This article investigates the concept of developing a game theoretic framework that is based on the application of buyer and seller utility functions to support the bidding process in government acquisition. The results of a literature survey of utility function approaches, with potential to provide a suitable foundation to a game theory framework for acquisition, are presented. The utility function methods found most promising were further adapted and tested: the Best-Worst method, the Multi-Swing Method, and Functional Dependency for Network Analysis. To test the scalability of the approach, the Best-Worst method is applied to a larger problem to show the extensibility on a government-relevant scale. The future application of these utility models is in support of a game theory framework that is envisioned to move bidding contractors closer to the government’s preferred negotiation point and expedite the decision-making process in government acquisition during competitive source selection.

DOI: https://doi.org/10.22594/dau.18-812.26.03
Keywords: Game Theory, Utility Theory, Optimization, Acquisition, Contract Negotiation
Game theory has been a research paradigm for studying conflict, bargaining, and negotiations for over 50 years. It is widely applied throughout the business domain to develop strategies that reflect priorities and tradeoffs. Bierman and Fernandez (1998) provide examples of successful game theory applications from business and industry. The U.S. Government has an opportunity to leverage game theory in the federal acquisition system to procure more preferable products and to increase the agility of the acquisition process as a whole, ultimately benefitting both government and industry. As programs become more technical and complex, game theory can help decision makers identify strategies and leverage information to make data-driven decisions that reflect government priorities and tradeoffs. This article considers the applicability of game theory to the competitive federal acquisition process to facilitate negotiation strategies, and provides research and testing of utility functions, which provide the cornerstone for a game theory framework. The research focus of this article is on utility functions for game theory models supporting acquisition and is a necessary first step in the development of an overall game theoretic framework for acquisition decisions.

The federal acquisition process is governed by a system of clearly defined rules and regulations codified in the Federal Acquisition Regulation (FAR, 2018). The codification and publication of, and adherence to, a uniform acquisition system establishes a common understanding or common knowledge of the rules of engagement. Common knowledge, a central tenet of game theory, encourages industry to develop and execute rational business strategies that differentiate solutions and reflect tailored cost, schedule, and performance tradeoffs. This creates the framework for achieving best value through competing strategies and decisions. As a rulebook, the FAR ensures fairness and transparency in the acquisition process for all players, including both industry and the government. Today’s structure and process of federal acquisition, coupled with the need for increased speed and agility, potentially provide the right environment for game theory to be effectively leveraged.
The focus of the game theory application in this article is in a competitive source selection acquisition where multiple vendors are bidding (or submitting a proposal) to provide a product that is a noncommercial product, still in the stages of development. Within a competitive source selection, multiple potential phases are in the acquisition life cycle where game theory has the potential to provide support. First, it can steer the decision makers towards the identification of the most meaningful program-specific evaluation criteria. Second, it enables a more objective and quantitative way to proceed through a source selection by better illuminating key attribute tradeoffs for the government. For developmental items, the government can use game theory to take control and drive the key features of the product during the competitive source selection.

"Common knowledge, a central tenet of game theory, encourages industry to develop and execute rational business strategies that differentiate solutions and reflect tailored cost, schedule, and performance tradeoffs.

In game theory, negotiations are facilitated by locating a solution that jointly maximizes the utility of each player. In government acquisition, there exists a game between each bidding vendor and the government. Utility for the government can be modeled through a utility function, which represents the government’s overall preference for a product. The utility function for the government is parameterized by the key attributes of the product being procured and maps levels of those attributes to a value of overall preference for the product. Utility functions for the bidding vendor can then focus primarily on the cost for achieving the product attribute levels in the government utility function solution. While generating their bid, vendors can then search for solutions that have attribute values that achieve a high value for the government’s utility function and result in low cost for their own utility function.

Federal source selections already adhere to a game theory configuration by the mandatory disclosure of evaluation criteria as key discriminators or attributes. By advertising its source selection criteria and relative order of importance, the government signals its tradeoff considerations. Vendors from industry also currently act like players in this type of game by tailoring and offering solutions to the government to meet these considerations.
Table 1 summarizes some of the similarities between game theory and the government source selection process, and shows how the government inherently implements several key aspects of game theory.

<table>
<thead>
<tr>
<th>Game Theory Principles</th>
<th>Government Source Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Players know the “rules of the game”</td>
<td>Federal acquisition process is governed by well-defined rules and regulations codified in the Federal Acquisition Regulation (FAR); both industry and the government (i.e., players) know the rules of the game.</td>
</tr>
<tr>
<td>Requires clear communication of attributes, priorities, and outcomes</td>
<td>Mandatory disclosure of evaluation criteria and relative importance ensures industry is informed of attributes (discriminators), priorities, and desired outcomes.</td>
</tr>
<tr>
<td>Players are rational and seek to maximize their expected outcome or utility</td>
<td>Government maximizes expected outcome or utility through best value tradeoff; industry seeks to maximize their objective function (e.g., profits, market share).</td>
</tr>
</tbody>
</table>

The current federal acquisition system follows a structured process outlined in the FAR to guide the government in navigating the complexities and cumbersome nature of the source selection process. The FAR presents no obstacles to adopting a game theory framework for acquisition support. Moreover, FAR Part 1.102, “Statement of Guiding Principles for the Federal Acquisition System,” outlines an opportunity to introduce the agility and efficiencies of game theory by allowing strategies, practices, or procedures that are in the best interests of the government (FAR, 2018). However, such strategies, practices, and procedures are restricted to those not specifically limited or prohibited by the FAR, Executive Order, or regulation. A game theory approach can also be consistent with FAR Part 15.305, “Proposal Evaluation,” as evaluations may be conducted using any rating method or combination of methods, including color or adjectival ratings, numerical weights, and ordinal rankings (FAR, 2018). Additionally, this approach is consistent with the “Department of Defense Source Selection Procedures” memorandum (DoD, 2016, Sec. 2.3), where evaluation criteria may be quantitative, qualitative, or a combination of both. Although numerical or percentage weighting of the relative importance of evaluation criteria may not be used in DoD, assigning quantifiable or value tradeoffs in evaluating an offeror’s proposal (found in the Subjective Tradeoff and Value Adjusted Total Evaluated Price [VATEP] source selection tradeoff approach) is allowable and harmonious with game theory.
Quantitative Decision Support in Acquisition

The game theory framework proposed herein consists first of generating a utility function parameterized by the key attributes affecting the acquisition decision. The output to this utility function is a level of preference for the government concerning a bid coming from a competing vendor. A bidding vendor uses this utility function as well in helping prepare their bid. Through an envisioned web-based application, each potential vendor can attempt to maximize the government’s utility function subject to their own cost function. Each vendor’s cost function is also parameterized by the same attributes or key criteria that parameterize the government’s utility function. This maximization can be performed with mathematical programming techniques to enable bids to be automatically generated. A general utility optimization formulation for each vendor specific to their own individual cost constraints follows.

