;

A METHODOLOGY FOR CREATING EXPERT-BASED QUANTITATIVE MODELS FOR EARLY PHASE DESIGN

A Thesis
Presented to
The Academic Faculty

by

William O. Engler, III

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Aerospace Engineering

Georgia Institute of Technology
May 2013

A METHODOLOGY FOR CREATING EXPERT-BASED QUANTITATIVE MODELS FOR EARLY PHASE DESIGN

Approved by:

Dimitri Mavris, Advisor
School of Aerospace Engineering
Georgia Institute of Technology

Daniel Schrage
School of Aerospace Engineering
Georgia Institute of Technology

Vitali Volovoi
School of Aerospace Engineering
Georgia Institute of Technology

Chris Raczynski
Systems Engineering Advancement Program
General Electric Energy

Kelly Griendling
School of Aerospace Engineering
Georgia Institute of Technology

Date Approved: 29 March 2013

Ad familia et Deum.

ACKNOWLEDGEMENTS

I would like to thank Dr. Dimitri Mavris for serving as my advisor and for supporting and educating me in innumerable ways all these many many years that I have been a member of the ASDL Family. I would also like to thank my committee, Dr. Daniel Schrage, Dr. Vitali Volovoi, Dr. Chris Raczynski, and Dr. Kelly Griendling, for giving their time and knowledge to make this research better from before the proposal until completion.

I would like to thank all the individuals who volunteered to give the data needed to test the research presented here. Thank you also to Taewoo Nam and Chris Perullo who helped me find the right data and models I needed to test the volunteers. Without the information these individuals provided, this research would have remained an untested theory.

I would like to thank my doctoral compatriots Elise Beisecker, Dan Crowley, Matt Daskilewicz, Daniel Cooksey, Pat Biltgen, John Salmon, Paul Brett, Scott Wilson, Kevin Johnson, Robert Combier, Ian Stults, Jonathan Murphy, Joe Iacobucci, and David and Michelle Ediger who helped to forge the path ahead of me and alongside of me. Their advice, insight, commiseration, and the occasional kick in the pants made this process. I also thank my friends who were wise enough to leave with their master's degree. In particular, Bryan Wood, Mark Smith, and Russell Joffrion.

Thank you to the chaplains and students of the Georgia Tech Catholic Center and especially the graduate student group, all of whom helped me to maintain my focus. Deo gratias.

Finally, I thank my family for encouraging and believing in me even when I had my doubts. My parents, brothers and their families, my fiancée and her family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

α	probability of type-I error
α_k	Krippendorff's alpha
$\Lambda_{c/4}$	wing quarter-chord sweep
ρ	density
ρ_{xy}	Pearson product-moment correlation coefficient
C_{D0}	zero lift drag coefficient
d_i	difference in ranks when calculating Spearman's rank correlation
L/D	lift to drag ratio
m	intermediate metric
M_{rel}	relationship matrix of a QFD
n_O	number of observations
Q	overall quality of a solution
R	system or customer requirement
R^2	coefficient of determination
r_s	Spearman's rank correlation coefficient
S	wing planform area
t	statistical parameter for Student's t-test
t/c_{avg}	wing average thickness to chord ratio
T/W	thrust to weight ratio
V	velocity
v	importance weighting of an engineering characteristic
V_{app}	approach velocity
w	importance weighting of a system requirement
x	design variable
Acq$	Acquisition Cost

AF-ICE	Air Force Integrated Collaborative Environment
AHP	Analytic Hierarchy Process
ALCCA	Aircraft Life Cycle Cost Analysis
ALTER	Approximation of Logical Trends from Expert-sourced Relationships
AR	wing aspect ratio
BFW	block fuel weight
BOGSAT	bunch of guys/gals sitting around a table
CAIV	cost as an independent variable
CFD	computational fluid dynamics
FEA	finite element analysis
FLOPS	Flight Optimization System
FPR	fan pressure ratio
IPPD	Integrated Product and Process Development
IPT	Integrated Product Team
IRB	Institutional Review Board
IRMA	Interactive Reconfigurable Matrix of Alternatives
JLTV	Joint Light Tactical Vehicle
M&S	modeling and simulation
MADM	multiple attribute decision making
MDO	multidisciplinary optimization
OEW	operating empty weight
PAT	Portfolio Analysis Tool
QFD	Quality Function Deployment
RDT&E	research, development, testing and evaluation
RFP	request for proposals
RMSE	root mean squared error
ROSETTA	Relational Oriented Systems Engineering and Technology Tradeoff Analysis

SOAR	Strategic Optimization for the Allocation of Resources
SP2	Strategic Portfolio Planning
TOFL	takeoff field length
TRL	Technology Readiness Level
TSFC	thrust-specific fuel consumption
W_{eng}	engine weight
W_{wing}	wing weight

SUMMARY

There are many occasions when engineers need to perform a rapid analysis and are willing to accept a larger amount of uncertainty or error in the results in exchange for speed and availability. Such occasions are most common in early design or when deciding whether or not to participate in a design competition at all. If the appropriate models needed to support this rapid analysis are not available, they must be created in a similarly rapid fashion using the resources at hand. This research presents a structured, repeatable process for developing quantitative, hierarchical models using experts as the sources of information. This method is the Approximation of Logical Trends from Expert-sourced Relationships or ALTER.

A survey of existing literature resulted in a five-step modeling framework. The mathematical structure of the model uses a series of hierarchical linear or quadratic transformations to relate design variables of a system to intermediate metrics to system-level performance metrics. Such structures are commonly used in methods such as Quality Function Deployment and the Strategic Optimization for the Allocation of Resources. Past research to develop and improve these two methods has been leveraged for the benefit of ALTER as well to identify alternatives at each of the steps and select the most appropriate path. Several areas required experimental investigation: the number of individual needed to participate, the level of expertise of those individuals, and how detailed the information captured from those individuals needed to be. These areas, in addition to the accuracy of the method as a whole, were tested using a relevant and well-understood aerospace engineering problem.

This experiment called for forty-two volunteers with varying levels of experience to provide information about the performance of a notional civil airliner using one of three different interfaces and scoring scales. The aircraft and its mission were defined using morphological analysis as a 225-passenger class aircraft flying a standard mission profile and divert with a range of 6000 nautical miles. An aircraft design and mission performance truth model was used as a point of reference to compare the accuracy of the experts. As part of this effort, an approach was developed to decompose the existing monolithic black-box truth model into the hierarchical format while maintaining its level of accuracy. The expert-based information was compiled and compared against a trusted truth model to test how well experts agreed with each other and as a group, how closely they matched the corresponding parameters of the decomposed truth model, and how accurate the outputs of the expert-based models were to the original unmodified truth model. Each of these tests were used to identify the effects of the different interfaces, different levels of expertise, and different number of experts included in the model.

The results of the experiment demonstrated that it is possible to create accurate performance models based solely on information from 5-10 experts with a moderate amount of experience (more than four years) in the specific field. The results further demonstrated a slight decrease in the accuracy of individuals with two to four years of experience relative to those with more than four years or less than two years of experience as a result of over-confidence in their own knowledge. Participants using a full integer scale from -9 to +9 gave more accurate information than using the more traditional reduced integer scale ($\pm 0, 1, 3, 9$) or a scale that included a graphical depiction of the model parameters. In this case, engineers were able to give more specific assessments with the full integer scale than the reduced scale. The graphical depiction tended to cause participants to estimate the information they provided based on the plot rather than the numeric value, reducing the accuracy.

Further conclusions were reached based on feedback from participants and post-processing of the information. Using a small number of the relationships to serve as tests of experts' ability demonstrated a discrete step change in accuracy of their estimates for all relationships between experts who passed three or four tests against those who passed zero or one test. In areas where participants initially had a great deal of disagreement, it is demonstrated that giving additional problem-specific information about the contributions of different effects significantly improved both the agreement between experts and overall accuracy of the information given in those areas. Overall, the method was shown to be capable of producing highly accurate performance models for a complex aerospace engineering conceptual design problem.

CHAPTER I

INTRODUCTION

Solving an engineering problem requires balancing the technical requirements of the solution with the time and money required to develop that solution. Depending on the problem, one of these three may be the most constraining forcing the requirements on one or both of the others to be relaxed. A difficult problem that needs to be solved immediately will cost more but an immediate need with a small budget will not be able to provide the same level of performance. This balancing act also applies to the analyses used to support the design decisions. When the development time is short, the time spent with modeling and analysis must also be short and designers may have to accept a less rigorous or detailed analysis.

In the early phases of aerospace design specifically, the realm of possible solutions for a given problem can be quite large and diverse. The analyses required to narrow this space intelligently must be able to quickly eliminate designs or regions that are not of interest to allow efforts to be focused on those that are more promising. They need to be flexible enough to handle innovative solutions without requiring detailed information that is not available early on. In exchange for high speed and flexibility, the results of the analyses may be less precise or accurate. When an existing model is not available, not applicable or does not meet the constraints, a new one must be created that does and that model creation process must be similarly fast and flexible.

This research proposes a structured method to rapidly create simple performance models for certain classes of engineering systems using expert opinion and experience. This method utilizes a hierarchical breakdown to calculate top-level system requirements from a set of design variables, using a set of intermediate metrics. A group of experts provide the model coefficients of the linear and quadratic models to describe the relationships between the levels. For convenience, the proposed method has been named *ALTER*: Approximation of Logical Trends from Expert-sourced Relationships.

This dissertation discusses the development of ALTER within a generic modeling process. During this development, alternatives are identified for specific steps in the process and compared against the needs. An experiment to test the method and the usefulness of the resulting model is described. The results of that experiment are analyzed and used to discuss the strengths, weaknesses, and areas of improvement.

1.1 Motivating Examples

Recent major aerospace projects have earned a reputation for taking very long periods of time to develop. Much of this time is spent in detailed design and testing. The initial design trade offs, on the other hand, take a relatively short period of time. The Joint Strike Fighter program took only a year between the initial request for proposals and the awards to build prototypes in 1996, despite the development issues it has had since then. It is this period of time when models need to be available quickly. In other cases, the pace may be even faster.

On January 26th, 2012, the United States Army and United States Marine Corps released the official request for proposals (RFP) for designs for the Joint Light Tactical Vehicle (JLTV) to replace the Humvee. This document set March 13th, 2012 as the deadline for submitting proposals, giving just thirty-three business days[70]. While there were draft versions of the RFP, the final version dropped the per-unit ceiling cost from the original value of $450,000 to $250,000 and relaxed the weight requirements for the vehicle[23]. These are significant changes in the design requirements over a very short period of time that likely required at least a partial redesign of a team's submission. However, a total program cost somewhere between $10 and $70 billion and a total production run of more than 32,000 vehicles (not including foreign sales) is sufficient motivation for companies to make whatever effort is necessary to submit a design[50].

Similarly, the United States Air Force produced a draft RFP with significant changes during the second round of competitions for the $35 billion KC-X aerial refueling tanker replacement project on August 6th, 2008. At that time, they planned to release the final version on August 15th with a deadline for proposals of October 1st: thirty-four business days later. The new RFP gave additional credit for exceeding the minimum requirement for fuel available for refueling, providing a bonus for oversized aircraft[7]. The revised RFP also changed the period for calculating the life-cycle cost from twenty-five to forty years and reverted from computerized cost modeling to "classic cost evaluation analyses"[27]. These changes might have led the teams to redesign or at least reevaluate portions of their design submissions.

These examples present two situations where a customer made major changes to design requirements and then gave potential suppliers a short period of time to revise their proposals. A quality proposal depends on developing a design to the point that any performance and cost metrics are believable by the customer and known to be feasible. If the design requirements shifted enough that previous design efforts are no longer applicable, a new design, and new analysis becomes necessary. Any analysis must be completed with adequate time remaining to understand and document the results into a proposal, further reducing the time available for design. The analysis needs to be as detailed as is possible subject to the time constraints.

In other problems, performance models are necessary as part of a larger modeling effort. A 2004 project, the Air Force Integrated Collaborative Environment (AF-ICE), was focused around the idea of bringing together models from multiple sources to be able to compare technologies in multiple disciplines and platforms. Within this agent-based simulation environment, multiple different vehicles were represented, each one with separate design parameters. Existing aircraft and missile models required information that was not available. For some platforms, a new model was necessary to capture the effects of technologies, driving the need to develop a traditional model. For other systems included in the simulation, any model that captured the broad trends would have been sufficient. Traditional models were created for those systems as well, but at the expense of time and manpower resources that could have been better spent elsewhere. In these cases, the ability to rapidly create models, even if they were less accurate, would have been of significant help to the overall process considering the short period of time available to perform the entire study.

Other product development activities have required a simplified cost model and performance models without being given information about the details of the system. In this case, several systems were already in existence and designed, but could not be named in order to eliminate the possibility of bias. Analytical cost models of systems are notoriously difficult to get from companies. On the other hand, many individuals know enough about the impact of cost to be able to create a model with enough detail to compare abstract concepts. Having a systematic approach to creating such models would have been incredibly helpful.

Each engineering problem is unique in its own ways, but many engineers re-use similar models and modeling techniques. When approaching a new problem, the needs, constraints, and objectives of that particular problem must be taken into consideration. This will both define what types of solutions are reasonable to consider as well as identify what analysis is necessary to convince stakeholders (whomever that might be) that a particular solution is the correct one.

1.2 Modeling and Analysis in Design

Engineering has included modeling and analysis since its origins when it was limited to building structures. Earliest models would have included scale replicas and rules of thumb necessary for finding the correct proportions to maintain both beauty and strength. These rules of thumb would have been based on past experience. Eventually these rules of thumb would be better understood based on the underlying physics of structures.

Outside of civil engineering and in the present day, the most accurate source of analytical information is historical data. This information comes from measurements and tests of a previously designed and built physical system under the same operating conditions. The information gathered from previous tests is as accurate as the sensors used to collect it. For a new design identical to the source of the historical data, this information is entirely applicable. For certain direct design derivatives, historical data may still be valid, such as information about the roll stability for a stretched version of an airliner. As the new design deviates from the ancestor, the reliability of the information degrades quickly. For most new designs, historical data may be used to calibrate models, but is not a direct part of the analysis.

