge.

Since internal and external surveys may use different scales, as shown in (1) and (7), it is necessary to transform these various scales to a common scale on which data from various surveys can be synthesized under a unified framework. This process mirrors the process of normalizing multiple criteria measured in different units, such as cost in pounds and speed in miles per hour. In this section, we investigate a common scale and data transformation techniques based on expert knowledge and common sense rules.

The aim of gathering evidence through internal and external surveys is to help prioritize the voices of customer. So, a common scale should reflect this aim and be intuitively understandable and easy to use. On the other hand, it should provide a basis such that expert knowledge and common sense rules could be used to transform various survey scales to the common scale in a flexible manner. As such, a set of five monotonic priority grades is suggested as a common scale for measuring criteria on which to assess voices of customers, defined by

$$H = \{\text{'}H_1\text{—No Priority'}, \text{'}H_2\text{—Low Priority'}, \text{'}H_3\text{—Average Priority'}, \text{'}H_4\text{—High Priority'}, \text{'}H_5\text{—Top Priority'}\} \tag{11}$$

For a specific application, each priority grade needs to be defined in detail and guidance should be provided about how to differentiate between these priority grades. This forms part of a process for gathering and structuring assessment knowledge. The assessment of a product on a statement A_l on the common scale is represented as follows:

$$S(A_l) = \{(H_1, \beta_1^l), (H_2, \beta_2^l), (H_3, \beta_3^l), (H_4, \beta_4^l), (H_5, \beta_5^l), (H, \beta_H^l)\} \tag{12}$$

As discussed in the previous sections, different types of ordinal scales such as shown in (1) and (7) as well as numerical data like mean scores and belief degrees need to be transformed to the common scale. In next sections, rule based techniques are investigated to transform survey data from their original scales to the common scale, while expert knowledge can be incorporated to guide the process and refine parameters.

Symmetrical qualitative data transformation

The ordinal five grade scale shown in (1) is symmetrical. It is suggested that a voice be given 'Top Priority' in future new product development if the voice is assessed to be 'Completely Satisfied' or 'Completely Dissatisfied', with the former counted as strengths and the latter as weaknesses, and that a voice be given 'No Priority' if it is assessed to be 'Neutral'. The following equivalence rules are then proposed to transform the survey data to the voice priority data to count the strengths of a voice.

- '$H_{1,3}$—Neutral' → 'H_1—No Priority'
- '$H_{1,4}$—Satisfied' → 'H_3—Average Priority'
- '$H_{1,5}$—Completely Satisfied' → 'H_5—Top Priority'

In the above, '→' stands for 'is equivalent to' in terms of utility. Let $u(H_n)$ be the utility of H_n. The above rules mean that implicitly $u(H_{1,3}) = u(H_1)$, $u(H_{1,4}) = u(H_3)$ and $u(H_{1,5}) = u(H_5)$, which assumes that the grades be all evenly distributed in the assessment space with the grades 'Completely Satisfied' and 'Top Priority' associated with the highest utility whilst the grades 'Neutral' and 'No Priority' with the lowest utility. Note that the above requires no explicit utility assignment. Also note that such an assumption may not always be acceptable and the rules should not be fixed but can be adjusted by practitioners as appropriate.

Similarly, the following equivalence rules are proposed to transform the survey data to the voice priority data to count the weaknesses of a voice.

- '$H_{1,3}$—Neutral' → 'H_1—No Priority'
- '$H_{1,2}$—Dissatisfied' → 'H_3—Average Priority'
- '$H_{1,1}$—Completely Dissatisfied' → 'H_5—Top Priority'

The above rules mean that implicitly $u(H_{1,3}) = u(H_1)$, $u(H_{1,2}) = u(H_3)$ and $u(H_{1,1}) = u(H_5)$.

Using the above transformation rules, an assessment measured on a survey scale as shown in (3), can be transformed to an equivalent assessment on the common scale as shown in (12) in terms of equivalence between the implicit utility of the original assessment and that of the transformed assessment, with

$$\beta_1^l = \beta_{1,3}^l,\ \beta_2^l = 0,\ \beta_3^l = \beta_{1,2}^l + \beta_{1,4}^l,$$
$$\beta_4^l = 0,\ \beta_5^l = \beta_{1,1}^l + \beta_{1,5}^l,\ \beta_H^l = \beta_{H_1}^l \tag{13}$$

Note that the above transformation is based on the assumption that both the survey grades and the priority grades are evenly distributed in the assessment space. This assumption does not necessarily need to be strictly followed. In practice, rules provided by decision makers based on their knowledge and preferences should be used to govern the data transformation. For example, suppose it is felt that instead of following the previous rules, the grade 'Dissatisfied' should be equivalently transformed to 'Low Priority' to a degree of a_1, 'Average Priority' to a degree of a_2, and 'High Priority' to a degree of a_3, or more concisely

- '$H_{1,2}$—Dissatisfied' → a_1'H_2—Low Priority' & a_2'H_3—Average Priority' & a_3'H_4—High Priority' with $a_1 + a_2 + a_3 = 1$

and similarly

- '$H_{1,4}$—Satisfied' → b_1'H_2—Low Priority' & b_2'H_3—Average Priority' & b_3'H_4—High Priority' with $b_1 + b_2 + b_3 = 1$

Then, the transformed assessment on the common scale can be generated as follows:

$$\beta_1^l = \beta_{1,3}^l,\ \beta_2^l = a_1\beta_{1,2}^l + b_1\beta_{1,4}^l,$$
$$\beta_3^l = a_2\beta_{1,2}^l + b_2\beta_{1,4}^l,\ \beta_4^l = a_3\beta_{1,2}^l + b_3\beta_{1,4}^l,$$
$$\beta_5^l = \beta_{1,1}^l + \beta_{1,5}^l,\ \beta_H^l = \beta_{H_1}^l \tag{14}$$

Equations (13) and (14) show that transforming survey data to the common scale is the process of using common sense rules and/or expert knowledge to interpret the data and turn them into decision evidence on which to assess and prioritize voices. It can also be observed that in the process the important features of the data such as their diversity as well as uncertainty (ignorance or proportion of missing data) are preserved.

The principle of data transformation governed by the above rules is illustrated in Figure 2, where the anchoring point links the 'No Priority' grade on the common scale to the middle assessment grade 'Neutral' on the survey scale. The 'V' shaped mapping function seems consistent with the expert intuitive reasoning process in interpreting the survey data, whilst the original qualitative nature of the data is preserved.

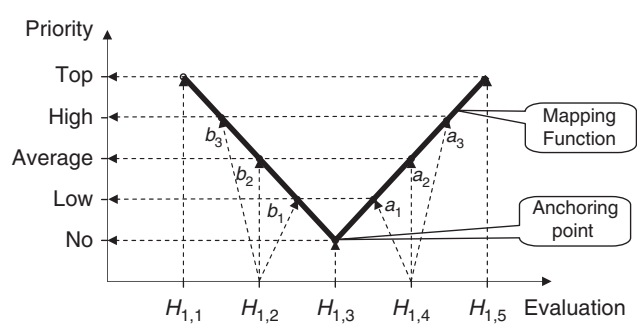

Figure 2 Symmetrical mapping function to transform survey data to the common scale.

Non-symmetrical qualitative data transformation

The ordinal scale shown in (7) is non-symmetrical and has 10 assessment grades. As such, it is not easy to decide an anchoring point, or an assessment grade in the non-symmetrical scale that should be mapped to 'No Priority' in the common scale. Such a decision needs to be made by the decision makers. Suppose the grade 'Average' of the scale shown in (7) is taken as the anchoring point. In other words, if a company product is assessed to have 'Average' performance on a voice, it is assumed that the voice will have little impact on the company's future product development and so will be given 'No Priority'.