\[
\text{Max } V(x) = \sum w_i v_i(x_i) \quad (1)
\]

subject to: \( \sum c_i x_i \leq B \)

where:
- \( x_i \) = is the level for attribute \( i \)
- \( w_i \) = the weight for attribute \( i \)
- \( v_i \) = the single attribute value function for the attribute \( i \)
- \( c_i \) = the offeror cost function for attribute \( i \)
- \( B \) = budget constraint for maximizing utility \( i \)

A known challenge is that Request for Proposal (RFP) language can make it difficult for potential bidders to extract what the most important attributes are for the government, especially when too many attributes are included or the evaluation criteria is unclear. Using this quantitative mathematical programming formulation allows bidders to move directly towards those key attributes through an automated and objective means, and increase transparency into the bidding process. The solutions to the optimization formulation in equation (1) provide stronger initial bids by the interested vendors. These solutions aren’t necessarily final solutions or final bids, but are an efficient means to generate an initial bid close to what the government is seeking. This can more efficiently set up the next stage of proposal tweaking and negotiation on both sides, resulting in streamlining the acquisition cycle.
Applicability of Game Theory and Utility Theory for Acquisition Support

While government solicitations or RFPs help identify the evaluation criteria and relative importance, the qualitative, subjective, and imperfect nature of evaluations may result in suboptimal tradeoff and selection analysis. A game theory framework with a utility theory foundation provides an objective and mathematical framework to provide the government with insight into potential attribute tradeoffs, as well as improve the transparency of the source selection process. Moreover, through integrating the government’s utility function with utility functions of cost established by each vendor, a process for generating better initial bids can be established that gets the government and bidding vendor at a much better starting point for negotiation. This high-level framework was originally suggested in Simon and Melese (2011) and is extended further in this research. The focus of the extension in this article is a further delineation of a methodology for better calibrating and applying the government’s utility function. Other extensions are integrated into the high-level framework discussion that follows and are also discussed further at the end of the article with future areas of research:

- Formulate the government’s preferences into a utility function parameterized by critical noncost attributes, criteria, or discriminators.
- Publicize government utility functions in an RFP for industry to formulate and submit bids that maximized the government’s utility function.
  - Each vendor calibrates their own cost function parameterized by the same attributes as the government’s utility function.
  - Software applications with embedded mathematical programming solvers automate the process of maximizing the utility function and generating the optimal bid for the vendor.
- Within its own web-based application, the government can view bids from all vendors simultaneously with respect to utility and cost. The government can drill into preferred utility/cost values and then examine possible attribute tradeoffs that can be used in the next phase of negotiation.
This automated optimization framework enables vendors to provide bids quickly and across a range of budget levels. Moreover, incoming bids from vendors can be viewed as a Pareto curve capturing utility as a function of cost. This leaves the government the ability to do tradeoffs between attribute values as well as between cost and attribute values to help provide a counter offer to the bids that are submitted.

A wealth of research questions must be addressed to design and implement this type of framework in a real-world acquisition scenario, the first of which pertains to the type of utility model and the method for calibrating the model to use as the foundation of such a framework. This article focuses on the utility models that can be used to best support this proposed framework.

Survey of Game Theory and Utility Theory Literature Relevant to Acquisition

The survey is focused on three topics: Multi-Criteria Decision-Making, Utility Theory, and Game Theoretic Applications to Government. As Multi-Criteria Decision-Making is needed to provide utility functions, which are needed to allow for the application of game theoretic approaches, the findings are then in this order. As previously implied, this selection is neither complete nor exclusive.

Literature on Multi-Criteria Decision-Making. As noted, a critical and challenging task in the acquisition process is the selection of evaluation criteria that reflects the preferences of the government. Wallenius et al. (2008) provide a good overview of the various multiple Multi-Criteria Decision-Making and Multi-Attribute Utility Theory methods currently in use. Such methods include the Analytical Hierarchy Process (AHP), Evolutionary Multi-Objective Optimization (EMO), or Multi-Objective
Linear Programming (MOLP), placing them into a historical perspective, as the same author group conducted a similar review in 1992 as well. Their overview is written from a management perspective. They also provide insight on how to apply their findings in the form of Multi-Attribute Utility, which connects their work to the next subsection. A more engineering-leaning perspective is given, among others, by Parnell, Driscoll, and Henderson (2011).

Velasquez and Hester (2013) conducted a literature review and analysis of multi-criteria methods. They observe that outranking methods, which were prevalent in early approaches, were overtaken by value measurement approaches. Further, they show that deficiencies can be overcome by combining approaches, although this requires a clear understanding of the advantages and disadvantages of the individual approaches, which are captured in a summarizing table in their conclusion section. Their paper assessed the more common methods of Multi-Criteria Decision-Making to benefit practitioners in choosing a method for solving a specific problem, and they state clearly that this can only be the first step in selecting the right approach.

Agarwal, Sahai, Mishra, Bag, and Singh (2011) provide an alternative viewpoint on the selection of the best Multi-Criteria Decision-Making method by focusing on the proper evaluation and selection of suppliers, which is highly relevant in acquisition as well. An additional insight provided by them is the need to evaluate the suppliers based on the inputs of the strategic, functional, and operational levels. They present that the “implication of lean manufacturing and popularly used JIT (Just in Time) approach has forced the researchers to shift the focus from the
efficiency-based model to a quality-based approach. The single criterion approach of the lowest cost supplier is no more accepted in this challenging and continuously changing environment. Thus, price or cost shifted down the line with respect to its importance in evaluating the suppliers, while the quality and delivery performance climbed up the hierarchy” (p. 808). This insight is relevant for the government as well, and needs to be addressed in the selection of the appropriate methods.

Both recent literature reviews show that a universal best solution does not exist, but that the selection of the best method is determined by the problem and may even require the use of a problem-specific hybrid solution that requires an in-depth knowledge of the problem as well as of the tools and approaches.

**Literature on Utility Theory and Utility Functions.** The literature highlighted in this section strictly pertains to the formulation of utility functions to reflect the preferences of decision makers. Slantchev (2012) defines preferences and utilities to support decision-making, including those to be made under uncertainty. As he is writing for political scientists, explanations and examples are easy to follow and do not require an in-depth education in game theoretic mathematical foundations.

If data are available that reflect preferences of earlier decision-making processes for either side of the negotiating partners, the methods and algorithms described by Afriat (1967) are still relevant. The application of big data methods supporting utility function definitions is a topic of ongoing research with no predominant methods emerging thus far, although it was initially believed that more literature would be found.

An interesting variant for multi-issue closed negotiations addressing multi-time as well as multi-lateral negotiation strategies is described by Matsune and Fujita (2017), who developed not only the concept, but demonstrated it in an agent-based simulation environment. Theoretically, nothing speaks against applying these ideas for acquisition-specific challenges as well, but no applications in this domain within the survey were found. What makes the application described in this paper so interesting is the ability to learn the opponents’ utility information from observing bidding choices within a strategy.