Physical experimentation represents a “lighter” choice over historical data. Rather than using a full-size production system, prototypes or scaled-down versions of portions of the system are built and tested. Examples of this would be a wind-tunnel model to test some portion of the aerodynamics or a structural lay-up of a beam to gather data about the strength and weight of the composites expected to be part of the design. These experiments require less time and resources than building the full system, but are still an involved process that is beyond the scope of most early-stage design efforts.

More typically, modeling and simulation (M&S) are the preferred tools for gathering analytical information while designing a system. Simulation refers to the instantiation and repeated use of a model to describe or predict the behavior of the system described by the model while modeling refers to the process of creating that model. M&S has the benefit of describing a system in a digital or mathematical framework without the need to physically build a representative object, greatly speeding up analysis.

Models in the general sense may refer to a wide variety of concepts from a scaled representation of a physical object or a diagram describing the flow of matter, energy or information to semantic models of language or equation-based mathematical models [56]. Within engineering, and especially when combined with simulation, the term typically refers to the mathematical-type.

Most of the models used in engineering perform analysis on a set of design parameters in order to predict the performance of a system defined by those parameters. There have been some attempts to reverse this process and produce models that identify design parameters based on input performance metrics. This is difficult because there are many solutions which can produce a particular performance point [47]. To reduce complexity of the models needed, only those models that use design parameters as inputs to produce performance metrics will be investigated here.

Models can be classified based on their fidelity, the degree to which a model accounts for the full physics and complexity of the system being modeled. An aerodynamics model which treats the air flow as having uniform properties for a given cross-section has a lower fidelity than a model which treats the air as a very large number of individual and discrete particles colliding against each other and the surface boundaries. The latter model accounts for more of the physics of the flow. Fidelity provides only a partial ordering. If two models each account for phenomena not included in the other, defining which has higher fidelity is difficult [96].

The use of a relative scale for describing the fidelity of a model makes it very difficult to rigidly define what is a low fidelity model and what is high fidelity model in an objective way. The answer may even change between disciplines or practitioners. Despite the trouble with a precise ordering, many models are classified into broad categories of high or low fidelity.

High fidelity models are typically those that use fundamental physics-based relationships and a great deal of iteration. Examples include structural finite element analysis (FEA) using programs like ANSYS or NASTRAN, computational fluid dynamics (CFD) using programs like FLUENT or FUN3D, or large multibody dynamics and orbital mechanics models. They try to mathematically replicate the exact physical behaviors present in the real world and often include a large amount of internal iteration. They are typically applicable across a wide range of problems and are very detailed and very accurate. This level of detail typically requires similarly detailed input information and a great deal of time to set up. The large amount of iteration means that it also takes a long time to perform the analysis of a single set of inputs.

Low fidelity models tend to make a number of simplifying assumptions to capture the broad main effects while neglecting smaller, minor details. Examples of low fidelity models include simplified fluid models like Bernouli's equation or the one-dimensional jet thrust equation, models that use block time calculations for dynamic behavior such as constant fuel burn during flight or constant speed during driving and "back of an envelope" equations. The simplifying assumptions required with such approaches significantly reduce the complexity of the problem, but may also reduce the applicability. These are generally much less detailed and less accurate than high fidelity modeling, but have the benefit of faster setup and runtime.

While not specifically high or low fidelity, surrogate modeling is an approach for creating regressed models based on data from a parent model over smaller ranges of a subset of the input variables. For designs within those ranges, the surrogate can take the place of the parent model, inheriting the parent model's accuracy and significantly reducing the time required to run the models. The downside of surrogate models is that, because they require a large number of previously-run data points to perform the regressions, they take longer to create than making a small number of runs of the parent model.

1.2.1 Model Development in the Design Process

It is easy to find agreement among engineers that modeling and analysis should occur. Many, but not all, established engineering development and design processes clearly depict this and specify when modeling should occur. These processes usually assume that any models required already exist and generally don't include a step for model development. It is useful to identify when in the process of engineering development a model should be created.

Boehm's Spiral Development process, shown in Figure 1, was one of the earliest specified processes for systems engineering, despite it being originally created for software engineering. Still, only slight changes are necessary to translate it to physical systems. In this process, design starts at the center with the requirements plan. The process rotates clockwise, increasing in detail with each cycle around. Note in particular that each spiral includes a simulation and modeling phase on the right. The prototyping phase just before this step is where, instead of a software prototype, a more detailed model of the system would be created.

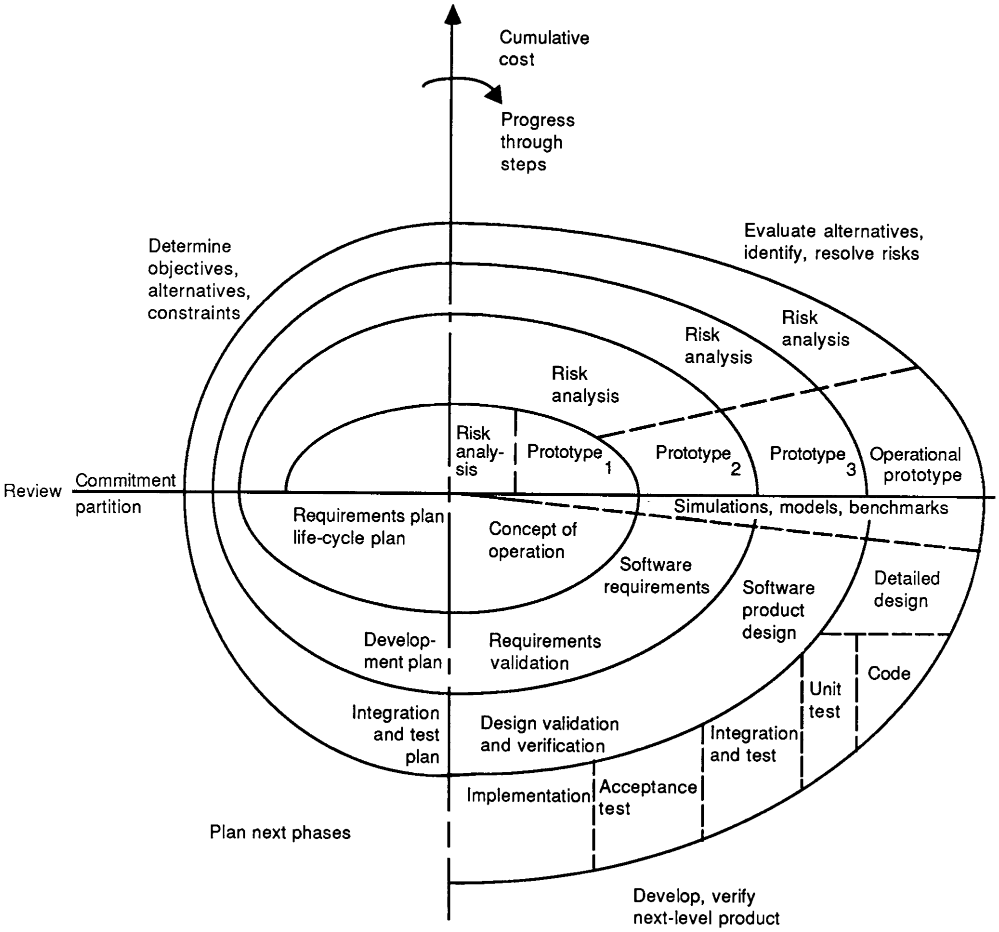


Figure 1: Boehm's Spiral Development Process [17]

Forsberg and Mooz developed the Vee model that is perhaps most common in systems engineering, shown in Figure 2. The fundamental driver is the decomposition of high level requirements into lower-level requirements down the left-hand side of the Vee until the problem can be solved at the base. The right-hand side shows the integration of the system components into the final system. Later versions and variations of this process include horizontal lines from the left side to the right side to indicate areas where integration, verification, and validation planning occur between each block of decomposition and the corresponding integration block. For this to be possible, modeling and analysis must occur at each stage of decomposition and testing at each stage of integration.[1]

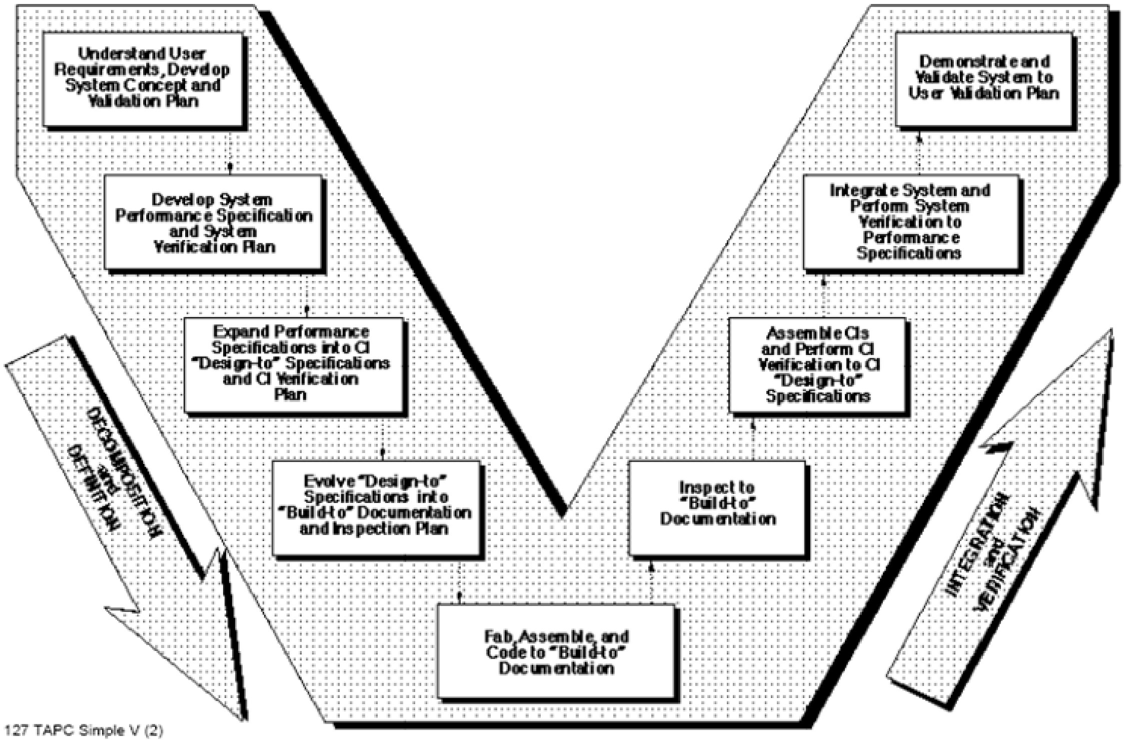

Understand User Requirements, Develop System Concept and Validation Plan
Develop System Performance Specification and System Verification Plan
Expand Performance Specifications into CI "Design-to" Specifications and CI Verification Plan
Evolve "Design-to" Specifications into "Build-to" Documentation and Inspection Plan
Fab, Assemble, and Code to "Build-to" Documentation
Inspect to "Build-to" Documentation
Assemble CIs and Perform CI Verification to CI "Design-to" Specifications
Integrate System and Perform System Verification to Performance Specifications
Demonstrate and Validate System to User Validation Plan
DECOMPOSITION and DEFINITION
INTEGRATION and VERIFICATION
127 TAPC Simple V (2)

Figure 2: Forsberg and Mooz's Vee Diagram [52]

The modeling process actually occurs within each of the decomposition blocks. The Vee process describes that decomposition must occur and the order in which it should occur, but not how that decomposition or integration should be performed. What is especially important to note here is that each block may require a separate model suitable to its level

[1] The testing during integration may be digital simulation rather than or in addition to physical testing.

of detail. The least amount of information is available in the upper left while a fully detailed manufacturing-level schematic (or the equivalent level of detail relevant to the problem) of individual parts should be completed at the bottom. The upper right has the detail of the schematic, but at the scale of the full system. This means that lower-fidelity models are most appropriate and needed at the beginning to be able to offer whatever information is available to the next decomposition steps.

The Georgia Tech Integrated Product and Process Development (IPPD) process is another approach design and is shown in Figure 3. This design process has been tailored more toward entirely digital design and further specifies how the design process interrelates with systems engineering and quality engineering methods. The primary process is in the center, going from *Establish the Need* through generating and evaluating alternatives until there is enough information to *Make a Decision*. The process is generic enough that it does not specifically call out analytical modeling, but the *Evaluate Alternatives* step is typically the place where this occurs by employing a multidisciplinary optimization (MDO) approach as shown on the right.