The transformation of the external survey data to the common scale will be based on the following principles. If the performance of a company product on a voice is assessed to be 'Unacceptable', this voice will be given 'Top Priority' for the company's future product development, to ensure that the area associated with this voice would be improved. Then, the following equivalence rules are proposed to transform survey data to the voice priority data on the common scale to count company weaknesses on a voice.

- '$H_{2,5}$—Average' → 'H_1—No Priority'
- '$H_{2,4}$—Below Average' → 'H_2—Low Priority'
- '$H_{2,3}$—Bad' → 'H_3—Average Priority'
- '$H_{2,2}$—Very Bad' → 'H_4—High Priority'
- '$H_{2,1}$—Unacceptable' → 'H_5—Top Priority'

In the above rules, it is assumed that the assessment grades are all evenly distributed in the assessment space, so implicitly $u(H_{2,5}) = u(H_1)$, $u(H_{2,4}) = u(H_2)$, $u(H_{2,3}) = u(H_3)$, $u(H_{2,2}) = u(H_4)$ and $u(H_{2,1}) = u(H_5)$.

On the other hand, if the performance of a company product on a voice is assessed to be 'Outstanding', this voice will be given 'Top Priority' to ensure that company strengths on this voice can be maintained. If the assessment grades are further assumed to be all evenly distributed in the assessment space, the following equivalence rules are proposed to transform survey data to the voice priority data on the common scale for counting the company strengths on a voice.

- '$H_{2,5}$—Average' → 'H_1—No Priority'
- '$H_{2,6}$—Above Average' → 0.2'H_1—No Priority' & 0.8'H_2—Low Priority'
- '$H_{2,7}$—Satisfactory' → 0.4'H_2—Low Priority' & 0.6'H_3—Average Priority'
- '$H_{2,8}$—Good' → 0.6'H_3—Average Priority' & 0.4'H_4—High Priority'
- '$H_{2,9}$—Very Good' → 0.8'H_4—High Priority' & 0.2'H_5—Top Priority'
- '$H_{2,10}$—Outstanding' → 'H_5—Top Priority'

In the above rules, it is assumed that the assessment grades are all evenly distributed in the assessment space, so implicitly $u(H_{2,5}) = u(H_1)$, $u(H_{2,6}) = 0.2u(H_1) + 0.8u(H_2)$, $u(H_{2,7}) = 0.4u(H_2) + 0.6u(H_3)$, $u(H_{2,8}) = 0.6u(H_3) + 0.4u(H_4)$, $u(H_{2,9}) = 0.8u(H_4) + 0.2u(H_5)$ and $u(H_{4,10}) = u(H_5)$. The data transformation function representing the above rules is illustrated in Figure 3.

Using the above transformation rules, an assessment measured on the survey scale shown in (8) can be transformed to an equivalent assessment on the common scale shown in (12), with

$$\begin{aligned}
\beta_1^l &= \beta_{2,5}^l + 0.2\beta_{2,6}^l, \\
\beta_2^l &= \beta_{2,4}^l + 0.8\beta_{2,6}^l + 0.4\beta_{2,7}^l, \\
\beta_3^l &= \beta_{2,3}^l + 0.6\beta_{2,7}^l + 0.6\beta_{2,8}^l, \\
\beta_4^l &= \beta_{2,2}^l + 0.4\beta_{2,8}^l + 0.8\beta_{2,9}^l, \\
\beta_5^l &= \beta_{2,1}^l + 0.2\beta_{2,9}^l + \beta_{2,10}^l, \ \beta_H^l = \beta_{H_2}^l
\end{aligned} \tag{15}$$

The above equations are based on the assumption of all grades being evenly distributed in the assessment space. Similar to the explanations made for (14), this assumption does not necessarily need to be followed strictly and the rules can be adjusted according to the decision maker's knowledge and preferences.

In a survey, both a company's and its competitors' products can be evaluated using the mean rating or belief degrees assessed to top and/or bottom grades. Various metrics can be used to measure differences between the evaluations of a company's and its competitors' products. Such quantitative metrics can also be transformed to the common scale using the rule or utility based data transformation techniques (Yang, 2001). It should be noted that in the above transformation functions the determination of the anchoring points and function shapes is domain specific and application dependent. They need to be decided by experienced engineers and managers who have appropriate knowledge and understanding of the implications and purposes of data transformation.

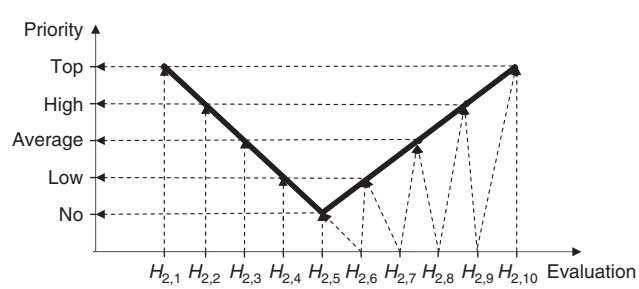

Figure 3 Non-symmetrical mapping function to transform survey data to the common scale.

Evidential reasoning based multi-criteria voice assessment

Evidence-driven decision modelling by belief decision matrix

Internal and external surveys can generate rich information, the main features of which can be preserved in the above-discussed data transformation process from various survey scales to a common scale. This provides a sound basis for evidence-driven decision modelling. The prioritization of the voices of customer is in essence a multiple criteria decision analysis (MCDA) problem. As discussed earlier, typical criteria include how well a company product performs on a voice and how differently it performs from its competitors. Such a MCDA problem can be represented using the following belief decision matrix.

Let A_l be the lth voice, $C_{i,j}$ the ith criterion created from the jth survey and N the number of grades. Then, a key voice prioritization problem can be represented using the following *belief decision matrix*:

$$BM = (S_{i,j}(A_l))_{L \times M} \tag{16}$$

In (16), L is the total number of voices, M the total number of criteria and $S_{i,j}(A_l)$ the distributed assessment of a voice A_l on the criterion $C_{i,j}$ represented on a common scale such as (11) by

$$\begin{aligned} S_{i,j}(A_l) = \{&(H_1, \beta^l_{i,j,1}), (H_2, \beta^l_{i,j,2}), \ldots, \\ &(H_N, \beta^l_{i,j,N}), (H, \beta^l_{i,j,H})\} \end{aligned} \tag{17}$$

$\beta^l_{i,j,n}$ can be calculated using Equations (13), (14), (15) or (17), depending upon the type of survey.

The major difference between a conventional decision matrix and the belief decision matrix is that in the former a single real number (score or utility) is used to measure the assessment of a voice on a criterion while in the latter a two-dimensional distribution is used. A belief decision matrix provides a natural representation of evidence generated from surveys and can record missing information, without distorting information or having to make unnecessary assumptions. This is desirable for supporting evidence-driven voice assessment. Clearly, conventional decision matrix is a special case of belief decision matrix.

For assessing voices, weights need to be estimated to represent the relative importance of one survey against another. Weight estimation is an important part in decision modelling and needs to be done properly. There are a plethora of methods for weight estimation in the literature (Saaty, 1994; Olson, 1996) but weight estimation is not the focus of this paper and hence not discussed here in great detail. In general, an external survey may be more heavily weighted than an internal survey; an independent external survey conducted by a non-profit making organization such as a consumer group may be more heavily weighted than an external survey conducted by a commercial consulting firm. Evaluation criteria for a particular survey also need to be weighted. For instance, equal weights may be given to the same group of criteria in a survey if there are no particular preferences given to the survey criteria.