While the mathematics behind utility theory and utility functions is well understood, how to elicit the knowledge about their preferences from decision makers is still a challenge in itself. Our survey of the literature did not reveal any predominant strategy. And this is a challenge the government must overcome in order to successfully apply game theory.
Literature on Game Theoretic Applications for Government Solutions. Obviously, every game theoretic insight has potential to be applied to better support government solutions, but two of the evaluated papers deserve special attention, as they directly apply game theory to acquisition and government decision-making.

Levenson (2014) provides an overview of the constraints of DoD procurement, showing why typical solutions from commercial markets are often not applicable and lead to undesired and unforeseen results. He describes the effects of fixed price and competitive price contracts, and concludes that “only when one or more competitors offer innovations that truly reduce the costs of development and production does the government substantially benefit from competition over sole-source procurement without the adverse side effects of cost overruns. Distinguishing between true innovation and optimistic cost estimating, however, can pose a challenge for DoD acquisition officials” (p. 437).

Blott et al. (2015) compiled a set of auction- and game theory-based recommendations for DoD acquisition by synthesizing literature into specific military acquisition categories: procurement with unknown cost and no risk, items with known costs and existent but understood stochastic risk, and items with unknown costs and/or unknown stochastic risk. Some examples further evaluate if multiple competing vendors participate, and if the lot is to be procured from several bidders, potentially at different stages of the project.

In summary, the literature survey provided some opportunities for immediate applications, but also revealed the need for continued research. The most pressing needs pertain to how to elicit preferences from decision makers and apply these utility function methods in an environment under the special constraints of government acquisition. The utility model approaches selected as having the most potential are described in detail in the upcoming section of this article, along with recommendations.

A known challenge is that Request for Proposal (RFP) language can make it difficult for potential bidders to extract what the most important attributes are for the government, especially when too many attributes are included or the evaluation criteria is unclear.
for preference elicitation to support each of these methods. These three utility function methods, with their corresponding preference elicitation procedures, are then tested and those results are presented later in this article.

**Research Methodology**

The methodology for this research first consisted of identifying three candidate approaches for utility function calibration and adapting them to be applicable to a government acquisition setting. The methodologies were first tested on a hypothetical scenario relating to the purchasing of a cell phone. The proposed utility functions were assessed by their goodness of fit and the time taken to perform the assessment procedure. After testing the three proposed methods in this setting, the method performing best with the two established measures was then applied in a realistic government acquisition setting.

The more realistic government acquisition scenario involved a significantly larger number of attributes. The selected method was then integrated with a screening and binning method to make it adaptable to a large number of attributes. Results from this experimentation are also provided in this article, which include the assessment time and goodness of fit of the resulting utility model.

For each of the two experiments that were performed, two experienced acquisition professionals from government were used as test subjects. For the first experiment, each subject tested the three different utility function approaches proposed in a private test setting. Even though the test subjects were experienced professionals, the first acquisition scenario involved the hypothetical purchase of a cell phone. The subjects were provided a background information sheet explaining the attributes in detail relating to the hypothetical product as a supporting testing instrument. Person-to-person interviews were then performed to conduct the assessment procedures of each of the utility function methods being tested. The standard protocol was that the experimentation was done in a private room with only one other team member in the room present to record responses. The subjects were told that the information they provided regarding their answers to the utility function assessment procedures would be protected and held as confidential.

For the second experiment, the test was performed on a real-life acquisition currently underway within the test subjects’ organization. The main limitation of both of these experiments is in the small sample size. Although
the sample size is too small to provide any statistical relevance, observed implementation by the test subjects shows the feasibility of a utility theory and game theory approach for the acquisition community, and the areas where future research is needed.

The remainder of this section provides the more detailed descriptions of the candidate approaches and their application in this setting. The three candidate approaches were selected from the methods covered in the literature review. The three chosen were selected due to their perceived potential being applied in an acquisition setting with qualitative adjustments to question phrasings and responses, and the potential to be calibrated through an assessment procedure that is not burdensome or time consuming.

- **Best-Worst Method** (Rezaei, 2015)
- **Multi-Swing Method** (Schmidt, 2017)
- **Functional Dependency for Network Analysis** (Garvey, Pinto, & Santos, 2014)

Testing of these methods in the next section will further reveal the features of the acquisition scenarios where each approach does well. In this research, all three approaches are applied to a small decision problem that simply involved five attributes to get the decision maker(s) accustomed to the steps and procedures needed to be conducted. Furthermore, the Best-Worst method was applied to a larger, more complex 20-attribute problem, representative of a major program (developmental item). These test cases are discussed in more detail in the next section.

Preliminary collaborations and discussion with a government sponsor helped to derive three best practices or considerations that impacted our utility function assessment procedure and resulted in the application of multiple assessment techniques.

The first is that the level of effort in developing the assessment procedure must be commensurate with the size, scope, and complexity of the acquisition. A day or longer interview process to fit a model may be realistic for a highly complex multibillion-dollar Acquisition Category (ACAT) I program, but is not realistic for all acquisition. The assessment procedure must be able to accommodate a more expedient timeframe for smaller or less complex acquisition, but have the flexibility to incorporate a higher fidelity of information for an assessment occurring over a longer period.
of time. An initial assessment procedure drawn out longer than expected may result in fatigue and complacency, which may lead to inconsistencies in preference articulation.

The second consideration is that assessment procedures must be adaptable so that they can be effectively applied to decision makers who are either more quantitative or qualitative in nature. Our research showed that most acquisition professionals are comfortable with relative importance and prefer qualitative descriptions of their preferences. Introducing descriptive adjectives in place of numerical values, in many questions, helped alleviate this issue.

The third and final consideration is that the assessment procedures must be applicable in the many acquisition situations where a large attribute set influences the decision. The size of this attribute set is often overwhelming for any decision maker in acquisition. Therefore, preference modeling methods must be able to screen out attributes of minimal significance to isolate the critical noncost attributes and the critical tradeoffs between those attributes. This supports an acquisition best practice of focusing on critical noncost attributes/evaluation factors and subfactors to avoid diluting the importance of key discriminators.

**Best-Worst Method Implementation**

The Best-Worst method originates from Rezaei (2015), and this research has extended the approach to work more smoothly for cases where many attributes are at hand and when the attributes are binary in nature (result in either a 0/1 or yes/no value). One of the Best-Worst method’s features is its ability to perform calibration in a short series of questions. Moreover, these questions can be phrased such that they are not overly burdensome to the decision maker(s). From our observations, having simple and clear acquisition questions to identify key discriminators facilitates the acquisition and conforms to best practices.

Consistent with source selection practices, the procedure for the Best-Worst method starts with selecting the attributes or discriminators that affect the decision. Then feasible ranges are assigned for each of these attributes. The next step is the assignment of weights for each attribute reflecting the preferences and importance. This applies specifically to the attribute to identify key discriminators and does not apply numerical weights to proposals in the source evaluation process. This step begins with selecting the most important attribute as well as selecting the least important attribute. From there, comparisons are made to understand the relative
importance of the most important attribute to each of the other attributes. In a similar manner, comparisons are then made to assess the relative importance of the least important attribute to each of the other attributes.