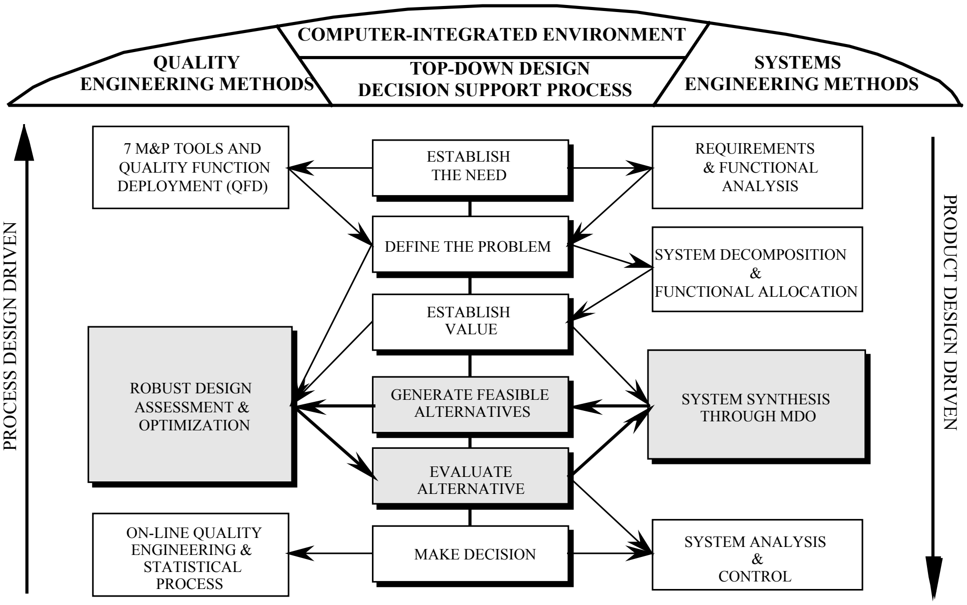

COMPUTER-INTEGRATED ENVIRONMENT
QUALITY ENGINEERING METHODS
TOP-DOWN DESIGN DECISION SUPPORT PROCESS
SYSTEMS ENGINEERING METHODS
PROCESS DESIGN DRIVEN
PRODUCT DESIGN DRIVEN
7 M&P TOOLS AND QUALITY FUNCTION DEPLOYMENT (QFD)
ESTABLISH THE NEED
REQUIREMENTS & FUNCTIONAL ANALYSIS
DEFINE THE PROBLEM
SYSTEM DECOMPOSITION & FUNCTIONAL ALLOCATION
ESTABLISH VALUE
ROBUST DESIGN ASSESSMENT & OPTIMIZATION
GENERATE FEASIBLE ALTERNATIVES
SYSTEM SYNTHESIS THROUGH MDO
EVALUATE ALTERNATIVE
ON-LINE QUALITY ENGINEERING & STATISTICAL PROCESS
MAKE DECISION
SYSTEM ANALYSIS & CONTROL

Figure 3: Georgia Tech IPPD Process [114]

For all three of these processes, the model must be created before analysis is performed, but after what must be modeled is determined. Boehm's Spiral makes this the easiest since each spiral around is just a further specification of the previous spiral. The modeling may be able to follow the same process as higher fidelity analysis is incorporated to the same model at each step. The Vee model makes this the hardest since what must be modeled at a lower step is not known until the previous decomposition step has been completed. The earliest that a model can be created for the Georgia Tech IPPD process depends entirely on the particular problem. If the *Generate Feasible Alternatives* creates discrete design families, the models cannot be created until a family is identified. If it is instead generating a set of possible design points within a continuous design space, the model can be created as soon as the design space is identified earlier in the process or even before the process starts.

1.2.2 Trade Between Fidelity and Speed

A simple bridge across a small creek to allow playing children to cross safely may be created with no more than a single plank of wood in a few moments. On the other hand, the Akashi Kaikyo Bridge carries thousands of vehicles across 3.9 km of swift, deep water in an earthquake-prone area with high winds every day and has done so for safely since 1998. The challenging design problem meant that the constructing of the Akashi Kaikyo Bridge cost more than 3.6 billion US Dollars and require 10 years to build, not including design time.[59] In an abstract sense, both bridges satisfy the same purpose: provide crossing over a body of water. However, the specific requirements associated with each are obviously and significantly different. The plank bridge only has to be durable enough to satisfy the temporary needs, but cannot be expensive or complex. If it washes away or breaks after a year of use, the consequences are small. The Akashi Kaikyo Bridge, on the other hand, must be safe and durable for a long period of time before any other requirements. While cost is important to the design, the project cost whatever was necessary to satisfy the other requirements.[2]

Similar examples of trades exist within aerospace, though the designs that seem to ignore cost completely for the sake of performance tend to stand out, such as the SR-71 Blackbird. It was designed to meet a unique technically-demanding problem and solved it with approaches that were both expensive and unconventional (such as secretly sourcing a titanium alloy for the skin from the very country it was to be used against) [22]. The low-cost and low-time extreme can be seen in a picture that has been circulated by email of an airframe interior being wired up. In this image, a shop broom, still with it's sweeping head, attached to the ribs of the fuselage to serve as a mount point for an electrical block. In this case, technical requirements for weight or strength were significantly lower than the problem of availability. These examples demonstrate the impact that differences in the requirements have on what solutions are feasible. The goal of engineering is to identify the solution that is good enough to solve the problem. A similar trade happens when selecting what type of modeling approach to use.

Figure 4 shows a notional diagram of the tradeoff between the time required to build and/or run a model (the trend tends to be true for both) and the uncertainty of the results of a particular modeling method. The ideal location is in the lower left corner with instant results that have perfect accuracy and precision. While the space is not entirely continuous, it is helpful to think of a Pareto frontier along the lower left edge. Different modeling approaches can move along this edge or move further to the interior with a model that is both slow and wrong, but a breakthrough in knowledge would be necessary to shift the frontier inwards.[3]

Two sources of analytical information discussed previously are not included here: historical data and surrogate models. The pre-existence of these two hides the sunk cost

[2] Based on current tolls to use the bridge and the cost of interest on the loans used to build it, the cost of the bridge will not be repaid before the end of its useful lifespan. This is further evidence that the cost of the project was not a primary constraint.

[3] A closed form solution of Naviers-Stokes equations would be an example of such a breakthrough.

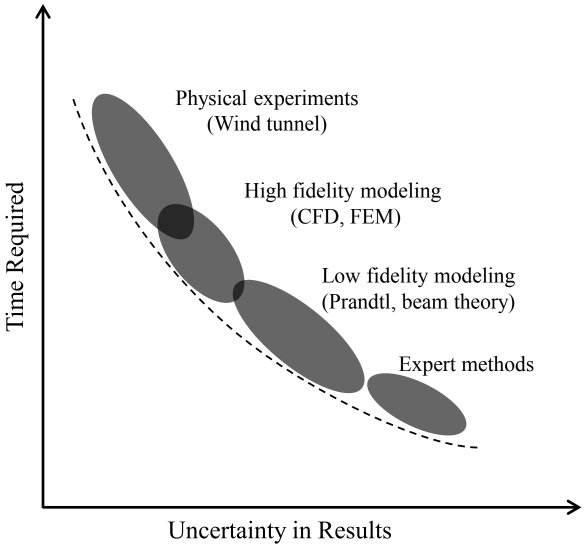

Time Required
Uncertainty in Results
Physical experiments (Wind tunnel)
High fidelity modeling (CFD, FEM)
Low fidelity modeling (Prandtl, beam theory)
Expert methods

Figure 4: Comparison of Modeling Methods by Speed and Certainty

associated with creating a prototype or building and running the model used to create the surrogates. However, if they are available already and are applicable, both may be nearer to the origin than the Pareto frontier would allow.

The claim of “Faster, Better, Cheaper” as made famous by NASA can apply to many parts of design due to the reduced bureaucracy associated with smaller budgets, in turn allowing engineers to focus more on development of good solutions[128]. This generally does not apply to modeling where the limits are brain power and man-hours creating a model, and computing power when performing the analysis. In fact, usually the three are in opposition. A higher fidelity (better) model requires more time and higher-powered computers, hence increase costs.

For the very early stages of design and other cases of minimal time available for analysis, we need modeling approaches that are quick to setup and to run. The design is still rough and we need “the detail, depth of fidelity, and precision of the models [to] be sufficient only to clearly distinguish between the options” [65]. Acknowledging the uncertainty associated with a lower fidelity model, decisions are still possible if the options are separated sufficiently [123]. Even if separation is not guaranteed, there is still a tradeoff between the consequences of being wrong and the consequences of taking too long to guarantee being right[121]. A sketch on the back of a napkin may be satisfactory to build a dog-house with low risk and low consequences of failure, but a scaled model and digital rendering would be expected for a multi-million dollar skyscraper.

Since less detail is required in the early phases, the time required to use high fidelity modeling is not easy to justify, which eliminates them for consideration. Surrogates of high fidelity modeling could be appropriate, but only if they already exist for the given problem. Low fidelity modeling, or a surrogate of a low fidelity model, is the best option if a suitable and applicable model is available.

In the event an appropriate low fidelity model is not available, the options are to use a higher fidelity model with the hope of finishing in time, modify an existing low fidelity model to fit the problem, or to develop a new suitable low fidelity model. Modifying an existing model depends on a complete understanding of the original assumptions and understanding the implementation. On the other hand, traditional model development from first principles can take months or years. Other solutions are necessary.

The companies considering bidding on the two examples each have experience in their respective industries. Through hiring and internal development, their employees have tacit knowledge of the design problem and the trades that will need to be made. The size of the budgets for each program motivate companies to utilize those employees that can be of help. This human capital is essential to creating a proposal and is available for model development. This research proposes a way to create models from the knowledge and experience of the experts already present and available within a given group.

1.3 Experts in Modeling

The term "expert" describes a person with exceptional knowledge of a particular subject.[4] This knowledge is a result of both experience and education. Experts are essential to creating and using models. Creating a model requires knowledge of the physical principles that define the system or phenomena of interest. Models that rely on simplified physics or aggregated effects require someone with experience to identify which aspects can be simplified. Experts are also necessary to use models effectively. They have knowledge of the uses and inputs for which a model is valid and what are reasonable results. If the model behaves unexpectedly, it requires an insight into the underlying mechanics of the model to identify the cause and correct the error or modify the model to account for the difference. If the unexpected result is not due to an error, it will take someone who understands the workings and implicit assumptions of the model to explain the results.

For some problems, it may seem like the expert is present to help with the model when the opposite is the truth. Most models were created by experts to aid in their work or to allow other experts to benefit from their experience. It was the experts of past and present who developed and accumulated the knowledge and understanding of nature to enable models to be created. A model is only a summary of a portion of the knowledge available through experts.

There is an anecdote about a student posed with the question of using a barometer to measure the height of a building. The student answers with several different, technically correct, approaches to employ the barometer while intentionally avoiding the method the professor was looking for. The final method he suggests is to find the superintendent (or janitor in some versions) of the building and offer him a shiny new barometer in exchange for telling the student the building's height [28].

One of the many lessons that can be taken from the anecdote is that a knowledgeable

[4] A discussion of who qualifies as an expert is presented in Section 2.1.

person can be just as accurate as other approaches to answering the same question with considerably less effort. In some cases, the person may be more accurate than the other approaches. In order to be trusted to provide accurate information, the person must be an expert. A random individual waiting in the lobby of the building would be less likely to know the height of the building and would not have the technical authority or reputation necessary to provide a convincing answer.

For engineering design, a single numerical answer is rarely sufficient. Multiple design points across a continuous design space should be considered. And while multiple experts may give correct values, they are less likely to agree upon a single best solution. Structured methods that use experts as the primary source of information are known as expert-based methods. These methods provide ways to capture the knowledge of experts so that it can be used repeatedly without the expert being present. Experts combine both historical data and theoretical knowledge in the forms of personal experience and education allowing them to accurately extrapolate outside of the problems with which they have been personally involved. This means that expert-based methods are applicable to a wide range of problems, so long as the design stays within the experts' realm of confidence. And most importantly for the problem at hand, the development process for expert-based models is fast: limited only by the speed of gathering information from individuals.

Consider another view of the anecdote with the stipulation that the student was given the same problem but without being given the barometer or any other tools. In this case, the student would either have to procure a barometer, long measuring tape, protractor, timer, or other tool to find the answer or could go to the superintendent as a first effort. If there was more than one student working together, one could find the superintendent while another sought out the tools to confirm the information. Cases where multiple individuals are available to distribute the analysis efforts help motivate this research. Expert-based models are not proposed as an alternative for all other models, but rather as a stopgap to fulfill a need until the other models are available or to narrow the range over which other models need to be used.

1.3.1 Limitations of Expert-Based Methods

No approach is without shortcomings and limitations. Fortunately, many of the shortcomings are easy to mitigate. Expert-based methods are no exception. Perhaps most obvious is that any method that uses experts can only produce results that are as reliable as the experts themselves. It is important to ensure that the individuals giving information are knowledgeable in the correct area. An expert on ship-building would not be qualified to give information on aircraft engine design. It also means that the more revolutionary the concept, the more likely it is outside any expert's confidence level. So long as experts are honestly confident about the information they are providing and multiple experts agree on a consensus, the resulting model can be trusted.

Ad hominem attacks are those that attack the individuals rather than their position or information. If any of the experts involved with a particular exercise have a poor reputation or have controversial opinions, the results may be discounted. The moderator or individual responsible for the exercise should be sure to minimize those people and include responses from enough other experts to make it clear that the result is from multiple trusted sources. This also helps balance out the effects of experts with different opinions that results from having experts of differing backgrounds. Ensuring that the experts who would have been responsible for verifying a traditional model are also present to serve the same purpose with the expert-based model is helpful.

Several expert-based methods prescribe that all those involved should meet together at a single location. This offers the benefit of being able to share viewpoints and exchange information beyond what is required for the model. If only a small number of individuals participate and they are all already collocated and able to quickly schedule their time to participate, this is an ideal situation. However, as the number grows or if participants are not at the same location, it is increasingly difficult and expensive to arrange a meeting that everyone can attend. Other methods avoid this problem by gathering data from individuals and combining it off-line or by only using meetings as a last resort. Similarly, there is a trade to be made between asking twenty experts to each give eight hours of their time versus asking one expert to spend 120 hours building a model. The twenty experts could produce a model in a single business day compared to the three weeks for the single expert, but at a cost of 33% more man-hours.

While computational models can be run for as long as the resources are provided, humans have limits on attention span and their productivity. Repeated questioning or data requests, especially when in a monotonous format, can tire an expert and result in degradation of the quality of their answers and thought processes. Expert-based methods must focus on the most essential information and make certain that the methods for gathering that information are user-friendly and intuitive.

Lastly, expert-based methods should only be used during design phases or on problems that are tolerant of some level of uncertainty. Such problems include pre-conceptual design, narrowing the design space for future modeling, creating preliminary results to allow other groups to proceed, and deciding whether to pursue a proposal effort. For final design or manufacturing decisions, the time should be spent using a traditional model. No matter how many experts are polled or how certain an expert may be, it is difficult to give the same confidence to a group of experts as to a fully validated and accredited physics-based model.

1.4 Scoping the Problem

At this point it is worthwhile to explicitly state the scope of the models desired as a result of this research.