A new generic ranking method for the evidential reasoning approach

The distributed assessments of a voice on multiple criteria elicited from a survey, represented by a belief decision matrix, need to be aggregated to generate a combined assessment of the voice from the survey. Eventually, an overall assessment of a voice from all surveys should be generated. In the aggregation process, it is desirable to preserve the distribution feature of the original assessments at various levels of an assessment hierarchy. The evidential reasoning (ER) approach is well suited for this purpose and its details can be found in the literature (Yang, 2001, Yang and Xu, 2002a, Yang *et al*, 2006; Xu *et al*, 2006; Wang *et al*, 2006). The basic aggregation steps of the ER approach includes the following steps:

Step 1: Assignment of the basic probability masses,
Step 2: Combination of the basic probability masses,
Step 3: Generation of the combined belief degrees and distributed assessment,
Step 4: Representation of the combined result as a distribution as follows

$$\begin{aligned} S(A_l) = \{&(H_1, \beta^l_1), (H_2, \beta^l_2), \ldots, \\ &(H_N, \beta^l_N), (H, \beta^l_H)\} \end{aligned} \tag{18}$$

Step 5: The ranking of voices and sensitivity analysis.

In (18), β^l_n ($n = 1, \ldots, N$) are the combined belief degree generated using the ER algorithm and β^l_H measuring the aggregated ignorance (Yang and Xu, 2002a). Distribution (18) provides a panoramic view about the combined assessment of the voice A_l with the degrees of strengths and weaknesses explicitly measured by the belief degrees. Although the distribution is useful to show the diversity of assessment information, it is not convenient for ranking voices. As shown in (4), the mean value of a distribution can be used as a mean evaluation rating of a voice. However, the mean value of a distribution depends on how the assessment grades are quantified. In this section, a new generic method is proposed to support the robust ranking of voices and sensitivity analysis.

Suppose H_n is quantified by $u(H_n)$, which should not be fixed to a single value but in general should be a variable interval, denoted by $[a_n, b_n]$ with $b_n \geqslant a_n$ and a_n and b_n not fixed. Suppose the grade H_{n+1} is also quantified by a variable interval $[a_{n+1}, b_{n+1}]$. Then, H_{n+1} is said to be preferred to H_n if and only if $a_{n+1} > b_n$ for $n = 1, \ldots, N-1$. Without loss of generality, suppose $u(H_n)$ is normalized so that $a_1 = 0$ and $b_N = 1$.

In the presence of ignorance, the lower and upper bounds of the belief degree to which the voice A_l is assessed to H_n is given by β_n^l and $(\beta_n^l+\beta_H^l)$, respectively; hence, the mean value of a distribution is characterised by a maximum and minimum value given as follows (Yang and Xu, 2002a):

$$u_{\min}(S(A_l)) = (\beta_1^l + \beta_H^l)u(H_1) + \sum_{n=2}^{N} \beta_n^l u(H_n) \tag{19a}$$

$$u_{\max}(S(A_l)) = \sum_{n=1}^{N-1} \beta_n^l u(H_n) + (\beta_N^l + \beta_H^l)u(H_N) \tag{19b}$$

Then, a voice A_l is preferred to another voice A_k if $u_{\min}(S(A_l)) \geqslant u_{\max}(S(A_k))$.

The pairwise comparison of A_l and A_k can be made directly using Equations (19a) and (19b) if the utilities of the grades are given precisely by the decision makers, or fixed to $u(H_n)=a_n=b_n$. There is rich literature about partially ranking alternatives using additive aggregations under imprecise weights, for example (Hazen, 1986; Yang and Sen, 1996; Dias and Clímaco, 2000; Park, 2004). In this section, we propose a new method that can be used to rank voices partially in a general manner without having to assume a specific utility function. In this method, we construct the following generic linear programming model for comparing two voices A_l with A_k

$$\begin{aligned} &\max \quad \sigma_{kl} = u_{\max}(S(A_k)) - u_{\min}(S(A_l)) \\ &\text{s.t.} \quad \mathbf{X} = [(a_1, u(H_1), b_1), \ldots, (a_N, u(H_N), b_N)]^T \in \Omega_d \end{aligned} \tag{20a}$$

$$\Omega_d = \left\{ \mathbf{X} \left| \begin{array}{ll} a_n \leqslant u(H_n) \leqslant b_n & n = 1, \ldots, N \\ b_n \geqslant a_n & n = 1, \ldots, N \\ a_{n+1} - b_n \geqslant \delta & n = 1, \ldots, N-1 \\ a_1 = 0,\ b_N = 1 & \end{array} \right. \right\} \tag{20b}$$

In (20b), δ is a small positive real number that should be large enough to make the difference between two adjacent grades significant and meaningful, with $0<\delta\leqslant 1/(N-1)$. If the optimal value of formulation (20a) for a sufficiently small δ is negative, or $\sigma_{kl}<0$, it means that $u_{\min}(S(A_l)) \geqslant u_{\max}(S(A_k))$ for any permissible quantification of the assessment grades (ie, as bounded by Ω_d). In this paper, if $\sigma_{kl}<0$, it is said that A_l is evidentially preferred to A_k since such a ranking does not depend on any particular values of $u(H_n)$. On the other hand, if $\sigma_{kl}>0$ no definite partial ranking between A_l and A_k exists, and assumptions would have to be made or preference conditions would need to be added to (20b) in order to rank A_l and A_k, which may be referred to as assumption-based or preference-based ranking.

In general, the larger the value of δ, the more powerful the formulation (20) is to differentiate voices but the less robust the differentiation is. For example, formulation (20a) will have a unique solution of $u(H_n)=a_n=b_n=(n-1)/(N-1)$ if δ is assumed to take the maximum permissible value of $1/(N-1)$, which is equivalent to assigning a single fixed value to each grade so that grades are evenly distributed in the utility space. On the other hand, a grade may be quantified to a fixed utility interval rather than a single utility. For example, each grade may be quantified to an equal utility interval so that $(a_n-b_n)=1/(N-1)$, which is equivalent to assign $\delta=0$, $a_1=0$, $a_{n+1}=b_n=(n-1)/(N-1)$ for $n=1, \ldots, N-1$, and $b_N=1$. Anyway, if it is necessary to make such assumptions for ranking voices, they should be made adequately to suit specific decision situations.

Preference conditions may be added to formulation (20b) for differentiating voices. For example, if an improvement from a low grade is more appreciated than from any better grade, then the following generic pseudo-concave conditions could be added to formulation (20b)

$$u(H_{n+1}) - u(H_n) \geqslant u(H_{n+2}) - u(H_{n+1}) \quad \text{for } n = 1, \ldots, N-2 \tag{21}$$

Conversely, if improvement from a high grade is more appreciated than from any worse grade, then the following generic pseudo-convex conditions could be added to formulation (20b)

$$u(H_{n+2}) - u(H_{n+1}) \geqslant u(H_{n+1}) - u(H_n) \quad \text{for } n = 1, \ldots, N-2 \tag{22}$$

On the other hand, if the degree of improvement from a grade to an adjacent better grade is bounded, the following bound conditions may be added to formulation (20b), with τ being a bound index and $\tau\geqslant 1$,

$$\frac{1}{\tau} \leqslant \frac{u(H_{n+2}) - u(H_{n+1})}{u(H_{n+1}) - u(H_n)} \leqslant \tau \quad \text{for } n = 1, \ldots, N-2 \tag{23}$$

The intelligent decision system for multi-criteria analysis

The above-described ER approach is not easy to implement manually. The Intelligent Decision System (IDS) software (Yang and Xu, 2005) has been developed to implement the ER approach. IDS is a general-purpose multiple criteria decision analysis tool and provides Windows-based graphical interfaces to build a decision model where all voices can be assessed on a hierarchy of criteria using the belief structures. Each voice can be assessed using a distribution on the defined common scale. The rank order of all voices can be generated on the basis of utility scores. IDS is introduced and used to support a case study as reported in the next section. Most of the

results for the case study are generated using the user-interfaces of the IDS software and illustrated in the next sections.