The question phrasing to the decision maker is the key to getting this approach to work effectively. The decision maker needs to be directly asked how much more important the most important attribute is with respect to each of the other attributes individually. Mapping qualitative scales to numerical scales was shown to work well in our studies for preserving rank order. For instance, levels, such as “just as important,” “slightly more important,” “more important,” “significantly more important,” and “extremely more important” were applied with good success while being mapped on a scale of [1-5].

The completion of the Best-Worst assessment procedure is to obtain a preference function in the form:

\[ V(x) = w_1 v_1(x_1) + w_2 v_2(x_2) + \ldots + w_n v_n(x_n). \]

This is consistent with the general form of the utility function presented in Equation 1 with \( x_i = \) the level for attribute \( i \), \( w_i = \) the weight for attribute \( i \), and \( v_i = \) the single attribute value function for the attribute \( i \). The Best-Worst procedure primarily focuses on the weights. Suggestions in this article for extending to the assessment of the single attribute utility functions \( v_i(x_i) \) focus on fitting a function across sample points for each individual attribute. Sampling can be effective with just four points on the utility curve. When doing a qualitative mapping, those points can be referenced as the min, midpoint, target, and max. On a [0,1] scale, those reference points were mapped to values of 0, 0.5, 0.75, and 1 respectively. The qualitative assessment questions can first focus on the target. Here the question is asked: “What is the value of this attribute that you would really want to have?” Then the level representing satisfactory for the attribute is assessed: “What level for this attribute is acceptable and would not hinder my use? It can be considered being like a minimum requirement that is not ideal
but gets the job done.” Then the max level for the attribute can be assessed: “What is the level for the most functionality that you could possibly handle/need—any more wouldn’t make life any better?” Finally, the minimum level for the attribute is assessed: “What is the maximum attribute level where there is zero utility or where you would have absolutely no use for this product if this attribute was at this level?”

The Best-Worst method was extended to be more applicable to acquisition consisting of a large number of attributes (>20). In acquisition, we observed with our government sponsors that the number of attributes was typically quite large. For a large number of attributes, the procedure was updated in the following manner:

1. Perform pairwise comparisons across adjacent pairs of attributes starting at attribute No. 1 and then work down the attribute list.

2. Bin the attributes based on whether the attributes were more important than two attributes, one attribute, or no attributes. End up with three bins: Prime, Mid, Low.

3. Reassess attributes in each bin to make sure they are in the right place.
   a. Ask for best and worst for each bin.
   b. Do pairwise comparison of best in mid and low bin, with worst in the higher level bin.
   c. Repeat 3a and 3b until no more changes are made.

4. Identify the attributes for inclusion into the Best-Worst method.
   a. Take all attributes in prime bin.
   b. Take best and worst in mid bin.
   c. Take best and worst in low bin.

5. Best-Worst method is then implemented on attributes in the prime bin.

6. Best-Worst method is then implemented on all other attributes kept after completing step 5.

7. Ask the level of difference between the worst attribute in prime and the best in mid. This level of difference then becomes the difference level for weights in the prime bin, and weights in the remaining bins are then scaled accordingly.
After these assessment procedures are made, the weights for the preference function can be solved through the optimization formulation outlined in Rezaei (2015). The pairwise comparisons given at the beginning of the assessment procedure can also be used to solve for the weights more effectively as well as for validation of the results.

**Multi-Swing Rollup Method Implementation**

The Multi-Swing Rollup Method (MSRM) is a new aggregation method for multi-attribute decision problems (Schmidt 2017). As previously discussed in the literature survey, rolling up multiple values into one representative value is a general challenge, as already discussed in our literature survey. The MSRM uses a Generalized Addition Tallying Organization (GATO) approach. While the classical approach uses the weighted sum of the contributing decision factors, MSRM/GATO uses a nonlinear combination in areas in which the simple addition leads to counterintuitive results.

MSRM starts with the definition of Multi-Swing tables to collect data and combine weights and utility functions into one user-driven process. These Multi-Swing tables are then multiplicatively rolled up and calibrated to fit to a percentage scale. The four steps of the methods are:

1. Selecting and quantifying the metrics for each contribution
2. Defining a scale for quantifying the overall score
3. Constructing the Multi-Swing tables for each contribution
4. Constructing and calibrating the rollup function

Selecting and quantifying the metrics for each contribution starts with identifying the qualities in which the user is interested. The result is a quality tree that identifies the contributions and the metrics used to quantify them. Then, defining a scale for quantifying the overall score of the attributes (not numerical scoring of proposals) ensures consistency when assessing the overall value increase or decrease when evaluating the individual contributions. MSRM recommends using a mapping of generally understood expressions to numerical values, such as ideal = 100%, very good = 90%, good = 70%, indifferent = 50%, poor = 30%, very poor = 10%, and not acceptable = 0%. The scale does not have to be symmetric. More important is that it reflects the weighting priorities and preferences of the user.
Constructing the Multi-Swing tables is conducted for each contribution, starting with defining a baseline with typical and acceptable values for each contribution. For each contribution, we next define a set of swing scores that can be better or worse than the baseline. For each contribution, a set of swing scores spanning all values that can occur in the selection process is collected and the swing rows constructed. If the value of a contribution is a showstopper, e.g., the battery life is too short to support operational use of the item, it is marked as such. In acquisition settings, every attribute that falls under a minimal value becomes a showstopper or an immediate rejection of the proposal.

The baseline and all swing rows are then captured in one Multi-Swing table. In this table, only one of the values in each row is changed in comparison with the baseline, so that a comparison with the baseline can be used to access an overall score using the expressions identified in step 2 of the MSRM. While the baseline may be seen as good, a lesser screen resolution may decrease the value to indifferent, poor, or may even become a showstopper, while longer life span may result in a very good overall value.

Constructing and calibrating the rollup function uses the Multi-Swing tables as its foundation, as each row in the Multi-Swing table captures how the overall value changes when we swing one contribution at a time. A multiplicative roll-up approach can now be applied to compute how the value changes when several of such changes occur at the same time. If, e.g., the resolution decreases, resulting in a change value decrease of 20%, and the battery life gets shorter as well, decreasing the value by 10%, then the occurrence of both changes should result in a decrease of 28%. The idea is to multiply the individual effects to generate the combined effects.

While the approach naturally results in the elimination of all showstoppers (as the multiplicative approach results in a zero whenever one of the contributions is not acceptable), the positive results can multiply up to more than 100%, which can be addressed using rescaled proportion retention multipliers that ensure that no combination exceeds the 100% limit.