The goal of these models is to support early phases of engineering design in cases where existing models are not available, appropriate or applicable. This means that they need to be faster and easier to create than the ideal low-fidelity models. It also means that they need to capture quantitative technical and physical aspects of an engineering system.

These models should be predictive and parametric in nature. They should take design variables as input parameters and output predictive estimates of the performance of the system which they model.

They should be mathematically static, deterministic, and continuous. In calculating system performance, the results should not be time-dependent and time need not be an input to the model. A given set of inputs should produce a single set of outputs and should do so consistently.

The method to create them should be consistent regardless of the particular system of interest. This reduces the training necessary prior to use the proposed method. It also limits the degree of customization to a particular problem.

1.5 Organization of This Document

This chapter is intended to present an introduction and motivation for investigating expert-based modeling for the use in early stages of engineering design.

Chapter 2 discusses the background of expert-based approaches in general and expert-based modeling in particular. Several existing approaches that may fulfill some of the needs of early engineering modeling are discussed and evaluated. The results of this evaluation provide the motivation for a focused set of research goals.

Chapter 3 uses those research goals and several preliminary experiments to identify the specific needs and to develop a method for gathering and translating expert-based information into a predictive model useful for engineering design. It starts with a general process for model development and then identifies the specific needs of each step to produce the desired model.

Chapter 4 identifies a set of research questions remaining after defining the modeling approach. Some of these questions focus on testing the performance of method while others call for a comparison of several remaining options. For those research questions suited to them, hypotheses were formed based on information from Chapter 3.

Chapter 5 describes the experiments that were performed to gather data from experts on a relevant aerospace problem to test the proposed method according to the research questions and hypotheses.

Chapter 6 analyzes and discusses the data collected from experts as part of the experiment. This analysis results in a number of conclusions, some which support the proposed method and some which show areas for improvement.

Chapter 7 summarizes the contributions of this research and limitations on where it can be applied. It also discusses areas for future development to expand the application of the proposed method.

CHAPTER II

BACKGROUND

The use of intelligent and experienced individuals as part of problem solving is not a new idea. A great deal of research has been done in the past regarding expert-based methods as a whole as well as their use within systems engineering specifically. This chapter examines the works of giants to identify the proper set of shoulders to stand on. It starts by considering who qualifies as an expert and how such individuals can be identified. It then surveys expert-based methods in general and narrows down to just those useful for engineering modeling until selecting a QFD framework for further development. Existing advances and enhancements to QFD-based research is examined and classified. The chapter ends by identifying what further research is needed beyond what is already found in open literature.

2.1 Identifying an Expert

Experts are the cornerstone of any expert-based method. Some methods, especially those focused around integrated product teams (IPTs), are less concerned with identifying experts than they are with ensuring a diverse group of stakeholders. For others, including those of interest here, each individual contributing information must be an expert in the appropriate field and should be recognized as such. Clearly not everyone is an expert and it is useful to know ahead of time what constitutes an expert and how to identify one. This identification is necessary to determine who should be asked to participate in an expert-based activity. Whether or not their data is included in the model can be determined later based on additional information they provide about themselves and about the problem, as discussed in Section 3.5. It is much easier to remove information from an expert than it is to add more experts after the fact.

The most basic qualifications of an expert are those in a general dictionary definition.

A survey of several dictionaries yields a combined definition of an expert as one with exceptional skill or knowledge displaying a mastery of a subject. As expected, this is insufficient to readily identify an expert since "mastery" and "exceptional" are just as ill-defined as "expert".

Considering their potential for influence, one would expect the court system to have a clear definition to help identify who may serve as an expert. Rule 702 of the Federal Rules of Evidence identifies that "a witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify..." [122], suggesting that any of the five qualifications may be sufficient. Some types of cases, such as medical malpractice, have more specific precedent that mandates certain types of experts testify. For the general judiciary, most of the guidance is about what type of testimony a judge should allow, focusing on the peer-review, acceptance and empirical testing of the methods used to reach the conclusions in their testimony. The requirements for allowable testimony are better defined than for the expert giving it[64]. The peer-reviewed requirement is essential outside of the legal field; any expert-based method, including the one proposed here, should be subjected to peer review and demonstrated experimentally before being used on a major acquisition problem.

For the purposes of this research, Rule 702 is not useful in pre-determining who is qualified to give information. While it identifies several metrics for qualification, it does not identify any minimum for these. Knowledge and skill are abstract notions that are difficult to measure objectively. Experience, training and education are easier to measure and are frequently used to judge expertise. Someone with a doctoral degree in a field is generally accepted as more of an expert than someone with a bachelor's degree in that same field. An individual with forty years of work experience in a field would be accepted as more of an expert than someone in their first year of employment in that field.

These measures are not sufficient on their own to guarantee someone is an expert. An individual with a research doctorate only needs to be an expert in his or her narrow area of research. They're also not always necessary to be able to collect correct information. There are individuals who have an intuitive knowledge of some types of systems without having advanced degrees or years of experience. Still, an expert-based method is defensible only if its experts are also defensible. So while the five measures given by the Supreme Court may not guarantee expertise, participants should qualify in at least some of them before participating, in particular in experience, training and education. A manager with no engineering experience and only a management degree would not be a valid choice as an engineering expert.

In some fields, there are recognized methods for demonstrating a certain level of expertise and excellence. In medicine, a doctor can become Board Certified. In certain engineering disciplines, the Professional Engineer license is necessary to be considered an expert. These certifications or licenses typically require proper academic credentials, exhaustive written and oral examination to test an individual's basic knowledge, and a period of performance underneath someone who has already been certified and trusted in the field. Depending on the field, they may also include practical tests using model patients for medical fields, test rigs for trades, or sample projects for engineers. These four requirements line up very well with those specified in Rule 702 as tests of education and training, knowledge, experience, and skill respectively. One could defensibly select experts from a pool of individuals certified or licensed in that particular field and specialty.

Unfortunately, not all fields (including aerospace) use such certifications, so it is difficult to rely on them universally. While it is possible to create processes for those fields to identify individuals with a suitable expertise, this process would be expensive and time consuming, making them impractical for any occasion when an expert-based method was the most appropriate approach. Beyond that, the best tests to determine whether or not someone could give correct data for a particular problem would be the same process as giving data for the problem. A different engineering problem or different testing format would not guarantee transferability. Grading such a test would require the information that would render the expert-based method unnecessary.

Time spent in a field is a frequent measure of expertise on its own. To test this, the United State Army conducted an experiment to compare the level of knowledge and intuition in combat scenarios of different groups of individuals using the Tacit Knowledge for Military Leadership Scale. On one extreme, first year cadets at the U.S. Military Academy (West Point) were compared against Lieutenant Colonels (with an average of 18 years experience). When the bottom 25% of each group was compared, there was a significant difference in the performance. But when the top 25% of each group was compared, the results were indistinguishable. This suggests that years of experience has an effect on the average expertise within a population, but that this is secondary to individual excellence. Individual excellence needs to be identified in another fashion. [83]

Pace and Sheehan offer advice on selecting experts for use in model verification and validation, some of which can be applied to model creation. One of the key ones is *recognized competence*. Competence itself is necessary to provide useful information, but the recognition of that competence also lends credibility to the model itself. This is one measure of defense against the *ad hominem* attacks mentioned earlier. This measure is effective for a more important reason. Individuals with knowledge in a field are best able to assess others in that field and can often do so without using formalized testing processes. Those who have worked together in the past have a feel for how frequently an individual's intuition match up with later modeling as well as how confident and comfortable they are in a subject area. [104]

Pace and Sheehan suggest capturing this effect by asking individuals to nominate others and include records of these nominations, along with their qualifications, as part of further analysis. Even *ad hoc* methods for selecting experts frequently rely on managers identifying and suggesting a particular individual within a department or group to serve as the representative of that group. While on occasion this falls to whomever is available, for the most part managers have a vested interest in selecting someone who will portray their group in a positive light and ensure that the information they provide is trustworthy enough to minimize future issues resulting from errors. Even if nominations are not used in selecting experts, asking participants to rate others is frequently used, either explicitly or implicitly for a number of methods, including the Delphi Method. [39]

Unfortunately, despite being the best-accepted method for identifying and rating experts, the reputation and opinions about experts cannot be easily included here. Such things inherently violate individuals privacy and are limited within academic research using human subjects. Additional implications are discussed in Section 5.2.1.

Most considerations of expertise focus on the minimum level of expertise required. Armstrong makes an argument that the level of expertise required to give useful information is often much lower than expected and that once someone meets some threshold of expertise, their ability to give information acts very much like a step function. Figure 5 shows his representation of this trend. Armstrong was looking specifically at forecasting a particular outcome from some level of knowledge, and the same trend may not be true for performance modeling. He also notes that in some cases there can be such a thing as too much expertise where individuals are less accurate, despite having more knowledge and experience. [4]

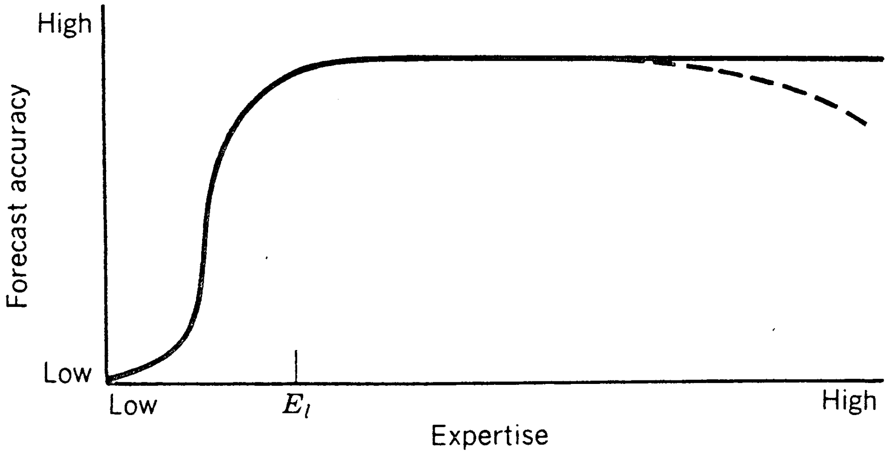


Figure 5: Level of Expertise Required for Accuracy [4]

Even if an individual has the perfect amount of expertise, this does not assure that they have the necessary knowledge relevant to the problem at hand. And even if they have the necessary knowledge, it does not guarantee that the information they give will be correct. A world-class doctor may misdiagnose a patient, a Professional Engineer may still make miscalculations and the engineers using this method may have individual biases or incorrect assumptions about a particular problem. The method itself must be robust enough to be able to handle individual errors while still providing reliable results.

2.2 Survey of Expert-Based Methods

The most basic and fundamental way to get expert-based information is to ask a supremely knowledgeable person to give the “right” answer to a particular question. Assuming the expert meets the qualifications discussed above, the response has a good chance of being correct. But for this information to be trusted with large budgets and major decisions, it needs to be validated by one or more other experts. Also, just as the results from scientific experiment that works only once cannot be accepted, if the process cannot be repeated with different (but equivalent) experts, the results should not be accepted either. Recalling the barometer anecdote from Section 1.3, the process needs to be repeatable for the process to be universally accepted.

In an effort to make results more consistent and reliable, expert-based methods have become more structured. These structured methods can be applied by different practitioners with different groups of participants to the same problem and should get similar results.

In general, these methods fall into one of three classes. The first class includes methods based around assessment, judging or scoring. This includes activities like classifying colors, scoring artistic performances, or medical diagnoses. There may be guides and rubrics, but the final output comes down to assigning a qualitative or quantitative description or score to something that is more abstract. The second class describes methods that predict values or outcomes. These methods use intuition and experience to estimate the outcome of some future event or how a particular thing or system of things will behave. Examples of this would include weather forecasting, risk assessment, and estimating how long it will take to travel to a location. These approaches frequently use a set of given input data beyond the individual's own observations in concert with individual experience. The third class are relationship-based methods that attempt to capture the way a set of measures interacts and impacts another set of measures. These methods tend to be limited to engineering and manufacturing problems.

The classes get closer to emulating traditional model in this order, with assessment methods being the least similar to traditional modeling and relationship-based methods being the most similar. The first two classes tend to be the bulk of the focus for most research. The limited applicability of relationship-based methods means that they are not as broadly researched outside of engineering, manufacturing, and occasionally business fields.

2.2.1 Methods in Everyday Use

There are several areas in everyday life where expert-based methods are the preferred choices for analysis. Medicine is a common example of a structured expert-based method. Medical school teaches doctors to be able to use standardized tests and symptoms to recognize illnesses reliably. Even so, for major diagnoses, the result is validated against another doctor's second opinion. To help doctors and to aid lay people, attempts have been made to create models based on this knowledge. WebMD is an example of a website that allows a person to input his or her symptoms and it will generate a list of potential ailments[130]. Unfortunately, the human body is still too complex and symptoms are too subtle for a model to take the place of the doctor. A typical set of symptoms may give 10-20 common ailments, some of which share similar treatments and some of which do not. Psychiatry has a similar problem, but with the difficulty of less concrete symptoms and few, if any, objective tests. The effort to structure the process of psychiatric diagnosis by the American

Psychiatric Society resulted in the *Diagnostic and Statistical Manual of Mental Disorders* as a guide.

Other areas attempt to use expert-based methods to predict the future based on past performance. Stock market analysis and investment planning is still primarily left up to individuals managing individual investment funds. While models and simulations have attempted to outperform the opinion of experts, individuals such as Warren Buffett have consistently outperformed the market average and those computer programs. The outcome of various sports games, matches and competitions as well as specific tournament performance, such as the National Collegiate Athletic Association men's basketball tournament, are frequently predicted by paid commentators. While much of the discussion and predictions may be primarily for entertainment purposes and are not always correct, it is an example from every day life.