Prioritization of the voices of customer—a case study

In this section, a case study for prioritizing the voices of customers for a world-leading car manufacturer is reported, in which some of the features of the above concepts and methods and their potential for wider applications are demonstrated. Due to commercial confidentiality, the following sections are intended to discuss the case study in general terms rather than describe it in any commercial detail. The data and information used in the case study are for illustration purposes only and are not meant to be directly related to any particular product or process of the company.

Data collection, transformation and sensitivity analysis

In one of the company's internal surveys (referred as Survey 1 from here on), the respondents are asked to evaluate the company's and competitors' vehicles on a five point symmetric scale. The grades of the five point symmetric scale are coded as: DAS, DA, N, A, and AS, in the increasing order of the agreement with the statement of the survey. The answers from all the respondents are individually recorded and are pre-processed as distributions in this case study. Figure 4 shows the distribution of the responses on a survey statement obtained for a vehicle of interest. From Figure 4, we can see that 5.26% of respondents selected DAS, 9.21% selected DA, 44.74% selected N, 34.21% selected A, and 6.58% selected AS, as their response.

As the five-point scale of this internal survey is symmetric, one could use the transformation function shown in Figure 2 (see the section '*Symmetrical qualitative data transformation*' for details), for transforming the data to the common scale as defined in (11). Using the transformation function shown in Figure 2, with $a_1 = b_1 = 0.15$, $a_2 = b_2 = 0.7$, and $a_3 = b_3 = 0.15$, the distribution shown in Figure 4 can then be transformed to a distributed assessment of the voice, as shown in Figure 5a. The distributed assessment in Figure 5a means that this voice should not be given very high priority as far as this internal survey is concerned. Note that the engineers and analysts of the company are free to use their own rules to transform the information. For example, if the mapping parameters a_i and b_i are given by $a_1 = b_1 = 0.3$, $a_2 = b_2 = 0.4$, and $a_3 = b_3 = 0.3$, then a new assessment will be generated as shown in Figure 5b, which has higher beliefs to the 'Low Priority' and 'High Priority' grades and lower belief to the 'Average Priority' grade than the case shown in Figure 5a.

In this case study, we used data from another internal survey, referred as Survey 2 from here on. The environment, questionnaire type, and the scale used in Survey 2 are different from those of Survey 1. For example, Survey 2 uses a non-symmetrical scale with five point grades coded as NS, SS, S, VS, and CS, in the increasing order of satisfaction to the respondent. As discussed in the section '*Non-symmetrical qualitative data transformation*', the following transformation rules are used to transform the data from the non-symmetrical scale of Survey 2 to the common scale.

- NS → 'Top Priority'
- SS → 'No Priority'
- S → 67% 'Low Priority' & 33% 'Average Priority'
- VS → 33% 'Average Priority' & 67% 'High Priority'
- CS → 'Top Priority'

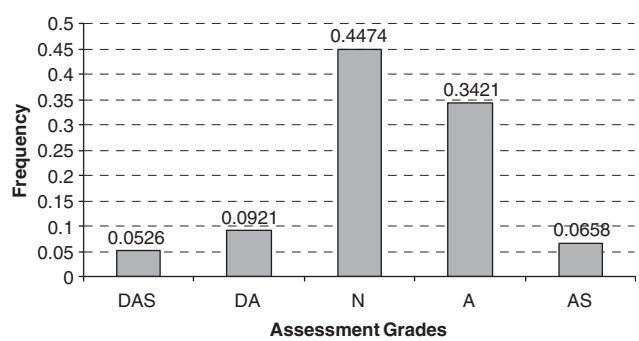

Figure 4 Distribution of an example Survey 1 data on its original scale.

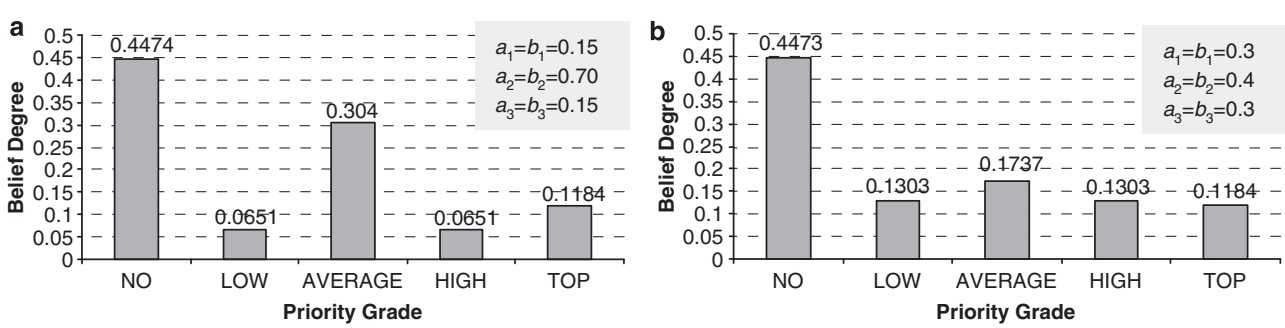

Figure 5 (**a**) Sensitivity analysis of transformed assessments.

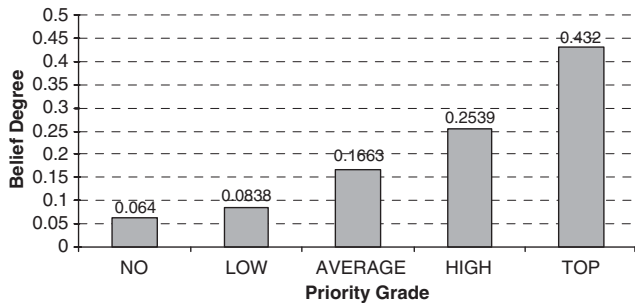

Figure 6 Distribution of example Survey 2 data on the common scale.

If a statement in Survey 2 has the following distribution of responses on its original scale: {(NS, 0.064), (SS, 0.064), (S, 0.125), (VS, 0.379), (CS, 0.368)}, then the ensuing distribution on the common scale using the above transformation rules is shown in Figure 6. From Figure 6, we can see that this voice should be given very high priority as far as the data from Survey 2 is concerned. Again, the true parameters used in the transformation rules need be decided by the decision makers of the company.

In this case study, we also used data from external surveys. Some of these external surveys provided data of the same qualitative nature as the internal surveys. Some other external surveys only provided quantitative data or mean scores. The techniques detailed in the literature (Yang, 2001) can be used to transform the quantitative data to the common scale.

Evidence-driven decision modelling for assessment of customer voices

In this case study, thousands of data sets generated from four internal and external surveys were used and transformed using the techniques discussed in the previous sections to formulate a number of decision models for assessing over 160 voices. It is not appropriate to present all the models in detail in this paper due to the confidentiality as well as the large scale. In this section, the generic structure and a typical small scale model will be discussed to illustrate the evidence-driven decision modelling framework and the reasoning-based decision support process for the prioritization of the voices of customer. While the entire MCDA problem constructed for this case study is large and complicated, its principle and main features can be demonstrated by a sub-problem investigated as follows, as illustrated using the IDS software.