One of the remaining challenges is the combinatorial explosion with the increasing number of contributions. Our initial application shown in the next section was limited to five attributes, but still required more than 45 minutes to build all Multi-Swing tables. On the positive side, the method allows the linear integration of new contributions after the initial set-up: a new attribute can be integrated without having to change the tradeoffs between the already existing attributes.
Functional Dependency for Network Analysis Implementation

The last approach utilized in our research was originally developed for a systems engineering setting, but due to its general applicability, we decided to include it in our evaluation. The application of the Functional Dependency for Network Analysis (FDNA) methodology involves:

1. Data gathering
2. Preference inference
3. Quantifying accuracy
4. Making predictions

The data gathering step involves constructing an experimental design to capture data on the different attributes of the product in accordance with the decision maker’s input. The preference inference step involves proposing specific preference models and using the gathered data to infer the defining parameters that are most consistent with the data. The quantifying accuracy step involves the application of cross-validation to assess the accuracy of the fitted preference model. Finally, the making prediction step entails converting a test case to the form selected in the first step and making predictions using the parameters inferred in the second step.
For data gathering with FDNA, it is necessary to create a dictionary of qualitative descriptions of product attributes and an assigned numerical representation to each. In the acquisition setting, the dictionaries are highly reusable according to our experiences, although no study has been conducted to verify this observation. As a general practical matter, many spirals of potential dictionaries should be generated and tested to ensure that the definitions are neither too narrow nor too broad so that the decision maker who is being modeled will be assigning a broad range of numerical preference scores to the anticipated set of optional designs. Then a set of optional designs of interest can be generated. Assuming the absence of a priori knowledge of the decision maker’s preference, the designs are randomly generated.

Motivated by the work of Garvey and Pinto (2009) and Servi and Garvey (2017), two different preference models are included in the approach. (Note that if there was a larger amount of experimental data, it would have been more desirable to use the precise FDNA model documented in the references. Due to the limited size of the data, two different aspects of the FDNA were used for this analysis.)

\[
f_s(P_i; \gamma) = \min_i [P_i + \beta_i] \tag{2}
\]

or

\[
f_s(P_i; \gamma) = \alpha_0 + \sum_i a_i P_i \tag{3}
\]

where \(P_i\) is the numerical level of preference of the \(i^{th}\) characteristic of the \(s^{th}\) experiment.

With the decision maker’s evaluation of different attribute combinations—the values of \(\beta_i\) or \(a_i\) computed using the training data, which are rows comprising all attribute values—and the resulting evaluation by the decision maker, estimating the preferences of the decision maker is possible.

One side result from the testing discussed in the next section is that the accuracy using equation 2 was found to be superior to that using equation 3. Therefore, it is recommended that when predicting the comparative preferences of the decision maker to two alternatives, the prediction is made using only equation (2) and, in the case of a tie, equation (3) is used to break the tie.
Experimentation with Utility Function Methods

The three candidate utility model assessment procedures were tested on two scenarios to test their applicability in the acquisition setting. Two acquisition specialists from a government sponsor supported the tests. The first scenario, described in the following discussion, is a cell phone purchasing example for initial testing of all three selected approaches. The second scenario, described in the next section, pertains to a real-life acquisition scenario, which adds complexities not existing when utility function methods are tested and presented in literature. The candidate performing best in the initial cell phone test scenario is then adapted for use to the real-life acquisition scenario described in the following section.

The test subjects were given the scenario of purchasing a new smartphone. The test instrument included a two-page description of the purchasing options and a description of the product attributes affecting the purchasing decision, which was provided to the subject. The five attributes that were presented as affecting their decision follow:

A. Weight [0 – 5 pounds]
B. Lifespan [0 – 10 years]
C. Screen resolution [0 – 4,000p]
D. Processing speed [0 – 10x]
E. Storage amount [0 – 500 Gigs]

Included in these five attributes are the ranges of values that each attribute could take during the test. After the subject read through the example and took the role of the purchaser, a series of questions was conducted via a person-to-person interview about their preferences in accordance with the assessment procedures for the three preference functions tested. The study was limited to two test subjects. The first test subject was used to help refine the question phrasing. The second test subject was an acquisition professional, and the results from their experimentation were used to compare each of the three methods. The test subjects were informed that all answers they provided regarding hypothetical preferences would be kept confidential and would not be released in any presentation of the results.
Testing of Best-Worst Method

The testing of the Best-Worst method began with the assessment of the weights for the five attributes of interest. The test subject was asked a series of tradeoff questions and identified processing speed as the most important attribute and lifespan as the least important attribute. Through a series of questions relating the level of importance of each attribute with respect to processing speed, the following vector was obtained representing how much more important processing speed attribute is with respect to each of the other attributes on a scale of 1 to 5: \( A_b = [5, 4, 1, 1, 2] \). The numerical values in this vector specify the relative importance of processing speed with respect to weight, lifespan, screen resolution, and storage amount. Table 2 depicts this in tabular format. These are the same mappings to numerical values introduced in the previous section.

<table>
<thead>
<tr>
<th>Attribute of Interest</th>
<th>Vector Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>5</td>
<td>Processing Speed “Extremely more important” than Weight</td>
</tr>
<tr>
<td>Lifespan</td>
<td>4</td>
<td>Processing Speed “Significantly more important” than Lifespan</td>
</tr>
<tr>
<td>Screen Resolution</td>
<td>1</td>
<td>Processing Speed “Just as important” as Screen Resolution</td>
</tr>
<tr>
<td>Processing Speed</td>
<td>1</td>
<td>Reference attribute defined against itself (i.e., Just as important)</td>
</tr>
<tr>
<td>Storage Amount</td>
<td>2</td>
<td>Processing Speed “Slightly more important” than Storage Amount</td>
</tr>
</tbody>
</table>

After this, the same questions regarding relative attribute importance were asked with respect to the attribute noted as the least important. This resulted in the following vector representing how much more important each attribute is with respect to the lifespan attribute (declared least important) on a scale of 1 to 5: \( A_w = [1, 1, 4, 5, 3] \). Following the approach outlined in Rezaei (2015), these two vectors representing relative importance between each attribute and the best and worst attribute, respectively, can be used to perform a least squares approximation to solve for the weights. The following numerical weights were obtained: \([0.04, 0.12, 0.28, 0.39, 0.17]\)

The final step was to solve for the single attribute value functions pertaining to each attribute. Here, the test subject was asked for each attribute to specify the min, midpoint, target, and max values for each of the five
attributes, and the question wordings introduced in the previous section were applied. Table 3 presents the values obtained for the mid, midpoint, target, and max for each attribute.

| Table 3: Subject Responses to Min, Midpoint, Target, and Max Levels for Each Attribute |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Min** | **Midpoint** | **Target** | **Max** |
| Weight | 1 lb. | 0.75 lb. | 0.5 lb. | 0.33 lb. |
| Lifespan | 1 year | 3 years | 4 years | 6 years |
| Screen Resolution | 400p | 720p | 1080p | 2000p |
| Processing Speed | 0.5x | 1x | 3x | 5x |
| Storage Amount | 64 gigs | 128 gigs | 250 gigs | 500 gigs |

The values in Table 3 are used to solve for the single attribute value functions \( v_i(x_i) \) for all five attributes in this test. For this application, a 2nd order polynomial was applied for fitting these single attribute value functions, and the method of least squares was used for fitting. The weights for all five attributes were then integrated into the single attribute value functions to obtain the following function for the preference model:

\[
V(x) = w_1v_1(x_1) + w_2v_2(x_2) + w_3v_3(x_3) + w_4v_4(x_4) + w_5v_5(x_5)
\] (4)

To test the accuracy of the Best-Worst method, a series of comparisons across six purchasing alternatives was performed. The following pairwise comparisons across all combinations of the purchasing options were performed by the decision maker.