In other areas, modeling has overtaken expert-based methods as the gold standard. Weather forecasting was once the responsibility of mystics until the patterns were understood well enough that meteorology came forth as a science. Using observations from geographically distributed locations with various instruments gives meteorologists the data necessary to combine with experience and their own expert intuition to create a forecast for the next few days. With time, computer models have gotten faster and more accurate and meteorologists have largely become the individuals responsible for developing new models, improving existing models, and interpreting and comparing the outcomes of different disagreeing models. At present, very few meteorologists continue to incorporate their own opinion into their forecasts. Those that do tend to be those with decades of experience with the weather within their particular region and in cases where localized but persistent effects have not been incorporated into detailed models. [68, 118]

The problem of travel routing and the time associated with travel was a common problem for vacationers. Locals who had learned the behaviors of traffic within a small region could offer expert opinion in the form of directions and estimated travel time. With the advent of computerized mapping software and satellite navigation, routing could now be performed automatically without needing to consult experts. Travel time was still hard to predict automatically until mass polling of traffic allowed these models to estimate travel speed at certain times of day based on historical or real-time information. Navigation units and mapping web sites are now able to estimate the travel time of different routes and select the route based on shortest actual travel time rather than defaulting to the defined speed limits.

2.2.2 Methods Used in Engineering

Expert opinion is an essential part of any model development exercise. The process of developing a model is more than copying equations from a sheet of paper or textbook into a series of programs and waiting for a computer to compute the results. Experts are necessary to ensure that relationships are applicable and that associated assumptions are valid. After creating a model, an expert is necessary to verify that it behaves as it was expected and that reasonable results are produced when reasonable inputs are used. Recalling the barometer anecdote, the student could ask the previous building superintendent the exact same question and compare their answers. If both superintendents give the same, or a very similar, answer, the answer has been validated and the method has shown itself to be repeatable. The goal of a well-accepted expert-based method is to demonstrate that the results have been validated and that the process is repeatable enough that the answer would continue to remain constant if different experts were asked, within some margin of error. Further discussion of experts in verification and validation can be found in Section 3.5.

2.2.2.1 To Identify Preferences

Expert-based methods are a common source to identify weightings or preferences. Multiple attribute decision making (MADM) methods use preferences to trade between several desired attributes. If no single physical relationship combines those attributes, it is essential for a decision maker (serving as the expert) to make those trades themselves. MADM approaches are often incorporated into interactive tools known as decision support systems. These systems combine the information from the decision maker with more complex supporting data in the background to produce graphical and intuitive suggestions and decision guidance.

One of the methods for incorporating expert information is the analytic hierarchy process (AHP), a well-known method to support decision-making. It does this by building a hierarchic view of the problem, the possible solutions to that problem, and the criteria associated with those solutions and problem. The solutions are given scores based on their criteria and the criteria are related to the goal. The criteria can be decomposed into subcriteria that split the weightings of their parent criteria. The result of an application of the method is an ordered list of alternatives. The solutions generally need to be a small number of choices, which makes it inappropriate for design space exploration where a traditional model would be used. [111] Since its introduction, a great deal of research has specialized the mathematics associated with AHP to apply it to logistics, manufacturing, business and higher education fields. It has been integrated with linear programming to solve problems directly, meta-heuristics, and data envelopment analysis. [69]

Preferences can be applied more directly with experts serving as the evaluator of a system directly. Buonanno used experts as the objective function in a genetic algorithm. His approach, known as the Hybrid Interactive Genetic Algorithm, showed graphical depictions of an aerospace vehicle (specifically supersonic business jet) to experts who then gave each design a qualitative score on a 5-value scale from “bad” to “best”. The hybrid aspect of the method was that it also included an analytic physics-based model behind it to more intelligently select which concepts to show to the experts. In this case, a great deal of experience with supersonic vehicles was necessary to be able to identify the best designs based solely on planform geometry. [25]

There are many other methods for utilizing experts to elicit or identify preferences or goals, but in general, these methods are simply more structured ways for an expert to defend a decision that has already been made.

2.2.2.2 *Forecasting and Reliability*

For extremely complex problems or problems with a great deal of unknown information, human judgment and intuition are frequently able to fill in the gaps. In these cases, even high fidelity modeling cannot account for all the interactions or accommodate the large numbers of unknowns. The most common problems here are those that involve predicting the future with some amount of certainty, such as economic forecasting or estimating reliability of complex systems. Weather prediction predicts the future in a way, but can only do so reliably for a few days. Longer ranged forecasts, such as those a few months or a few years out, are still beyond present scope.

All expert-based methods that include more than one participant seek, in some way, to build consensus between the experts or at least produce a result that balances the information from each expert. However, for methods that are focused on forecasting future events, the consensus itself is the primary goal, sometimes balanced with robustness. The best known method for building consensus is the Delphi method. Originally developed by the RAND Corporation shortly after World War II, the goal was to estimate the impact of technology on future conflicts. Much of the original research it was used for was classified and was not openly released until the early 1960s. The original method used a series of paper surveys to collect information on a particular topic from a group of experts. Once the surveys were collected, the information was collated and then redistributed along with another survey to the same experts to give them an opportunity to update their estimates. This process continued until the estimates converged within an acceptable margin. [39]

Since that time, the Delphi method has been expanded to a wide variety of other uses and has been updated to make use of digital technology. Its widespread use has resulted in many methodological advancements, but the method still revolves around being an off-line, iterative method for polling experts about a small number of specific values.

The Cooke method is similar to the Delphi method in that it is an off-line method for polling experts about a small number of values. However, it is not iterative and always requires a three-point estimate for each value: the value and the 5th and 95th percentile confidence bounds. This separates each expert's estimate and the expert's certainty. Another standard part of this process is to ask a series of control questions that have known historical values. The experts' estimates and bounds are then combined based on a weighting such that those experts who showed themselves to be most reliable on the control questions have the greatest influence on the combined result. Obviously this is limited to cases where pre-existing data is available on a very similar problem as the one for which the exercise is being performed.[6]

There are a number of other methods for collecting expert opinion for the purpose of forecasting future events. Bedford *et al.* performed an excellent survey of methods for estimating future reliability of systems during the systems engineering process. Their analysis considers elicitation exercises through the design process and found that, when using experts in concert with analysis, the feedback loops provided much more accurate design decisions for problems where reliability was a driving factor.[11] A particularly unique concept are prediction markets where forecasts are formulated within a forum based on the stock market. This approach has been shown to be very accurate, but is limited to a small number of simple responses, such as elections.[13]

2.2.2.3 Relationships Between Inputs and Outputs

The previous two classes of expert-based methods are responsible for generating the inputs to a process (in terms of priorities) or outputs of analysis that does not exist (for reliability and forecasting). Model development is the process of identifying and determining the relationship between the inputs and outputs of that model. There are fewer methods that are intended to capture those relationships without creating a full physics-based (or other theory-based) model. A relationship-focused method is the best chance for model-development. Quality Function Deployment (QFD) is the most widespread method for relating top-level priorities to lower-level product and process characteristics. The Strategic Portfolio Planning (SP2) methods are based on a QFD framework but have been specialized to technology and portfolio planning. The Portfolio Analysis Tool (PAT) is a bottom-up method designed around being easy to interpret visually. Since these methods are the most applicable, they merit a more detailed discussion, found in Sections 2.2.4 through 2.2.6.

2.2.2.4 As Part of Integrated Product/Process Design

Integrated Product and Process Development (IPPD) is a systems engineering or management approach to improve customer satisfaction by integrating all acquisition activities around a multidisciplinary team of stakeholders known as Integrated Product Teams (IPTs). It grew out of Concurrent Engineering and became part of the best-practices for many companies in the late 1980s and early 1990s. The United States Secretary of Defense mandated its use throughout the Department of Defense acquisition process in May 1995[99]. The Joint Strike Fighter project was one of the first major military programs to use IPPD and was widely praised as an example of the benefits of IPPD[100]. There are a number of different flavors and processes depending on the company and industry that developed them. The version of IPPD developed and taught at the Georgia Institute of Technology is shown in Figure 3 on page 10 [114].

IPPD commonly uses expert-based methods including the Seven Management and Planning Tools and Quality Function Deployment as part of requirements definition process. This step is shown in the upper left of Figure **??** where it is the links between steps of *Establish the Need* and *Define the Problem*. Because this part is focused on eliciting the customers' requirements and preferences, any method used here must be built around the idea of using expert information (with the assertion that customers are experts at knowing what they want). Several of these methods also include input from designers, managers and experts in other areas to give a unified view of the problem. Though much of the literature about IPPD came out of the 1990s, it is still used on aerospace problems today[115, 48].

It is worth noting that, in the Georgia Tech IPPD process and in others, the information gathered in the requirements definition portion is used to define the problem and create a list of requirements but is not directly used after that step. It may be used again when iterating or revisiting requirements, but once requirements are set and agreed upon, they are the only information that remains from that exercise. Reusing this information during the steps of *Establish Value*, *Generate Feasible Alternatives*, and *Evaluate Alternatives* is another major motivating factor for this research. If this information already exists and can reduce the level of effort required later, there is that much more reason to incorporate it into a model.

2.2.3 Expert Models and Expert Systems

In searching the literature, the terms “expert model” and “expert systems” are used extensively. At first glance, one would expect that they are similar to expert-based models and expert-based methods. While they do include information from experts, they focus on capturing the decision-making abilities of the experts or their behavior under certain conditions. The resulting model is intended to be a replacement for an expert within a larger model or process as an artificial intelligence[129]. This field also includes knowledge systems and knowledge engineering. The test of the accuracy of one of these types of models is how well the model replicates the expert’s behavior where the test for an expert-based model within engineering design is how well the expert and the resulting model replicate true physics or a higher fidelity model. While expert models rely on expert knowledge to create and hone their accuracy and realism, the objectives are sufficiently different that the research cannot be easily transferred. Within this document and research, the preferred terms remain “expert-based models” and “expert-based methods” for models and methods that use experts as the source of information without intending to model the experts themselves.

An extension of models of expert systems are used to capture the process side of product manufacturing. These methods attempt to capture how exactly a product will be built to most closely match the design specifications. Where product modeling and design determines what you want, process modeling describes how you get it. With the addition of computer simulation and computer-aided graphics, this has developed into the field of virtual manufacturing where the entire manufacturing process can be defined and explored prior to any tooling including the actions, abilities, and judgements of the experts. As manufacturing methods have gotten more complex themselves, they have had a greater impact on the ability to meet tighter tolerances, individual unit cost, and learning-curve effects. All three of these are important aspects of final performance, but are extremely difficult to account for in the early phases of design. The uncertainty present in a conceptual design are larger than any manufacturing tolerances. The cost and learning-curve effects are important to correctly estimate what a customer will be expected to pay for the finished product, but any modeling that requires virtual manufacturing is far to detailed for the rapid modeling this research is attempting to provide.

Some expert-based method, such as QFD, do account for some aspects of process development in a very generic sense.

2.2.4 Quality Function Deployment (QFD)

Quality Function Deployment, almost universally referred to by its abbreviation QFD, was first developed by Yoji Akao in Japan in the late 1960s to improve the “quality of design” and support statistical quality control (which would later become total quality control). It came out of a desire to begin designing quality into products and to develop quality control process charts prior to the design reaching manufacturing. It was limited in use to Japan until the method traveled to the United States in 1983 where it was embraced and quickly became well-published and researched. The name refers to its process of decomposing (deploying) the problem to account for the sources of quality (or sources of failure to produce quality) from the customer's requirements down to manufacturing characteristics. [2] In practical usage, QFD translates customer requirement priorities into prioritized technical goals for development and production through a series of relationship matrices.

The QFD process is inherently hierarchical. The top-level customer requirements are related to engineering characteristics or technical measures. These engineering characteristics are then related against parts characteristics. Parts characteristics are related against the process or manufacturing operations needed to produce those parts. And lastly, the process operations are related against production requirements. This process is illustrated in Figure 6 [30]. This is the standard form that is most commonly used for product development, but there are frequently customizations to suit a particular problem. Many early engineering design problems, especially those that are performed for the sake of design rather than for the sake of manufacturing a product, only focus on the first set of relationships and ignore manufacturing concerns. For large and extremely complicated products, the problem may be decomposed and split among many other metrics, including parts failure modes, multiple levels of engineering characteristics, cost effects, and product functions to produce a matrix of matrices commonly known as the Thirty Matrix Approach. [110]

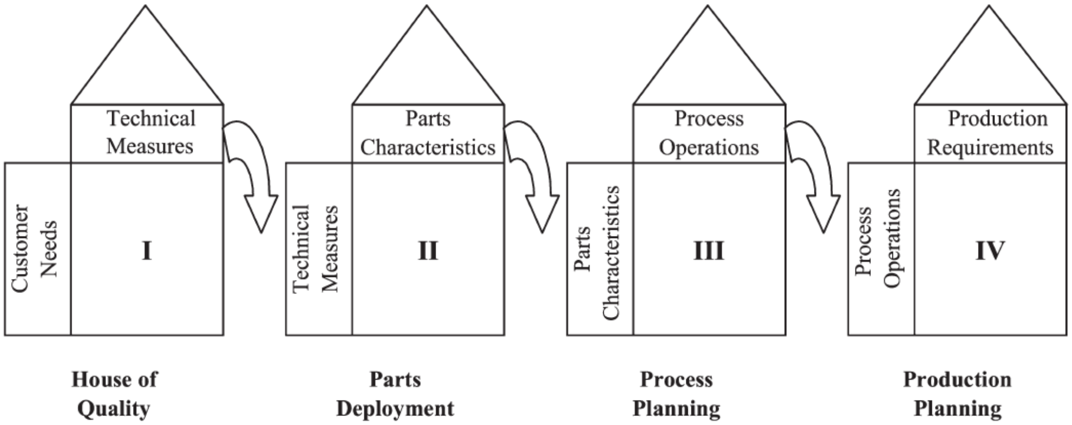


Figure 6: Four Phases of a QFD [30]

Each of these levels is useful in itself, especially when starting from the top. Each level, particularly the top one, can be known as a House of Quality (HoQ). Figure 7 shows the standard parts of this house[124]. The "Hows" and "Whats" vary with each level according to the particular decomposition. In this particular level, they correspond to the customer attributes or requirements and the technical requirements or engineering characteristics, respectively. The customer assessment is also known as a competitive assessment where a producer ranks competitors products in the same customer requirements it is using to judge its own product. The importance values of the customer requirements are used to prioritize particular aspects of a design over others.