The hierarchy for assessing four voices using evidence generated from two surveys is shown in Figure 7, which is the IDS main window, including a bottom-right tree view for displaying a hierarchy of assessment criteria, a bottom-left list view for listing all voices to be assessed, a menu bar for listing all the IDS functions and a shortcut bar for quick access to frequently used IDS functions. In Figure 7,

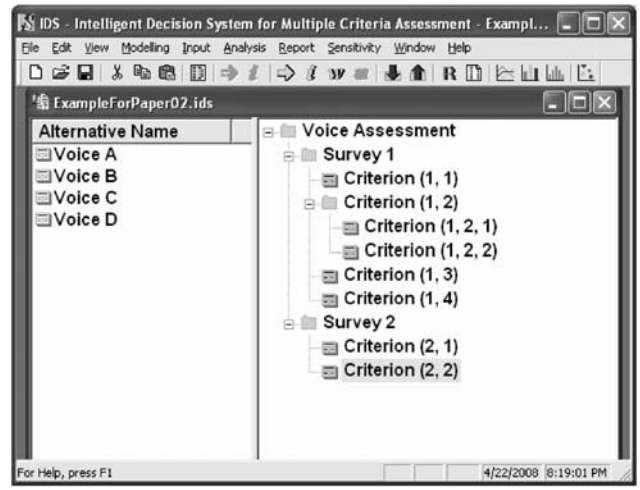

Figure 7 The IDS main window for the voice assessment sub-problem.

there are seven criteria at the bottom of the hierarchy that are derived from the evidence generated from Survey 1 and Survey 2. Survey 1 has five bottom-level criteria and Survey 2 has two bottom-level criteria.

In Figure 7, 'Criterion (1, 4)' is the evaluation obtained directly from Survey 1 and transformed to common scale using symmetrical qualitative transformation rules discussed in the section '*Symmetrical qualitative data transformation*'. 'Criterion (2, 2)' is obtained by non-symmetrical qualitative transformation rules discussed in the section '*Non-symmetrical qualitative data transformation*'. The remaining criteria in Figure 7 are proprietary information and cannot be discussed in detail here. Some of the criteria in Figure 7 involve comparison of company's product with relevant competition. While the exact criteria used for the case study and later implemented by the company are confidential, in principle one could use average ratings and statistical comparison between company's and competition's product as criteria. Also note that one could use different criteria for different voices, but our case study as reported in this paper was a coincidence where all voices had the same criteria hierarchy. In fact, in the processes of validating and implementing the methodology for all the voices identified by the company, different criteria hierarchies have been used for different groups of voices.

Given the nature of the two surveys, weights are suggested by the company staff in such a way that Survey 1 is regarded to be more reliable and better trusted than Survey 2. For this case study, weight is used as a proxy measure to assess the degree of relative reliability of and trust in surveys, and the weight for Survey 1 was thus judged to be twice as large as that for Survey 2 as a starting point for investigation. The weights for Survey 1 and Survey 2 are thus normalized to $\frac{2}{3}$ and $\frac{1}{3}$, respectively, which are subject to sensitivity analysis. The other criteria in the same groups of the hierarchy are all equally weighted in this case study.

The original data sets generated from the two surveys were pre-processed into the distributions on the survey scales. The pre-processed data are then transformed to the distributed assessments on the common scale (11) as evidence for assessing the four voices on the seven bottom level criteria, summarised in a belief decision matrix as shown in Table 2. The numbers in the brackets are belief degrees associated with the evaluation grades on the common scale in the order of {NO, LOW, AVERAGE, HIGH, TOP}.

The 'Unassigned Priority' (or Unknown) is not explicitly given in Table 2 and can be calculated by one minus the sum of the five numbers in an assessment. In the assessment of Voice *B* on 'Criterion (2, 2)', for example, the total belief degree is 0.96, which means that 4% of the responses are missing from the original Survey 2 data. The missing information of 4% recorded in the belief decision matrix will be preserved in the aggregation process.

Reasoning-based assessment aggregation, voice ranking and sensitivity analysis

In this section, we use the sub-model as discussed in the previous section to demonstrate how to apply the ER approach and the IDS software for aggregating assessments from the bottom level criteria of the hierarchy progressively to the top level voice assessment, ranking the four voices, and conducting sensitivity analysis. The following figures were all generated using the IDS software. In the case study, the original survey data sets were provided in Excel files. The data were transformed to the common scale using Matlab programmes. The generated data were then read into the IDS software through its data input interface.

For each voice, the evidential reasoning approach is employed to first combine the assessments on the two bottom-level criteria 'Criterion (1, 2, 1)' and 'Criterion (1, 2, 2)', resulting in an aggregated assessment on the higher level criterion 'Criterion (1, 2)'. The assessments on Criterion (1, 1), Criterion (1, 2), Criterion (1, 3) and Criterion (1, 4) are then combined to generate an aggregated assessment on Survey 1. Similarly, the assessments on Criterion (2, 1) and Criterion (2, 2) are combined to generate an aggregated assessment on Survey 2. The aggregated assessments on Survey 1 and Survey 2 are shown in Table 3.

In Table 3, it can be observed that the assessments of the four voices are all complete on Survey 1 but incomplete on Survey 2, though the overall degrees of uncertainty (unknown) are relatively small. This may be explained by the fact that Survey 1 is a monitored or supported process,

Table 2 A sub-belief decision matrix for assessing four voices

Criterion	*Voice*			
	A	*B*	*C*	*D*
Criterion (1, 1)	{0.01, 0.01, 0.05, 0.25, 0.69}	{0, 0.01, 0.04, 0.29, 0.66}	{0, 0.03, 0.13, 0.35, 0.49}	{0.00, 0.01, 0.13, 0.46, 0.39}
Criterion (1, 2, 1)	{0, 1, 0, 0, 0}	{0, 0, 0, 0, 1}	{0, 0, 0, 0, 1}	{0, 0, 0, 0, 1}
Criterion (1, 2, 2)	{0, 1, 0, 0, 0}	{1, 0, 0, 0, 0}	{0, 0, 0, 0, 1}	{0, 0, 0, 0, 1}
Criterion (1, 3)	{0, 0, 0, 0, 1}	{0, 0, 0, 0, 1}	{0, 0, 0, 0, 1}	{0, 0, 0, 0, 1}
Criterion (1, 4)	{0.15, 0.06, 0.26, 0.06, 0.47}	{0.14, 0.06, 0.27, 0.06, 0.47}	{0.32, 0.07, 0.35, 0.07, 0.19}	{0.16, 0.07, 0.32, 0.07, 0.38}
Criterion (2, 1)	{0.91, 0, 0, 0, 0}	{0, 0, 0, 0, 0.98}	{0, 0, 0, 0, 0.98}	{0.97, 0, 0, 0, 0}
Criterion (2, 2)	{0.01, 0.05, 0.13, 0.21, 0.51}	{0.01, 0.04, 0.10, 0.17, 0.64}	{0.04, 0.07, 0.13, 0.18, 0.56}	{0.02, 0.05, 0.13, 0.21, 0.56}