A. [2, 2, 2000, 0.75, 256]
B. [0.5, 5, 720, 2, 128]
C. [4, 1, 4000, 1, 64]
D. [1, 4, 720, 1, 256]
E. [2, 3, 1080, 2, 256]
F. [0.5, 3, 4000, 4, 64]
The results of these pairwise comparisons were applied to generate a ranking. The rankings obtained here were compared to rankings generated through the preference model sampled under these same alternatives. In addition, the proportion of the pairwise comparisons that was consistent between the decision maker and model was measured. Fifteen different combinations of pairwise combinations resulted from the six purchasing alternatives previously shown. The preference model resulting from the calibration involving the Best-Worst method resulted in consistency amongst all 15 pairwise comparisons. This meant that when the subject specified, for example, that alternative B was more preferable than alternative A, that the preference model outputted a larger value when inputting the attribute levels for alternative B than when inputting the attribute levels for alternative A. As naturally follows, the rankings for all six alternatives were consistent as well. The test demonstrated promise in the Best-Worst method to generate an accurate model in a short amount of time.

The use of a software tool for calibration streamlines the process, and also ensures both repeatability and scalability that can be extended to more complex acquisition to increase the expected outcome.

The entire assessment procedure was done in roughly 45 minutes for this scenario involving five attributes and consisted of a single decision maker articulating preferences. Multiple decision makers participating in the utility model calibration is possible if they can provide an agreed-upon answer for each question in the procedure. The process can be supported by graphical user interface (GUI), and our research additionally resulted in the development of a tool design to support data collection, calibration, and presentation of the results. The use of a software tool for calibration streamlines the process, and also ensures both repeatability and scalability that can be extended to more complex acquisition to increase the expected outcome. Future researchers seeking to enhance this method may consider extending the assessment procedure to explore the impacts of integrating interaction effects between attribute preferences. These interaction effects between attributes are not captured by the current form of this utility model, but with the support of a software tool, more complex assessment procedures to obtain interaction effects are obtainable. In various real-world scenarios, dependencies between attributes do exist where one needs, for example, both attributes A and B to be high; otherwise, both should be
low relative to C. Decision makers don’t always immediately reveal or are cognizant of these situations, but through a well formulated assessment procedure and sophisticated human-computer interaction design, this could be achievable within an assessment procedure of under 1 hour.

**Testing of MSRM**

The Multi-Swing Rollup Method was applied in the same setting as the Best-Worst method, using the same test subjects to conduct the experiments. Using the same attributes as enumerated in Table 3, we defined one positive and one negative swing for each attribute, as shown in Table 4.

| TABLE 4. ATTRIBUTES AND SWING STATES (GREEN VARIATIONS ARE POSITIVE, RED ARE NEGATIVE) |
|---------------------------------|-----------------|-----------------|
| Baseline | Variations |
| (W) Weight | 0.5 | 0.33 | 0.75 |
| (LS) Lifespan | 2 | 1 | 4 |
| (SR) Screen Resolution | 1080 | 720 | 2000 |
| (PS) Processing Speed | 4 | 2 | 5 |
| (SA) Storage Amount | 256 | 128 | 500 |

Next, we defined the utility factor terms to be used to rate the comparisons between the baseline and the swings. In the discussion with the experts and decision makers, we ended up with a table with which the test group was comfortable that showed the semantic equivalencies between different families of terms describing comparisons, status descriptions, and grades.

Having five attributes with two swing states results in 11 entries. Using the utility terms, each entry was compared individually with the baseline to identify the overall change in utility by changing one attribute. Table 5 shows the individual utility contributions in percent that resulted from our discussions with the test subjects.
**TABLE 5. UTILITY VALUE CHANGES IN PERCENT RELATIVE TO THE BASELINE**

<table>
<thead>
<tr>
<th>Utility Value Comparison</th>
<th>Utility Value Status</th>
<th>Utility Value Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>significantly better</td>
<td>ideal</td>
<td>A</td>
</tr>
<tr>
<td>much better</td>
<td>very good</td>
<td>A/B</td>
</tr>
<tr>
<td>better</td>
<td>good</td>
<td>B</td>
</tr>
<tr>
<td>little bit better</td>
<td>above average</td>
<td>B/C</td>
</tr>
<tr>
<td>average/mid-point</td>
<td>average</td>
<td>C</td>
</tr>
<tr>
<td>little bit worse</td>
<td>below average</td>
<td>C/D</td>
</tr>
<tr>
<td>worse</td>
<td>poor</td>
<td>D</td>
</tr>
<tr>
<td>much worse</td>
<td>very poor</td>
<td>E</td>
</tr>
<tr>
<td>showstopper</td>
<td>showstopper</td>
<td>F</td>
</tr>
</tbody>
</table>

Using this information, the full Multi-Swing rollup table with all entries (243 in this case) can be created. The resulting table contains all possible combinations of Multi-Swings plus the baseline. The resulting overall utility is calculated by multiplying the individual changes. When ordering the table, the entry with all negatives obviously is the lowest, and the entry with all positives the highest, but all possible permutations in between are listed as well, showing the ranking of all alternatives, including the selected subset used in the Best-Worst method. This shows that there is at least consistency in the preference articulation since the same ranking was derived between these two methods.

The assessment procedure conducted with our decision makers took less time than for the Best-Worst method, but only because several of the results could be reused. In an internal comparison with in-house experts, the amount of time needed for the first two methods was approximately the same for the cell phone example.