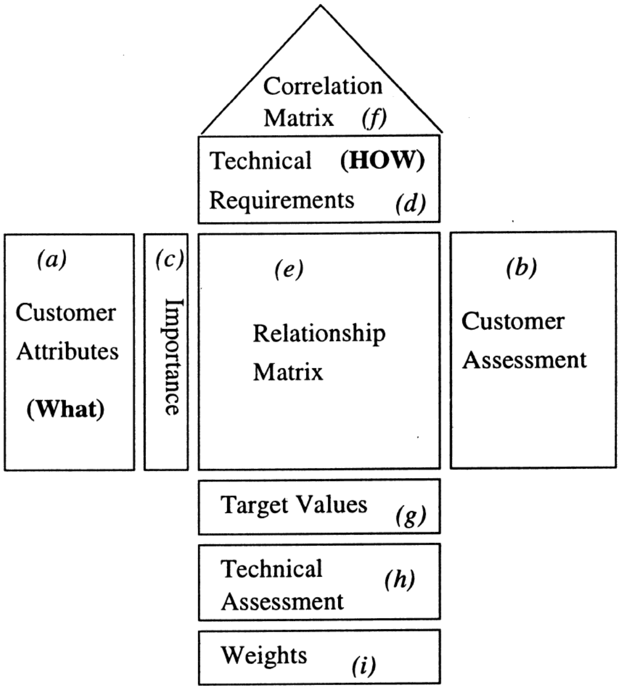


Figure 7: Parts of a House of Quality [124]

The central matrix, the relationship matrix, is where the relationships between the "Whats" and the "Hows" are mapped using qualitative scores. Traditionally these scores are symbols based on symbols used in horse racing for win, place and show. These roughly correspond to describing the relationships as "high", "medium", "low" and "none" (or "strong", "moderate", "weak", or "none"). The "roof" of the HoQ describes the correlations and trades between the "Hows". Though not shown here, more recent versions include a "greenhouse" to describe the correlation between the "Whats". When deploying to the next level down, the roof becomes the greenhouse of that level. The target values are early estimates of what the goals for the engineering characteristics should be and the technical assessment is an estimate of how difficult it will be to meet them. Lastly, the weights are produced by matrix multiplication of the vector of importances of customer requirements times a numeric representation of the relationship matrix.

There are several approaches for creating the QFD. One of the more structured methods is with the use of the Seven Management and Planning Tools. These seven tools include an affinity diagram and the tree diagram to identify the "Hows" and "Whats", the interrelationship digraph for generating the correlations matrix, the matrix diagram for creating the central relationship matrix, and a process decision program chart and activity network for program support. [20] With time and experience, a small group of professional can skip some of these steps and perform a QFD more directly.

Since its original development, the QFD has become a field unto itself and expanded with a great deal of research. Its versatility has allowed it to be adapted to a wide variety of uses beyond quality control issues associated with manufacturing. It has been readily incorporated into requirements definition and Integrated Product/Process Design and is widely used within engineering. Raytheon Missile Systems has recently integrated QFD as a central part of its Mission System Engineering process [101]. It also is relatively simple to perform. This versatility, widespread use and simplicity makes it an excellent choice as a baseline. Unfortunately, the information flow in QFD is entirely from the priorities at the top level down to priorities at the lower levels. For predictive modeling, the information must flow from the bottom up.

While the usefulness of a QFD has been repeatedly demonstrated, the accuracy has not. The improvements in quality and manufacturing reliability for products developed and created using QFD is the most common measure of success of a QFD. Other groups find that any perceived gain in insight that participants have is beneficial enough to repeat the process for future products, whether or not the actual product is improved. What is not tested is whether or not the relationships present and the prioritizations that result from them match up to true behavior or importances. Since this research will be focusing on the ability to produce an accurate model, it will be necessary to develop an approach to test those relationships.

2.2.5 Strategic Portfolio Planning (SP2) & Strategic Optimization for the Allocation of Resources (SOAR)

The Strategic Portfolio Planning (SP2) process is another relationship-centric method, but used for technology portfolio planning and allocating resources. It was originally developed by Chris Raczynski and Michelle Kirby [75]. Raczynski further developed the method by improving the optimization techniques and refining other aspects to changing many minor aspects of the method into the Strategic Optimization for the Allocation of Resources (SOAR) process [107]. The two methods are nearly identical in the ways that are relevant to this research. While SP2 is the original method and is better known, SOAR is documented in greater detail. The two methods will be referred to as a set as either just SP2 or SP2/SOAR throughout the document.

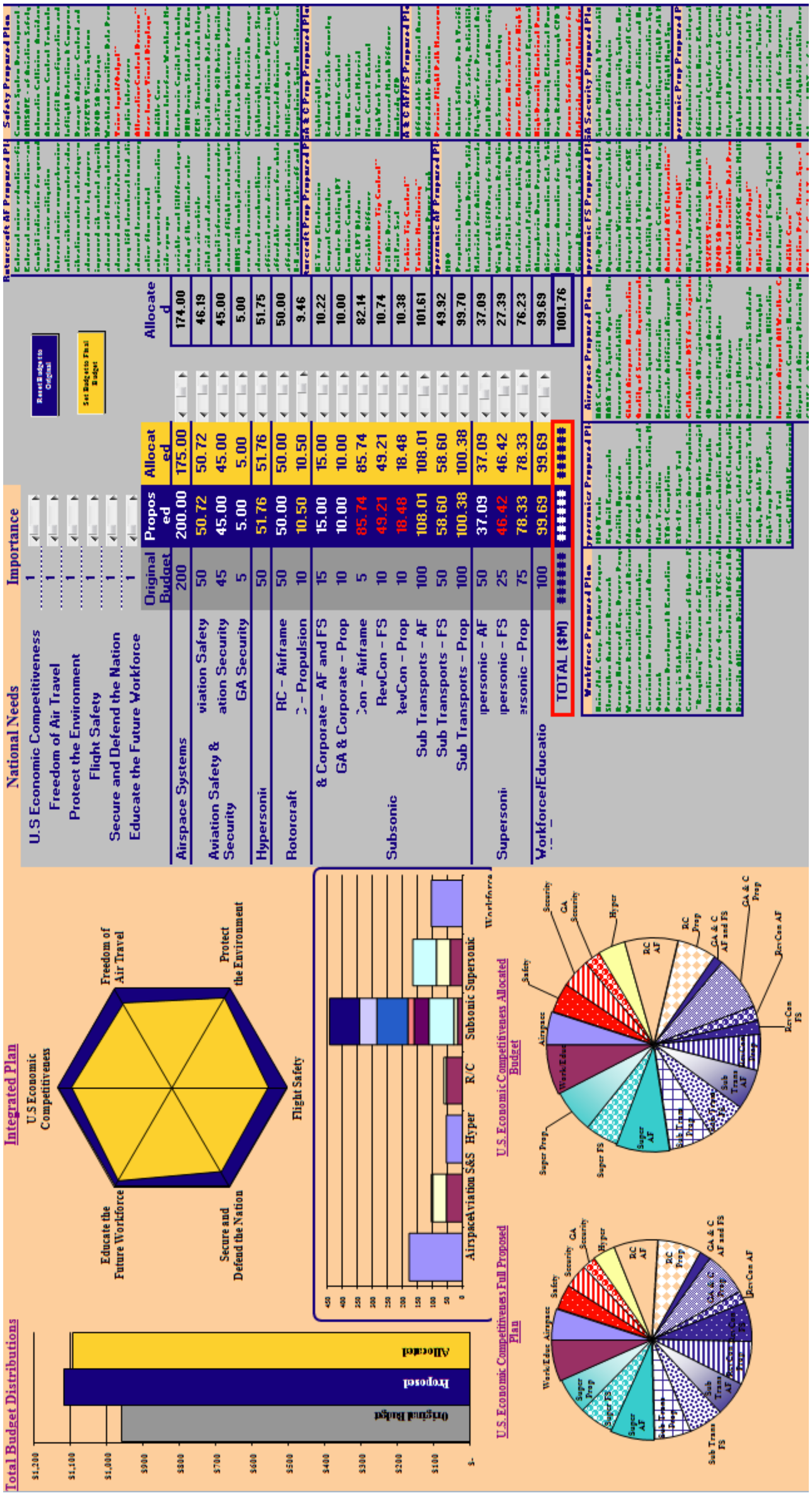
The methods are based on a QFD framework, using it to decompose the high-level goals of an organization into lower level goals which are then mapped to technologies and programs. The prioritizations on the technologies are then used to drive an optimizer to solve the knapsack problem with a budget (and sometimes time) constraint. The technologies are scored based on a set of surveys to the technologists responsible for them or who would be most knowledgeable about them. The relationship matrices are completed using one or more workshops where participants use electronic voting devices to provide a score for each individual relationship. After the voting process, the votes are analyzed. When the majority of the experts agree, the scores are accepted. When there are large disagreements or unusual distributions, the matter is opened up for discussion to allow participants to explain their reasoning and build consensus by agreement.

The relationship matrices, technologies, and optimization are combined into an interactive visual tool to support decision makers and further analysis. These decision support tools are frequently referred to as "calculators" and are useful for understanding problems long after their initial development. An example of one is shown in Figure 8 [108]. This interface makes it easy for non-technical managers and decision-makers to understand the trades between the top-level organization priorities and available budget.

There are slide bars in the center of the interface to allow decision makers to adjust each of these and see the perturbations in which technologies are selected on the right side of the interface. The left side of the interface shows typical summary graphics. Of particular interest is the radar-gram, since this depicts a measure of how well the current solution set meets the goals. Generating these values requires flowing up information, an essential capability needed for the modeling desired.

Since publication of SP2 and SOAR, further research has developed further incremental improvements that have yet to be published. Such improvements include constraints on available manpower, more detailed classification of available resource, and the inclusion of product development projects.

The ability for SP2/SOAR to perform bottom-up information flow is a significant motivation to consider using them as a baseline for further modeling. The tested improvements for workshop information collection are a useful starting point for selecting methods for this research. The down-side of SP2/SOAR is that they are very much focused on solving the technology selection problem. In order to use it as a starting point for system modeling, significant portions of the methods would have to be ignored or removed. It would be far easier to start with QFD as a common ancestor and pull individual improvements from SP2/SOAR as needed.


Total Budget Distributions
Original Budget
Proposed
Allocated
Integrated Plan
U.S Economic Competitiveness
Freedom of Air Travel
Protect the Environment
Flight Safety
Secure and Defend the Nation
Educate the Future Workforce
Airspace
Aviation S&S
Hyper
R/C
Subsonic
Supersonic
Workforce
U.S. Economic Competitiveness Full Proposed Plan
U.S. Economic Competitiveness Allocated Budget
Reset Budget to Original
Set Budget to Final Budget

National Needs | Importance
U.S Economic Competitiveness 1
Freedom of Air Travel 1
Protect the Environment 1
Flight Safety 1
Secure and Defend the Nation 1
Educate the Future Workforce 1

		Original Budget	Proposed	Allocated	Allocated
Airspace Systems		200	200.00	175.00	174.00
Aviation Safety & Security	viation Safety	50	50.72	50.72	46.19
	ation Security	45	45.00	45.00	45.00
	GA Security	5	5.00	5.00	5.00
Hypersonic		50	51.76	51.76	51.75
Rotorcraft	RC - Airframe	50	50.00	50.00	50.00
	:- Propulsion	10	10.50	10.50	9.46
Subsonic	& Corporate - AF and FS	15	15.00	15.00	10.22
	GA & Corporate - Prop	10	10.00	10.00	10.00
	;on - Airframe	5	85.74	85.74	82.14
	RevCon - FS	10	49.21	49.21	10.74
	tevCon - Prop	10	18.48	18.48	10.38
	Sub Transports - AF	100	108.01	108.01	101.61
	Sub Transports - FS	50	58.60	58.60	49.92
	Sub Transports - Prop	100	100.38	100.38	99.70
Supersonic	ipersonic - AF	50	37.09	37.09	37.09
	ipersonic - FS	25	46.42	46.42	27.39
	ersonic - Prop	75	78.33	78.33	76.23
Workforce/Educatio		100	99.69	99.69	99.69
	TOTAL ($M)	######	######	######	1001.76