Table 3 Aggregated assessments at the survey level

Survey	*Voice*			
	A	*B*	*C*	*D*
Survey 1	{(No, 0.0350), (Low, 0.2382), (Average, 0.0697), (High, 0.0673), (Top, 0.5898)}	{(No, 0.1332), (Low, 0.0135), (Average, 0.0641), (High, 0.0706), (Top, 0.7187)}	{(No, 0.0641), (Low, 0.0212), (Average, 0.0988), (High, 0.0886), (Top, 0.7274)}	{(No, 0.0328), (Low, 0.0163), (Average, 0.0935), (High, 0.1076), (Top, 0.7498)}
Survey 2	{(No, 0.4667), (Low, 0.0230), (Average, 0.0628), (High, 0.1044), (Top, 0.2541), (Unknown, 0.0891)}	{(No, 0.0048), (Low, 0.0136), (Average, 0.0394), (High, 0.0664), (Top, 0.8562), (Unknown, 0.0197}	{(No, 0.0160), (Low, 0.0280), (Average, 0.0492), (High, 0.0718), (Top, 0.8202), (Unknown, 0.0148)}	{(No, 0.5043), (Low, 0.0259), (Average, 0.0628), (High, 0.1016), (Top, 0.2753), (Unknown, 0.0302)}

while Survey 2 is a postal survey with the respondents not directly supported by the survey team, leading to no answers to some survey questions. It can also be observed that the evidence gathered from the two surveys assesses Voice B and Voice C to be of very high priority. However, the evidence gathered from Survey 1 is in conflict with Survey 2 for assessing Voices A and D. Survey 1 assesses Voice A and Voice D to be of very high priority but Survey 2 assesses them to be of rather low priority. Such conflicting evidence is common in this case study. The evidential reasoning approach preserves such diversity in its evidence combination process, which is desirable to support informative decision making.

The initially aggregated assessments are then further combined to generate an assessment for each voice on the overall criterion 'Voice assessment' as the following overall distributed assessments:

$$S(A) = \{(H_1, 0.1164), (H_2, 0.1856), (H_3, 0.0655), (H_4, 0.0721), (H_5, 0.5444), (H, 0.0161)\}$$

$$S(B) = \{(H_1, 0.0865), (H_2, 0.0109), (H_3, 0.0483), (H_4, 0.0574), (H_5, 0.7938), (H, 0.0031)\}$$

$$S(C) = \{(H_1, 0.0442), (H_2, 0.0183), (H_3, 0.0731), (H_4, 0.0707), (H_5, 0.7913), (H, 0.0024)\}$$

$$S(D) = \{(H_1, 0.1210), (H_2, 0.0168), (H_3, 0.0820), (H_4, 0.1012), (H_5, 0.6737), (H, 0.0054)\}$$

From the above overall assessments, it seems reasonable to assume that Voice B and Voice C are of higher priority than Voice A and Voice D and that Voice D is of higher priority than Voice A. However, it is difficult to compare Voice B and Voice C from their distributions. From the above observation of the four assessments, we could assume the following partial ranking order of the four voices: Voice $C \sim$ Voice $B \succ$ Voice $D \succ$ Voice A, where the symbol '$\sim$' means 'is equivalent to' and '$\succ$' means 'is preferred to'. In order to rank and prioritize these voices with confidence, more detailed analysis than the intuitive observation of the distributions is required.

A conventional approach for ranking voices based on distributed assessments is to quantify the assessment grades first and then calculate a mean score (or utility) for each voice. If the assessment grades are quantified so that a change from any grade to its adjacent grade is equally appreciated or depreciated, then we would have a pseudo-linear utility function as shown in Figure 8. The utility function is said to be pseudo-linear because it looks like a linear function if the grades are evenly distributed along the horizontal axis, though they do not have to be evenly distributed. Under the assumption of pseudo-linear utility function, the four voices are ranked as shown in Figure 9, or Voice $C \succ$ Voice $B \succ$ Voice $D \succ$ Voice A. The small grey areas on the top of the bars in Figure 9 show the degrees of ignorance in the assessments, though the ignorance does not affect the ranking in this case.

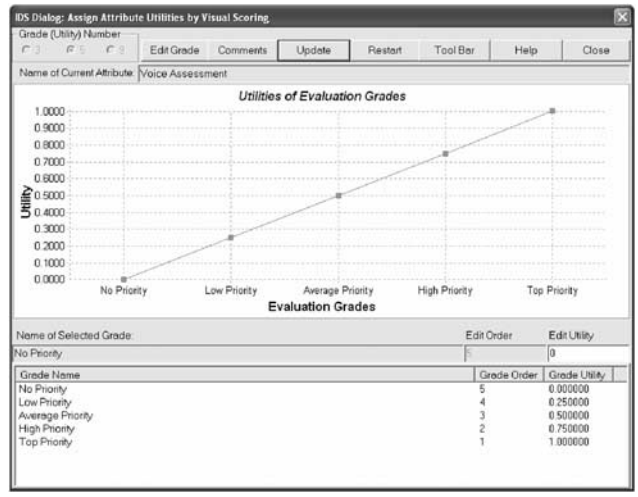

Figure 8 Pseudo-linear utility function.

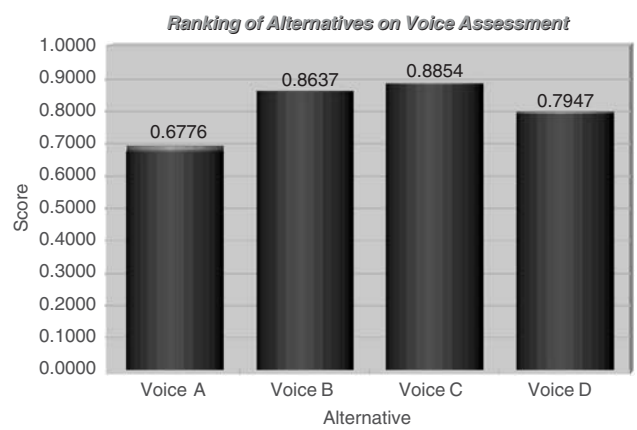

Figure 9 Pseudo-linear ranking.

If the grades are quantified so that a change from a low grade is more appreciated than from a higher grade, then we will have a pseudo-concave utility function for the qualitative grades, for example as shown in Figure 10. Given the *fixed* pseudo-concave utility function as shown in Figure 10, the four voices are ranked as shown in Figure 11, or Voice $C \succ$ Voice $B \succ$ Voice $D \succ$ Voice A.

The above rankings were generated by assuming that Survey 1 is twice as important as Survey 2. It is worth investigating the robustness of the rankings by changing the weight. If the pseudo-linear utility function is assumed,

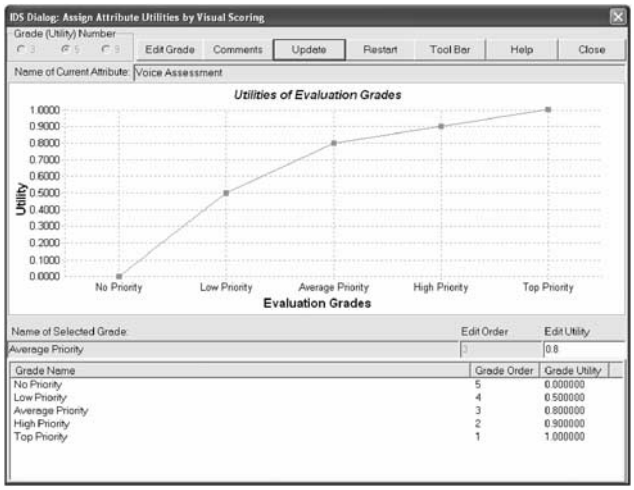

Figure 10 Pseudo-concave utility function.