**Testing of the FDNA Method**

The FDNA method consists of two different preference models that were motivated by the work of Garvey and Pinto (2009) and Servi and Garvey (2017). First, the term needs to be defined. The dictionary shown in Table 6 is comparable to the terms defined in Table 7 for the utility terms used in the MSRM.
TABLE 6. DICTIONARY ASSIGNING NUMERICAL PREFERENCE LEVEL TO PREFERENCE LEVEL OF QUALITATIVE CHARACTERISTICS OF CELL PHONE

<table>
<thead>
<tr>
<th>(W) Weight</th>
<th>(LS) Lifespan</th>
<th>(SR) Screen Resolution</th>
<th>(PS) Processing Speed</th>
<th>(SA) Storage Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - Bad</td>
<td>heavy</td>
<td>good images</td>
<td>email, word OK</td>
<td>some added apps, best photos</td>
</tr>
<tr>
<td>1 - OK</td>
<td>not heavy</td>
<td>good for printing</td>
<td>very good nongames, slow video</td>
<td>apps and photos</td>
</tr>
<tr>
<td>2 - Good</td>
<td>light</td>
<td>great images, good enlarging</td>
<td>good for video</td>
<td>huge for apps and photos, some videos</td>
</tr>
<tr>
<td>3 - Great</td>
<td>ultralight</td>
<td>very good for enlarging</td>
<td>everything great</td>
<td>virtually unlimited including videos</td>
</tr>
</tbody>
</table>

TABLE 7. UTILITY VALUE TERMS

<table>
<thead>
<tr>
<th>Weight</th>
<th>Lifespan</th>
<th>Screen Resolution</th>
<th>Processing Speed</th>
<th>Storage Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative</td>
<td>-6.7</td>
<td>-26.7</td>
<td>-26.7</td>
<td>-46.7</td>
</tr>
<tr>
<td>positive</td>
<td>13.3</td>
<td>20.0</td>
<td>33.3</td>
<td>33.3</td>
</tr>
</tbody>
</table>

Next, we generated possible solutions for the five attributes important for the selection of the cell phone: weight, lifespan, screen resolution, processing speed, and storage amount. Table 8 shows the 27 generated cases, using the index numbers defined in the dictionary to specify the solution. The decision maker then graded the various solutions as captured in the column “evaluator.”
<table>
<thead>
<tr>
<th>No.</th>
<th>Weight</th>
<th>Lifespan</th>
<th>Screen Resolution</th>
<th>Processing Speed</th>
<th>Storage Amount</th>
<th>Evaluator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

It is possible to search exhaustively for the integer values of $\beta_i$ most consistent with the data in terms of the mean sum of squares error, $(\alpha_0 = 1, \alpha_1 = 8, \alpha_2 = 6, \alpha_3 = 0, \alpha_4 = 0)$ as well as analytically solving for the values of $\alpha_i$ most consistent with the data, $(\beta_0 = -1.6204, \beta_1 = 0.2101, \beta_2 = 0.2219,$
$\beta_3 = 0.4290, \beta_4 = 0.4375, \text{ and } \beta_5 = 0.5724$. This leads to mean sum of square error of 0.19 when using equation (2) and a worse mean sum of square error of 0.44 when using equation (3).

For FDNA, however, the more precise approach to quantifying the error is using the method of cross-validation. Here, the values of $\beta_i$ or $\alpha_i$ are computed using a random set of $8/9$ of the data in Table 8, and then the accuracy of the prediction is computed using the $1/9$ of the data not trained on. This was repeated numerous times. This led to the conclusion that the mean sum of square when using equation (1) was 0.19 (and a standard deviation of 0.24) and when using equation (2) was 0.54 (with a standard deviation of 0.23). The conclusion, for these data, is that equation (1) leads to a superior model of this decision maker, which means that $f_s(P_s, y) = \min_i [P_i + \beta_i]$ is the better model to capture preferences.

### Main Results from Utility Function Testing

The results of the application of all three methods for the cell phone example used were consistent, showing that the expert group had a consistent understanding of their preferences. Our application had too few data points to allow for any statistical analysis, but the Best-Worst method resulted in perfect accuracy with respect to the 15 sample pairwise comparisons. It also had the fastest calibration time and could be completed in less than 1 hour, even including the time for the subject to fully understand the testing scenario.

The following observations from discussions with the experts were recorded after the tests were performed. The Best-Worst method helped to create a clear understanding of the preferences and how the attribute metrics can reflect them. The identification of the best and the worst solution, followed by a pairwise comparison with the other attributes and alternatives, helped to shape the discussion. The experts trusted this candidate approach the most, as it delivered clearly documented and fully traceable results, without any hidden assumptions.

The Multi-Swing method was applied at the same sessions as the Best-Worst method, which created a small advantage for the second method, as the mindset of the evaluators was already prepared for better understanding the attribute preferences. The amount of data needed from the experts was significantly less, which was overall appreciated, but as the computation of the ranks was based on numerical values reflecting the utility values (captured in Table 7), some skepticism remained. The source of such skepticism was apparently a question of whether the values in Table 7 were a full reflection of the experts’ values in comparison to the researchers’
calibration of the parameters of the model. More transparency, however, would not be in full alignment with current procurement regulations, which do not allow the numerical representation of these values.

"Due to the initial pairwise comparisons being made between adjacent attributes, a comparison of every attribute with the most important attribute and every attribute with the least important attribute is unnecessary."

Finally, the FDNA was conducted based on the data received from the experts for a consistency check. The results were presented to the experts and accepted, and the candidate approach was identified as a potential future solution. The idea to semiautomate the process based on validated training data was identified as promising. However, given the test subject’s confidence with the Best-Worst method and its perfect validation score, it was adapted and tested on a government acquisition scenario. The results to this test are detailed in the following section.

**Testing on a Government Acquisition Scenario**

The second experiment involved an acquisition scenario consisting of 20 attributes to show the scalability of approaches. Similar to the first experiment, two subjects were used with the results from the first subject yielding refinement to the question phrasings; and the results from subject two yielding more general observations about the performance of the approach. The second experiment is a more challenging experiment requiring a lengthier assessment procedure to make comparisons among a very large set of attributes. Another twist to this problem was that these attributes had a threshold value (minimum acceptable value) and an objective value (desired operational goal). The government was only interested in solutions in which the attribute exceeded the threshold value. So, each attribute has two levels (0, 1) to represent whether it met the government objective or not.

A new modification of the Best-Worst method involving the screening of attributes was applied to handle this scenario of having a large set of attributes. The types of attributes of the government study cannot be discussed in this article, but the implementation of the procedure can be
discussed. In order to describe the results without mentioning variable names, another cell phone example is described in place of this actual acquisition scenario.

According to various consumer reports, 20 attributes of cell phones have been identified as decision factors for purchasing a special product (Table 9).

**TABLE 9. ATTRIBUTES/DECISIONS FOR PURCHASING A SPECIAL PRODUCT (EXAMPLE: CELL PHONE)**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Screen resolution</td>
<td>10. Ability to work with multiple windows</td>
<td>15. Operating system used</td>
<td>20. Accessible GPS functionality</td>
</tr>
</tbody>
</table>

The first portion of the assessment procedure involved doing pairwise comparisons across the adjacent pairs of attributes starting with the first attribute. The resulting table of the results to these pairwise comparisons in our illustrative example helps further explain the approach (Table 10).
After this initial pairwise comparison is done, the attributes were binned based on whether they were more important than two attributes, one attribute, or no attributes. This resulted in three bins, which are named Prime, Mid, Low, respectively. To further exemplify the approach, the resulting bins are shown in Table 11.
The next step was to reassess the attributes in each bin to make sure they were allocated properly, as it is possible that an attribute is neighbored by exceptionally strong or weak neighbors—resulting in placing it into an inappropriate bin. To address this challenge, we asked the test subject to first identify the most important and least important attribute in each bin. If the most important attribute of a lower bin was evaluated by the decision maker to be more important as the lowest attribute in the higher bin, a change of bins was conducted. The process repeated until no attributes could be exchanged between bins in this manner. The final binning of attributes for this test is shown in Table 12. It shows that in our example, the importance of the grip factor and the headphone technology was originally estimated too high, while the importance of the battery life was underestimated.