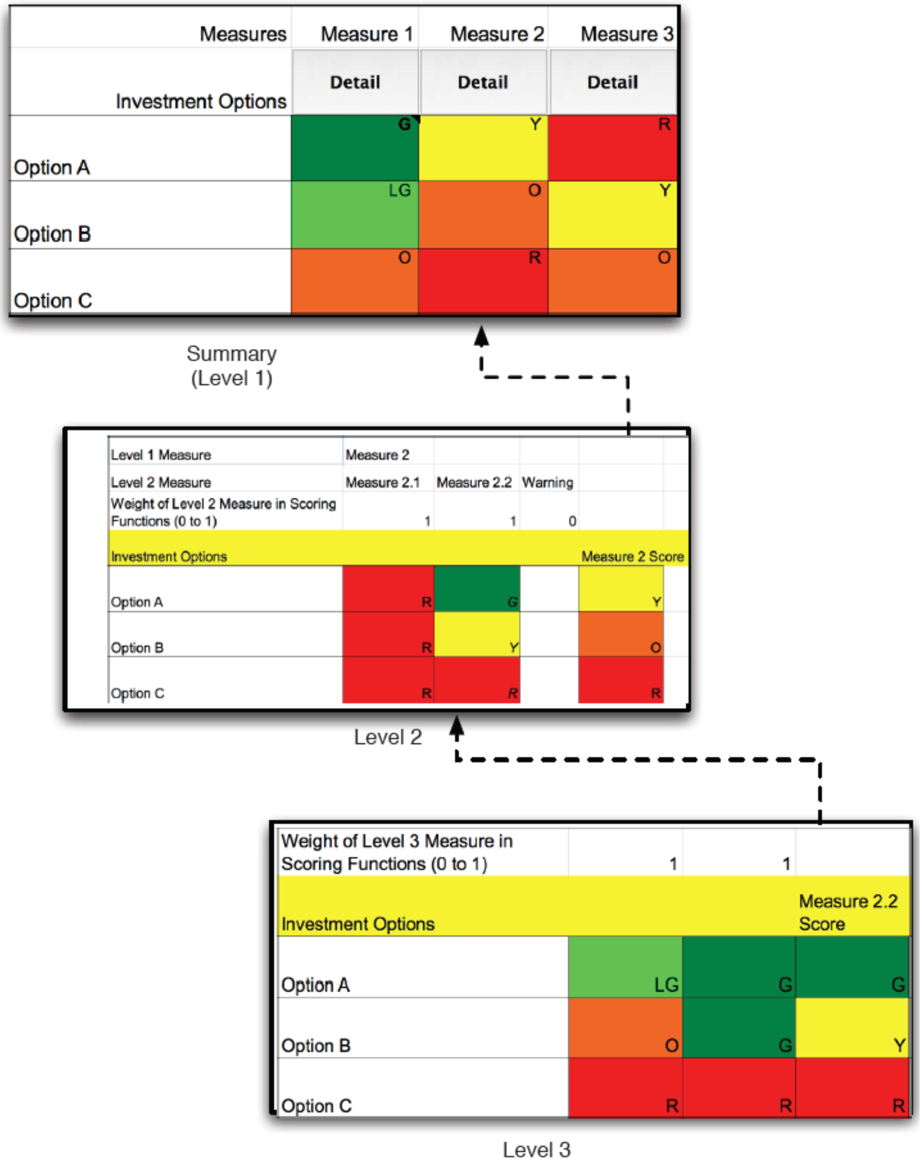
Figure 8: Example SP2 Decision Support Tool, reproduced from [108]

2.2.6 Portfolio Analysis Tool (PAT)

The Portfolio Analysis Tool (PAT) is an Excel-based tool and associated method for comparing a portfolio of options or alternatives based on heuristic measures and submeasures. It was originally developed by Paul K. Davis and Paul Dreyer in 2005 as a missile defense portfolio analysis tool to produce analysis for executive-level decisions-makers [44]. It was designed to use data from system specifications, test data, models, or expert judgment. Its use of expert judgment with a bottom-up flow is what makes it of interest here. As a portfolio analysis tool, it assumes the presence of a set of options with associated costs as an input. If this were extended as a general-purpose bottom-up model, this set of options would include the infinite set of points (that could be represented by a large finite number of samples) within a particular design range of interest.

PAT was designed from the beginning to flow up information into graphical summary tables to make it easy to assess options at a glance. This makes the results simple to interpret, but at the cost of detail and number of possible options. The top-level measures of interest are decomposed into sub-measures, which may be further decomposed into sub-sub-measures. For the lowest level measures of the tree, each alternative is scored using either an actual value or a subjective score. Actual values are mapped to scores by testing whether or not it meets a threshold and then assigning a score based on where it lies between the threshold and the goal. Any value not meeting the threshold or exceeding the goal is given the minimum or maximum score, respectively. These scores are mapped to stoplight colors, as shown in Figure 9. [41]

The top-level measures are aggregated into a single effectiveness score between zero to one for each option in the portfolio. Cost is included as a separate metric that allows a cost effectiveness score to be calculated. The tool also includes support for applying a budget over time to track effectiveness over time. The use of colors and simple results plots make it work very well for summarizing data for a small number of measures and a small number of options. However, as both the number of measures and the number of options increases



Measures	Measure 1	Measure 2	Measure 3
Investment Options	Detail	Detail	Detail
Option A	G	Y	R
Option B	LG	O	Y
Option C	O	R	O

Level 1 Measure	Measure 2			
Level 2 Measure	Measure 2.1	Measure 2.2	Warning	
Weight of Level 2 Measure in Scoring Functions (0 to 1)	1	1	0	
Investment Options				Measure 2 Score
Option A	R	G		Y
Option B	R	Y		O
Option C	R	R		R

Weight of Level 3 Measure in Scoring Functions (0 to 1)	1	1	
Investment Options			Measure 2.2 Score
Option A	LG	G	G
Option B	O	G	Y
Option C	R	R	R

Figure 9: Flowing Up Information Within PAT, reproduced from [41]

to what would be necessary for design space exploration, the use of stoplight charts as summary statistics becomes infeasible.

Also, PAT's fundamental requirement of threshold values at each level of the model to flow up information is useful for concept selection or comparison, but not for design space exploration. For variables and metrics which must always be maximized or minimized, the requirement of foreknowledge of the goals is problematic. Many design problems also include variables that aim for a particular target or sweet spot of optimality that would be more difficult to capture in PAT.

These limitations along with the limitations of working within the existing tool and the difficulty of expanding it, make PAT an unappealing candidate as a starting point for expert-based modeling. While it would be possible to correct many of these limitations, the effort associated would be similar to starting from scratch.

When considering which of these three choices, if any, to use as a basis for expert-based modeling, it is desirable to minimize the number of changes that must be made. PAT would require significant number of changes and would still likely be limited in its application and usefulness. SP2/SOAR demonstrate a number of the needed qualities for expert-based modeling. However, the method comes with the baggage of being tightly tied to portfolio planning and the methodology requirements that go with it. QFD, though more primitive, has a great deal of history and familiarity behind it that provide an excellent starting point. That it has been expanded and used as a base for many other methods is promising And since QFD is the ancestor of SP2/SOAR in many ways, it is possible to back-port some of the desired features to the proposed method. For these reasons, QFD has been selected as a point of departure for this research.

2.3 Advancing QFD

Once QFD has been identified, it is important to identify what changes are necessary to produce the desired models. One of the first issues to note is that QFD is built for general product development, not product design. Development is the process of identifying, prioritizing, and reaching goals from first mental concept to items on shelves (or the equivalent for aerospace systems). Design is certainly part of development, but instead focuses on identifying the value of physical characteristics that produce a certain performance or capability. Especially at the conceptual level, this revolves around sizing and synthesis.

The first difference is the choice of what items are included at each level. Section 2.2.4 identifies the standard four relationship matrices and five levels of attributes associated with a typical QFD. These were primarily manufacturing-focused and not as useful for problems within early phases of engineering design where manufacturability is less pressing. Table 1 identifies a translation of the first four levels to attributes more suitable for engineering design. The top level remains customer requirements, but may be generalized to system requirements. These are the top-level cost and usage descriptors. These are the values that tend to come from models that incorporate economics and usage scenarios.

Table 1: Translation from Traditional QFD to Engineering Design QFD

Traditional QFD	Engineering Design QFD
Customer Requirements	Customer or System Requirements (R)
Engineering Characteristics	System Performance or Intermediate Metrics (m)
Parts Characteristics	Design Variables or Subsystem Performance Metrics (x)
Process Operations	Subsystem Design Variables

The second level has transitioned from the frequently vague and qualitative engineering characteristics to system performance metrics or intermediate metrics. The primary difference between characteristics and metrics is that metrics are defined as a measurable quantity while characteristics may not be. These are the outputs of typical aerospace vehicle performance analysis.

The third level has transitioned from parts characteristics to design variables and subsystem performance metrics. The previous levels used interchangeable terms, but this level includes both terms. Design variables are those physical, often geometric, quantities that designers control directly. Subsystem performance metrics are quantities that are not controlled directly, but are the result of the design of a subsystem. A common example would include engine performance metrics.

The fourth level allows for the decomposition of the subsystem performance metrics into subsystem design variables. The variables here would only map to the associated subsystem performance metrics and not to the design variables in the previous level. For problems that do not require detailed subsystem design (typical of very early phase design), the fourth level may not be used.

The mappings between each of these levels need to capture the driving effect rather than the correlation between the two. This means that when an expert provides data for the relationship between a system performance metric and a customer requirement, the expert should only include the direct causal relationships as much as possible.

The next sections discuss the efforts of others to advance QFD and QFD-based methods to be more useful, accurate, and applicable. The first section discusses the scope of literature and narrowing it. The sections following it focus on several major areas of improvement selected based on their acceptance or applicability. Additional methods beyond these improvements will be brought up in Chapter 3 to solve specific problems as they are identified.

2.3.1 Narrowing the Scope of Literature

QFD as a discipline has become quite extensive and sizeable. The versatility of QFD has caused research to spread out over many fields. As a measure of the size, Carnevalli and Miguel studied 157 publications over just the seven year span from 2000 to 2006 in their survey of the field[29]. Chan and Wu lamented that their 2001 survey of QFD literature was incomplete with only 650 publications[31]. A search on Google Scholar for "quality function deployment" lists 23,000 papers, books or other publications about or referencing QFD[58]. The US-based QFD Institute, the QFD Institut Deutschland and other regional quality engineering and QFD groups each hold annual conferences solely for topics directly related to QFD. Journals such as *International Journal of Quality and Reliability Management*, *Quality Engineering*, *Quality Progress*, *Total Quality Management*, and others include QFD-related research as a considerable portion of what they publish. This is to say that the realm of research that has been performed and is still being performed with QFD is beyond the scope of what could be considered for any single survey.

It is important to narrow this field to highlight the past research that will be most useful here. Many papers describe case-studies or specific implementations of QFD on a particular problem. These capture the individual successes or failures and may include some information on lessons-learned from the particular group. While these are useful lessons-learned, they focus on how to do QFD rather than advancing the state of the art. They also tend to be come to many similar conclusions.

There are several improvements that are widely used in new QFD exercises beyond the traditional House of Quality approach. These help to account for deficiencies in judgement, reduce variability of results to noise in expert information, and extend the usefulness of QFD as a discipline. The next few sections discuss these approaches and their applicability to expert-based modeling.

2.3.2 Fuzzy Logic

A common problem with expert-based information, especially qualitative and categorical information, is that words have slightly different meanings to different people. A common example involves asking several people to describe the temperature of the room. One person's "warm" may be another person's "hot" and one person's "comfortable" may be another person's "cold", even if all four are sitting in the same room. A similar effect occurs when asking individuals to describe the relationships in a QFD as high, medium or low. The effects of both are that, even though the numeric temperature is constant for all, the apparent temperature has a great deal of variation. Using the apparent temperature to determine the physical temperature introduces a great deal of uncertainty.

One solution to reduce uncertainty is to require all experts to agree on a single descriptive word. This process can take a while and may not always be possible. Another, more technical solution is to give each descriptive word a fuzzy meaning rather than a precise one. Each term is given a membership function where, for that term, a range of assigned value is likely. This is often centered around a particular value and decreases in probability further away from that value. Figure 10 shows three of these membership functions where the vertical axis is the utility coefficient and the horizontal axis indicates the value. In the same plot is a bolded line that shows the aggregate of the three, assuming that three experts each selected one of the terms. [127]

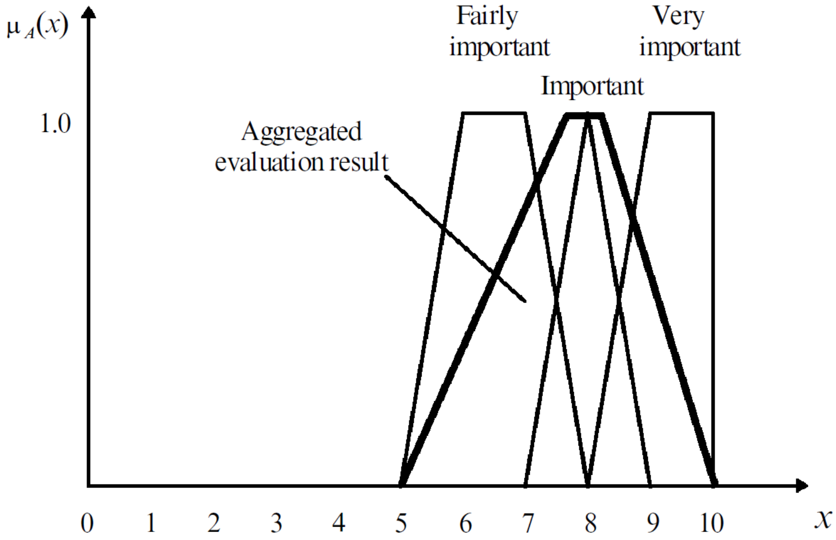


Figure 10: Fuzzy logic membership functions and their aggregate [127]

Once membership functions have been defined and applied, a mixed integer linear program can be used to optimize the engineering characteristics instead of the typical matrix math prioritization [134]. Temponi found that using fuzzy logic allowed a poorly defined and time-consuming requirements and relationship definition process to be more mathematically rigorous and faster (since less agreement between participants is necessary) [124]. The benefit here comes only when still using and combining non-specific scores. This approach is also intended to focus on creating a single estimate of the value rather than finding a numeric combination of multiple experts. While traditional QFD revolves around these scores, for an engineering model, more specific, defensible scores may guarantee that an expert provides the data that says exactly what they mean without the need for an intermediate translation.

2.3.3 Analytic Hierarchy Process

The prioritization values for customer requirements are often determined by experts, making the rankings at lower levels of QFD to be a complex translation based on a group voting scheme. Hazelrigg suggests that this means that QFD results and rankings are therefore bound by Arrow's Impossibility Theorem and that, as a result, are guaranteed to be erroneous [67]. Arrow's Impossibility Theorem states that no group voting scheme for making a social choice can be fair while still meeting all five of his conditions [5]. Saltmarsh performed an analysis of the effects of this on QFD-based portfolio management voting systems that showed the possibility of solutions changing order as a result of changing the number of participants and nothing else [112]. Separately, Anderberg claimed that individuals are unable to describe the differences (or provide independent weightings) between multiple objects accurately, but that individuals are only able to accurately rank items in a particular order [3].