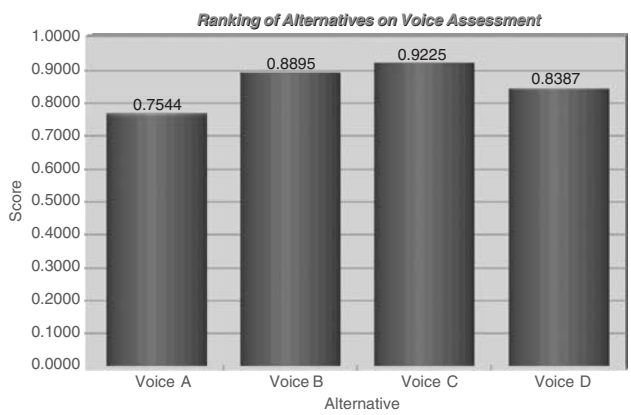

Figure 11 Pseudo-concave ranking.

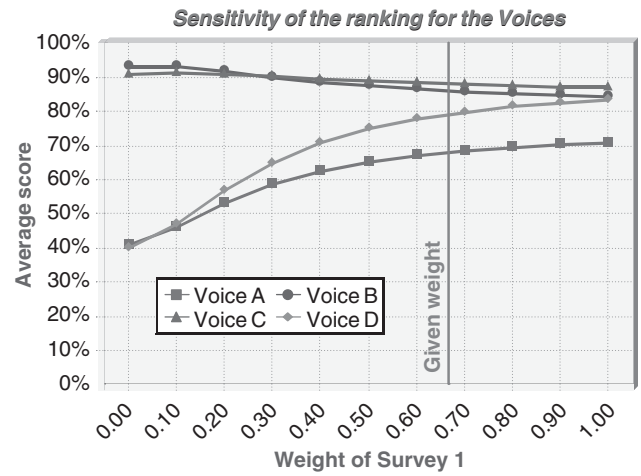

Figure 12 Sensitivity of pseudo-linear ranking.

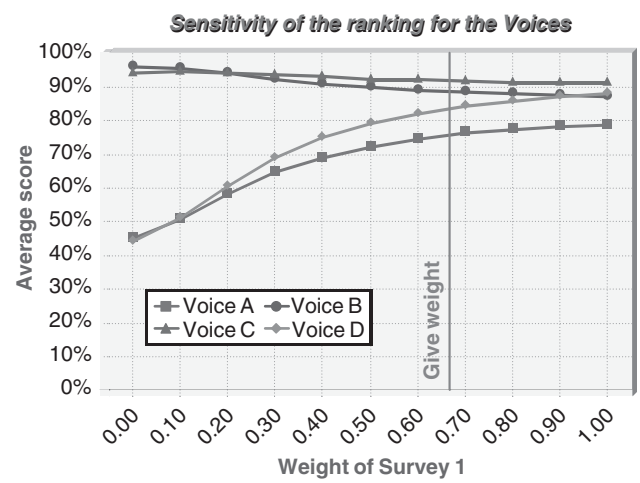

Figure 13 Sensitivity of pseudo-concave ranking.

the ranking of the four voices changes as shown in Figure 12 with the weight of survey 1 changed from 0 to 1. This shows that the ranking is insensitive to the change of the weights around the currently given weight of two thirds to survey S1. A similar feature can be found if the *fixed* pseudo-concave utility function is assumed, as shown in Figure 13. This analysis shows that the ranking is robust to weight changes around the currently given weights to the two surveys.

In the above analyses, we assumed that the precise utilities of all the grades were estimated. Such an assumption is questionable in that the robustness of the ranking generated on the basis of the assumption needs to be examined. In general, as discussed in the section '*A new generic ranking method for the evidential reasoning approach*', the utility of a qualitative grade should be a variable interval but not fixed to a single value, unless the decision makers can provide sufficient and precise preference information to estimate a unique utility function for all assessment grades. In what follows, we employ the generic ranking method proposed in the section '*A new generic ranking method for the evidential reasoning approach*' to generate the ranking of the four voices and investigate its robustness without making the above strict assumptions unnecessarily.

Take the pairwise comparison of Voice C and Voice A for example. From their overall assessments, the minimum mean utility of Voice C and the maximum mean utility of Voice A are given as follows:

$$\begin{aligned} u_{\min}(S(C)) = {} & 0.0466u(H_1) + 0.0183u(H_2) \\ & + 0.0731u(H_3) + 0.0707u(H_4) \\ & + 0.7913u(H_5) \end{aligned}$$

$$\begin{aligned} u_{\max}(S(A)) = {} & 0.1164u(H_1) + 0.1856u(H_2) \\ & + 0.0655u(H_3) + 0.0721u(H_4) \\ & + 0.5605u(H_5) \end{aligned}$$

Table 4 Pairwise comparison indices of the voices

Voice A	*Voice B*	*Voice C*	*Voice D*
$\sigma_{AA}=0.0000$	$\sigma_{AB}=-0.0170$	$\sigma_{AC}=-0.0192$	$\sigma_{AD}=-0.0011$
$\sigma_{BA}=0.2463$	$\sigma_{BB}=0.0000$	$\sigma_{BC}=0.0032$	$\sigma_{BD}=0.1182$
$\sigma_{CA}=0.2511$	$\sigma_{CB}=0.0433$	$\sigma_{CC}=0.0000$	$\sigma_{CD}=0.1174$
$\sigma_{DA}=0.1745$	$\sigma_{DB}=-0.0064$	$\sigma_{DC}=-0.0085$	$\sigma_{DD}=0.0000$

Table 5 Evidential partial ranking of the voices

Voice A	*Voice B*	*Voice C*	*Voice D*
—	$B \succ A$	$C \succ A$	$D \succ A$
—	—	—	—
—	—	—	—
—	$B \succ D$	$C \succ D$	—

The model for comparing Voice C with Voice A is then given by

$$\begin{aligned} &\max \quad \sigma_{AC} = u_{\max}(S(A)) - u_{\min}(S(C)) \\ &\text{s.t.} \quad \mathbf{X} = [(a_1, u(H_1), b_1), \ldots, (a_N, u(H_N), b_N)]^T \in \Omega_d \end{aligned}$$

Set δ in (20b) to be a sufficiently small positive number that makes differences between grades meaningful, for example $\delta = 0.1 \times (1/(5-1)) = 0.025$. Then, the optimal value of the linear programming problem is given by $\sigma_{AC} = -0.0192$, which means that it is evidentially true that Voice $C \succ$Voice A.

Similarly, we can construct a linear programming model for comparing each pair of the four voices. The results are summarized in Table 4. We can then generate the evidential preference relations between certain pairs of voices as shown in Table 5, resulting in the following partial evidential ranking of the voices: Voice $C \sim$Voice $B \succ$ Voice $D \succ$Voice A, which is consistent with our earlier visual observation.

To differentiate Voice B from Voice C, additional preference information need be added. For instance, if the utility function of the assessment grades is assumed to be generically pseudo-concave, as defined in (21), then the following linear programme can be constructed:

$$\begin{aligned} &\max \quad \sigma_{BC} = u_{\max}(S(B)) - u_{\min}(S(C)) \\ &\text{s.t.} \quad \mathbf{X} = \in \Omega_d \oplus (21) \end{aligned}$$

The optimal value of the above problem is given by $\sigma_{BC} = -0.0019$, which means that Voice $C \succ$Voice B given the assumption of a *generic* pseudo-concave utility function. If σ_{BC} were still non-negative, then more strict preference information would be needed. If precise grade utilities are estimated, a complete ranking of the voices will always be generated. For this part of the case study, such estimation was not needed.