At this point, the Best-Worst method can be executed across a subset of these attributes. The first step in doing this is to identify the most important and least important attribute in each of the three bins. Due to the initial pairwise comparisons being made between adjacent attributes, a comparison of every
attribute with the most important attribute and every attribute with the least important attribute is unnecessary. In the prime bin, each attribute is compared with the most important and least important attribute. Then for the mid bin, only the most important and least important attributes are compared with the least important attribute in the prime bin. Likewise, these attributes are compared with the most important attribute in the low bin. Then finally, the most important and least important attributes of the low bin are compared to the least important attribute in the mid bin. These measures of relative importance are again done on a scale of 1–5. After all of these comparisons are made, enough information is extracted to perform a least squares estimation to approximate the weights for all 20 attributes.

The nice feature about having attributes with a binary value is that assessing a single attribute utility function for each attribute is unnecessary. If the attribute meets the threshold, then the utility is mapped to a value of 1; and if it does not meet the threshold, it is mapped to a value of 0. Therefore, the weights can be used with 0, 1 terms for the attribute values directly to result in this equation for the preference model: \( V = \sum_i w_i t_i \) where \( t_i = 1 \) if the value for attribute \( i \) meets its threshold level for the objective.

In this experiment, the most notable result of this new assessment procedure was that the interaction time with the test subject to train this preference model with 20 attributes was done in less than 1 hour. For this amount of time, the decision maker can expect to stay engaged for the duration of the assessment procedure, and remain accurate and limit inconsistencies. Another promising feature of this method is that the initial pairwise comparisons can be held out from training of the model and used for validation. When reserving 8 pairwise comparisons for validation, the model resulted in making the same decision with the pairwise comparison as the decision maker in 7 out of 8 (87.5%) of the test cases.

**Conclusions**

This article investigates the utility function approaches to be used as a foundation for a game theory framework in support of government acquisition. Through accurate and efficient calibration of utility functions, there is a strong potential to develop a framework that can more effectively illuminate strategies that move industry into the government’s preferred negotiation point and expedite the decision-making process in acquisition. These utility functions can be used for vendors to generate optimal bids when integrated with their own utility functions (which are cost functions parameterized by the same
attributes) in a mathematical programming framework. This concept is only realizable if the utility functions can be accurate representations of the decision maker’s preferences and if the calibration of these functions can be performed in a reasonable amount of time.

All outputs from the government GUI are traceable, well documented, unambiguous, and repeatable, which can support the government as well when protests occur.

The primary objective of the research was to examine the performance of three utility function calibration procedures in a government acquisition scenario. To support this objective, testing was performed by working closely together with a government sponsor, but not with a sufficient number of subjects to deliver statistically significant results. Concretely, three potential utility function calibration procedures from literature were adapted and applied to examine the accuracy and assessment time of each approach. The three methods were first applied in a test involving the acquisition or purchasing of a cell phone. The Best-Worst method showed robustness in handling a small or large number of attributes effectively. The MSRM demonstrated the ability to capture sharp drops in utility in individual attributes. This is an important feature when some attributes present thresholds where the entire product becomes unusable by the government. The FDNA method showed the ability to work effectively when the decision maker is more qualitative than quantitative in nature.

The Best-Worst method was assessed as having the best validation test results so the procedure was also extended to be more applicable to acquisition scenarios involving a large number of attributes (20 plus attributes). This was accomplished by providing an efficient method for screening and grouping attributes at the beginning of the procedure. These situations are more common than the examples that are often provided in the toy examples from the literature, which contain a small number of attributes. This expanded scenario was then tested on a real-world government acquisition example and was shown to be capable of calibrating an accurate utility function in under 1 hour of engagement time with the decision maker.
In conclusion, applying the game theory framework mentioned in this article to government acquisition holds promise when using a government utility function as a foundation, which is calibrated by the Best-Worst method. This leads to future work involving the integration of the utility function into a decision support framework that can enable potential bidders to maximize the fitted government utility function with respect to their own specific cost functions. The specific cost functions are then parameterized by the same attributes as the utility function. The sampling procedure to calibrate these cost functions along with the best optimization algorithm to apply are the next steps of our research. The optimization algorithm must have the ability to generate solutions in near real-time in order for this decision support framework to be usable and effective.

In addition, research is ongoing pursuing the implementation of the game theory framework into a web application consisting of two front-end GUIs: one for the government and one for the bidding vendor. The GUI to be used by the vendors is envisioned to help prepare bids by generating automated solutions through the maximization of the government’s calibrated utility function. In this GUI, the vendor submits information to calibrate a cost function parameterized by the same key attributes as the utility function, and then the optimization is performed. The GUI on the government side enables the user to visualize all incoming bids simultaneously and perform sensitivity analysis on their highest valued bids. Lastly, all outputs from the government GUI are traceable, well documented, unambiguous, and repeatable, which can support the government as well when protests occur.
References


Author Biographies

Dr. Scott Rosen

is the chief engineer for the Operations Research Department of The MITRE Corporation with more than 10 years of application experience. He holds a PhD and MS in Industrial Engineering and Operations Research from Pennsylvania State University and a BS in Industrial and Systems Engineering from Lehigh University.

(E-mail address: srosen@mitre.org)

Ms. Kelly Horinek

is a principal analyst in the Center for Acquisition and Management Sciences Division of The MITRE Corporation. She holds a BBA in Marketing from Texas A&M University, is a Certified Professional Contracts Manager, and a former Contracting Officer.

(E-mail address: khorinek@mitre.org)

Mr. Alexander Odeh

is a principal economist and business analyst in the Center for Acquisition and Management Sciences Division of The MITRE Corporation. He holds a Bachelor in Economics from McGill University and a Master’s in Public Policy from The George Washington University.

(E-mail address: aodeh@mitre.org)
Dr. Les Servi

is a group leader in the Operations Research Department of The MITRE Corporation. He holds a PhD and MS in Engineering from Harvard University, and an ScM and ScB in Applied Mathematics from Brown University. He is a Fellow of INFORMS (Institute for Operations Research and the Management Sciences), a former editor of three journals, and a past member of two Defense Science Board studies.

(E-mail address: lservi@mitre.org)

Dr. Andreas Tolk

is a technology integrator for the Modeling, Simulation, Experimentation, and Analytics Division of The MITRE Corporation. He holds a PhD and MS in Computer Science from the University of the Federal Armed Forces, Munich, Germany. He is a Fellow of the Society for Modeling and Simulation and senior member of the Association for Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE).

(E-mail address: atolk@mitre.org)