Concluding remarks

In this paper, a new methodology composed of an innovative evidence-driven decision modelling framework and a reasoning-based decision support process was proposed and investigated for the prioritization of the voices of customer. The features of data generated from both internal and external customer satisfaction surveys were analysed. Due to the different features of such surveys, it is necessary to transform survey data measured on various scales to a common scale. In this paper, several novel yet natural and pragmatic rule-based transformation functions were formulated and investigated, which provide a sound basis for rigorous data transformation that is flexible to incorporate expert judgments and domain specific knowledge systematically and consistently. These functions are the first of the kind that can capture domain knowledge and mimic expert judgement processes for transforming raw data to decision information in a pragmatic yet rigorous manner. The evidential reasoning approach was used to aggregate multiple assessments of a qualitative nature and handle various types of uncertainties such as ignorance caused due to missing data. A new generic ranking method was proposed that can be used to rank voices without making assumptions for the quantification of assessment grades unnecessarily, while the decision makers' preferences can still be incorporated into the ranking process if available. The case study illustrated the application of the decision modelling framework and decision support process to the prioritization of the voices of customer for a world-leading car manufacturer. The case study shows the scope and the potential of applying the methodology in a wider range of decision situations where large data sets and domain specific knowledge are available for informative and robust decision making.

It may be noted that a ranking of voices produced using the new methodology can identify the most important areas in which a manufacturer gets very good marks (features that should be kept) or very bad marks (features that should be improved). As a consequence, at the top of this single ranking these two different types of voices will emerge, without being distinguished if the data transformation functions discussed in the section 'Rule-based data transformation for evidence-driven decision modelling' are used. This issue may need to be addressed in implementing the methodology. There are a couple of ways to handle this issue. First, it can be demonstrated that the two types of rankings can be easily achieved by changing the transformation functions. Alternatively, based on the methodology developed in this paper, a software tool is developed. In that tool, engineers can easily trace back the ranking and

see if a voice comes out at the top because of bad marks or good marks. By tracing back, the engineers, given resource constraints, can decide how many voices with bad marks should be improved and how many voices with good marks should be kept as they are. The tool has been well appreciated by the engineers at the car manufacturing company as it enables several scenario studies by varying the weights for the data sources and criteria. The scenario studies not only improve the quality of analysis but also provide a structured and uniform way for analysing voices of customers.

Acknowledgements—This work was funded by General Motors Company (GM) under the Grant no: GM ISL-105/CL-06-08 and also partially supported by the UK Engineering and Physical Science Research Council (EPSRC) under the Grant no: EPSRC EP/F024606/1. The authors are grateful for the technical support to this research by the many staff members of the GM R&D labs in both India and the US. The detailed and constructive comments from the anonymous reviewers are greatly appreciated.

References

Besterfield DH, Besterfield-Michna C, Besterfield GH and Besterfield-Sacre M (2003). *Total Quality Management*. 3rd edn, Prentice Hall: New Jersey.

Büyüközkan G and Feyzioğlu O (2005). Group decision making to better respond customer needs in software development. *Comput Ind Eng* **48**: 427–441.

Chan LK, Kao HP, Ng A and Wu ML (1999). Rating the importance of customer needs in quality function deployment by fuzzy and entropy methods. *Int J Prod Res* **37**: 2499–2518.

Chan LK and Wu ML (2002). Quality function deployment: A literature review. *Eur J Opl Res* **143**: 463–497.

Chan LK and Wu ML (2005). A systematic approach to quality function deployment with a full illustrative example. *Omega* **33**: 119–139.

Chen CY, Chen LC and Lin L (2004). Methods for processing and prioritizing customer demands in variant product design. *IIE Trans* **36**: 203–219.

Chen LH and Weng MC (2003). A fuzzy model for exploiting quality function deployment. *Math Comput Model* **38**: 559–570.

Chen LH and Weng MC (2006). An evaluation approach to engineering design in QFD processes using fuzzy goal programming models. *Eur J Opl Res* **172**: 230–248.

Chin KS, Yang JB, Lam J and Guo M (2009). An evidential reasoning-interval-based method for new product design assessment. *IEEE Trans Eng Mngt* **56**: 142–156.

Dias LC and Clímaco JN (2000). Additive aggregation with variable interdependent parameters: The VIP analysis software. *J Opl Res Soc* **51**: 1070–1082.

Griffin A and Hauser JR (1993). The voice of the customer. *Market Sci* **12**(1): 1–27.

Grigoroudis E and Siskos Y (2002). Preference disaggregation for measuring and analysing customer satisfaction: The MUSA method. *Eur J Opl Res* **143**: 148–170.

Hazen GB (1986). Partial information, dominance and potential optimality in multiattribute utility theory. *Opns Res* **34**: 296–310.

Ho ESSA, Lai YJ and Chang SI (1999). An integrated group decision-making approach to quality function deployment. *IIE Trans* **31**: 553–567.

JD Power and Associates (2007). *Guide to Automotive Studies*. www.jdpa.com.

Keeney RL and Raiffa H (1993). *Decisions with Multiple Objectives*. 2nd edn, Cambridge University Press: Cambridge.

Kwong CK and Bai H (2002). A fuzzy AHP approach to the determination of importance weights of customer requirements in quality function deployment. *J Intell Manuf* **13**: 367–377.

Olson DL (1996). *Decision Aids for Selection Problems*. Springer-Verlag: New York.

Park KS (2004). Mathematical programming models for characterizing dominance and potential optimality when multicriteria alternative values and weights are simultaneously incomplete. *IEEE Trans Syst, Man Cybern—Part B: Cybern* **34**: 601–614.

Saaty TL (1994). *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*. RWS Publications: Pittsburgh, PA.

Temponi C, Yen J and Tiao WA (1999). House of quality: A fuzzy logic-based requirements analysis. *Eur J Opl Res* **117**: 340–354.

Wang YM, Yang JB and Xu DL (2006). Environmental impact assessment using the evidential reasoning approach. *Eur J Opl Res* **174**: 1885–1913.

Winston WL (1994). *Operations Research—Applications and Algorithms*. Duxbury Press: Belmont, CA.

Xu DL, Yang JB and Wang YM (2006). The evidential reasoning approach for multi-attribute decision analysis under interval uncertainties. *Eur J Opl Res* **174**: 1914–1943.

Yang JB (2001). Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainty. *Eur J Opl Res* **131**: 31–61.

Yang JB and Sen P (1994). A general multi-level evaluation process for hybrid MADM with uncertainty. *IEEE Trans Syst, Man Cybern* **24**: 1458–1473.

Yang JB and Sen P (1996). Preference modelling by estimating local utility functions for multiobjective optimization. *Eur J Opl Res* **95**: 115–138.

Yang JB and Singh MG (1994). An evidential reasoning approach for multiple attribute decision making with uncertainty. *IEEE Trans Syst, Man Cybern* **24**: 1–18.

Yang JB and Xu DL (1998). Knowledge based executive car evaluation using the evidential reasoning approach. In: Baines T-B and Zhao Z-X (eds). *Advances in Manufacturing Technology XII*. Professional Engineering Publishing: London, pp 741–749.

Yang JB and Xu DL (2002a). On the evidential reasoning algorithm for multiattribute decision analysis under uncertainty. *IEEE Trans Syst, Man Cybern—Part A: Syst Humans* **32**: 289–304.

Yang JB and Xu DL (2002b). Nonlinear information aggregation via evidential reasoning in multiattribute decision analysis under uncertainty. *IEEE Trans Syst, Man Cybern—Part A: Syst Humans* **32**: 376–393.

Yang JB and Xu DL (2005). *The IDS Multi-criteria Assessor Software*. Intelligent Decision System Ltd: Cheshire, UK.

Yang JB, Wang YM, Xu DL and Chin KS (2006). The evidential reasoning approach for MADA under both probabilistic and fuzzy uncertainties. *Eur J Opl Res* **171**: 309–343.

Received October 2009;
accepted May 2010 after one revision