Several individuals identified the Analytic Hierarchy Process (AHP) as a potential solution that would reduce the variation in rank reversals and provide a way to use the pairwise comparisons for experts to produce better weightings. The process here, follows the typical AHP process where, for m requirements, the participants each complete a $m \times m$ pairwise comparison matrix **A** such that each element $a_{ij} = 1/a_{ji}$ and $a_{ii} = 1$. Each column i is divided by the sum of the values in that column (apply the one-norm), producing a new matrix where the column sums are all equal to one. Calculating the average of each row in the new matrix produces the weightings for customer requirements.[33]

To improve the accuracy and robustness, Raharjo proposed a method to use an analytic network process, an advancement over AHP. This demonstrated slight improvements in accuracy, but at the penalty of additional information required [109].

Franceschini argues that the danger of noncoherence between judgments and the difficulty in adjusting them increases with the use of AHP. He further argues that AHP makes it easy to process a "heap of data of somewhat scant significance" and produce results that assume a great deal more than was present in the original information. [53] There are two even larger detractors for using AHP in this fashion for this particular problem. The first is that the prioritizations and weightings are only useful for the top-down information flow and the method here has already identified a need to produce bottom-up information. The second is that, even if AHP were modified and used to generate the relationship matrix, the time associated with the process would likely result in the modeling exercise taking longer than a similar traditional modeling process, negating the need to use AHP at all.

2.3.4 Kano's Model of Customer Satisfaction

Another problem with the weightings of the customer requirements is that they do not have equal impact on customer satisfaction. While different weightings account for priorities, they do not accurately capture the difference between requirements with a particular threshold, desirements, and "as good as possible" goals.

Matzler and Hinterhuber first suggested the use of Kano's model of customer satisfaction in QFD to help account for these differences [87]. Kano's model was first introduced in 1984 and classified customer preferences into five categories: attractive, one-dimensional, must-be, indifferent, and reverse [73]. Three of these are depicted in Figure 11. Attractive requirements are those where exceeding the requirement brings about delight, but not meeting it has no penalty. One-dimensional requirements are those where customer satisfaction is directly related to how well a requirement is met. Must-be requirements are threshold requirements where exceeding the threshold brings no further satisfaction, but failing to meet it brings dissatisfaction. Indifferent requirements are those which the customer does not care about and would lie along the horizontal axis in the figure. Reverse requirements are those where customer satisfaction is opposite of the direction of the requirement's improvement. The last type is usually explained with customers who dislike a product that is too high-tech or too fast.

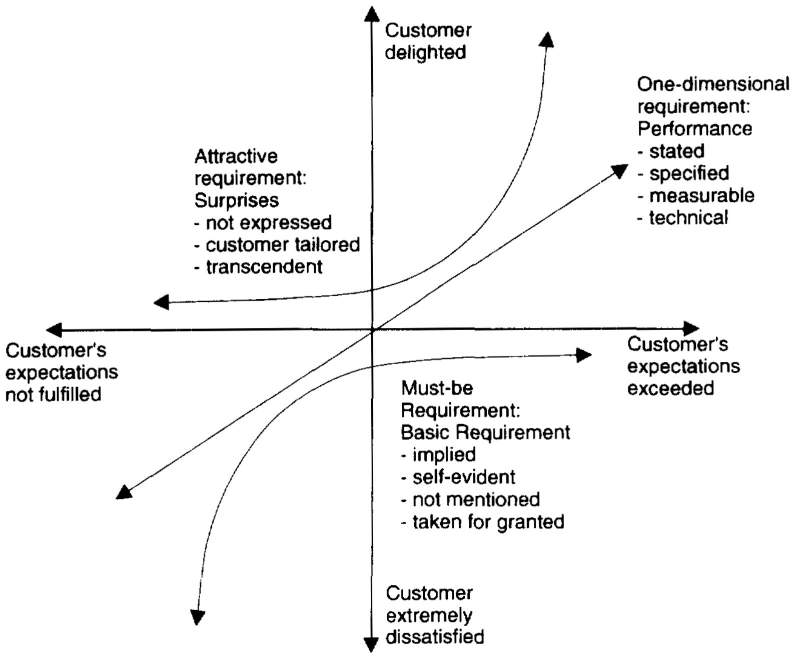

Customer delighted
One-dimensional requirement: Performance
- stated
- specified
- measurable
- technical
Attractive requirement: Surprises
- not expressed
- customer tailored
- transcendent
Customer's expectations not fulfilled
Customer's expectations exceeded
Must-be Requirement: Basic Requirement
- implied
- self-evident
- not mentioned
- taken for granted
Customer extremely dissatisfied

Figure 11: Kano's model of customer satisfaction [87]

Differentiating the customer requirements into these five types allows for the design process to better focus attention on different areas. The drivers move from a single weighting on customer requirements to a pair of requirements, one indicating the impact on customer satisfaction, and one for the impact on customer dissatisfaction. When included with the competitive assessment, this can identify which requirements need attention to exceed a competitor's offering and which would have no effect (and thus need no improvement).

With time, requirements shift from one category to another. What is now considered an attractive requirement may soon become a must-be requirement. This problem impacts aerospace vehicles as well, as military requirements and abilities change or passenger airlines shift flight types or passenger types. Raharjo identifies a method to include the current category and probability of future category to better account for these inevitable changes [109].

Others have identified the power of this method toward better identifying the most essential engineering and manufacturing characteristics. It adds more information to the results than is required to create it. However, for the case of producing a bottom-up model, this approach suffers the same problems as AHP in that it requires additional time and information that is not helpful in model-creation.

2.3.5 ROSETTA

The QFD is inherently subjective since it depends on expert-based assessment of the relative importance values and relationships between levels of attributes to create it. Several efforts have been made to include objective quantitative data into a QFD. Sohn used statistical information about local traffic accidents to create the relationship matrix. The reduction in types of accidents were the customer requirements and the conditions of accidents (road type, road width, traffic control devices) were in the place of engineering characteristics. He then used the QFD structure to prioritize changes in road and traffic design to reduce certain types of accidents.[119] This approach was one of the first efforts to use the QFD framework to support numerical analysis, but did not extend it beyond that problem.

Mavris and Griendling developed a generalized approach to use modeling and simulation to populate a QFD independently of Sohn's efforts and named it the Relational Oriented Systems Engineering and Technology Tradeoff Analysis (ROSETTA) environment [91]. The name stems from the idea that, just as the Rosetta Stone translated between several language, ROSETTA translates from modeling to QFD, using theoretical math as a common language.

Where Sohn used existing data and put it into a QFD format, ROSETTA uses a design of experiments to generate a large data set of points from a higher fidelity modeling and simulation environment with the primary purpose of incorporating it within ROSETTA. This data set is then regressed into a set of surrogate models using the Response Surface Methodology. Surrogate models provide nearly instantaneous calculation of the same quality of relationships present in the full model, but over a reduced range and reduced number of variables.

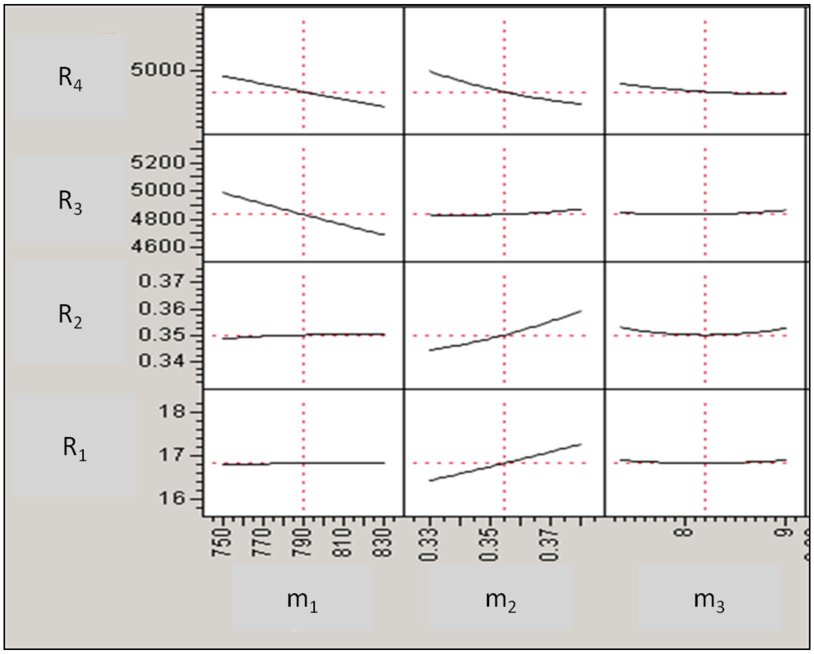


Figure 12: Example Prediction Profiler for ROSETTA [60]

When the partial derivatives of the surrogate models of each response with respect to each metric are plotted simultaneously, they produce a set of graphs similar to Figure 12, sometimes known as a prediction profiler. The partial derivatives are defined in terms of the values of the three metrics on the bottom as indicated by the vertical red dashed lines. In this case we see the four requirements R_i defined in terms of the intermediate metrics m_j. As the values of each of the metrics changes, the partial derivatives are recalculated at the new point. Translating this back into terms of a QFD involves demoting the continuous partial derivatives into numerical scores, even if they remain dynamic.

The correlation matrix, or roof, of the QFD is formed using the same set of surrogate models, but in a different fashion. The high speed of the surrogate models allows a very large number of data points to be generated quickly. A large Monte Carlo simulation is performed over the ranges of the design variables for which they are valid. For the lowest level of design variables which serve as the inputs of both the original model and the surrogate model, there is no correlation since those values were set by a design of experiments for independent variables. At higher levels, the correlations show up as a result of the physics of the model. These can be plotted in a multi-variate plot and Pearson's correlation coefficient can be calculated for each relationship. The upper-diagonal portion of this plot, as shown in Figure 13 within the blue border, is analogous to the correlation matrix present in the typical QFD.

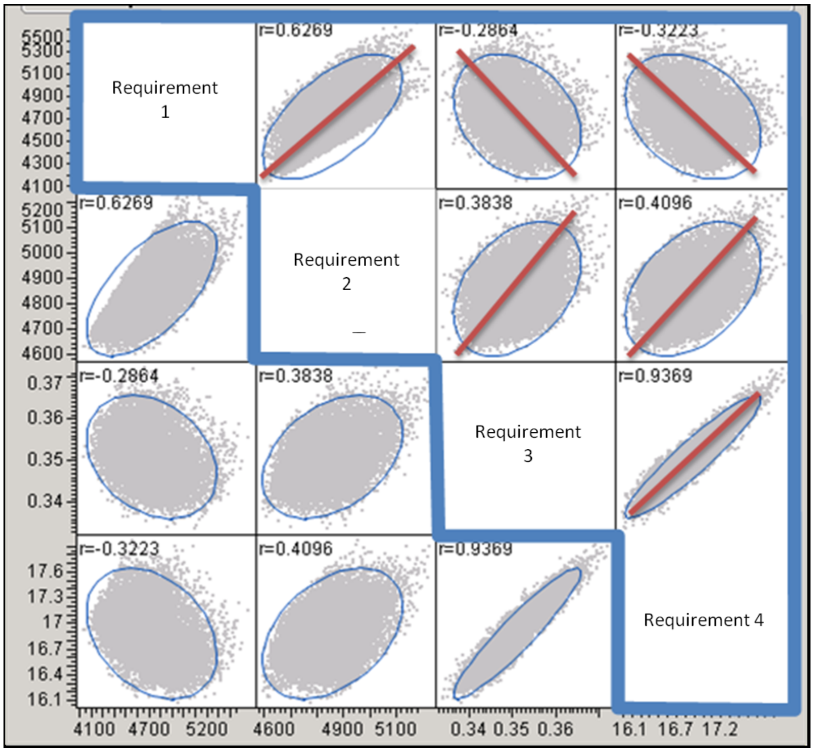


Figure 13: Example Multivariate Plot for ROSETTA, reproduced from [60]

While the current state of the art of ROSETTA cannot capture interdependencies between the design variables, it is possible to do so once constraints are applied. A correlation in the roof of a QFD does not mean that two things are impossible to change without changing the other, but rather that the two tend to both change at the same time. Often this is due to constraints at a higher level, even ones that are so obvious they aren't always included in the modeling environment. For example, the size of wings on an aircraft and the payload of the aircraft are highly correlated, not because they have to be. One could build an aircraft with a giant payload and tiny wings or one with almost no payload and huge wings. The first would not be able to fly and be better described as a cargo container while the second would just be a very inefficient way to move cargo. The reason wing size and payload are generally correlated is because of external constraints or optimization that tries to find the most efficient design. If such optimization and constraints are applied rather than a pure bottom-up calculation of the surrogate models, the design variables will be correlated as well.

ROSETTA also can use these surrogate models as closed form-equations to provide calculations of the priorities of the engineering characteristics. The overall quality of a solution Q is calculated as a weighted sum of customer requirements R_i with corresponding weightings w_i. These weightings are equivalent to the weightings on customer requirements typically seen in a QFD.

$$Q = \sum_{i=1}^{n} w_i R_i \tag{1}$$

Taking the partial derivative of Q with respect to the intermediate metrics m_l generates Equation (2).

$$\frac{\partial Q}{\partial m_l} = \sum_{i=1}^{n} w_i \frac{\partial R_i}{\partial m_l} \tag{2}$$

This means that $\frac{\partial Q}{\partial m_l}$ is equivalent to the priorities of the intermediate metrics and if all values of $\frac{\partial R_i}{\partial m_l}$ were limited to constant values on a predetermined scale, it is the same as the relationship matrix of a QFD. ROSETTA also includes relationships between metrics and captures this as $\frac{\partial m_k}{\partial m_l}$. To cover all combinations, a second summation is needed to produce Equation (3).

$$\frac{\partial Q}{\partial m_l} = \sum_{i=1}^{n} \sum_{k=1}^{p} w_i \frac{\partial R_i}{\partial m_k} \frac{\partial m_k}{\partial m_l} \tag{3}$$

Each metric is a function of the design variables x, allowing further differentiation against design variables and then against subsystem design variables until the lowest level of detail is reached. This differential equation produces QFD-like values and representations from existing models and closed-form equations.[91] A visual representation of these calculations are shown in Figure 14.

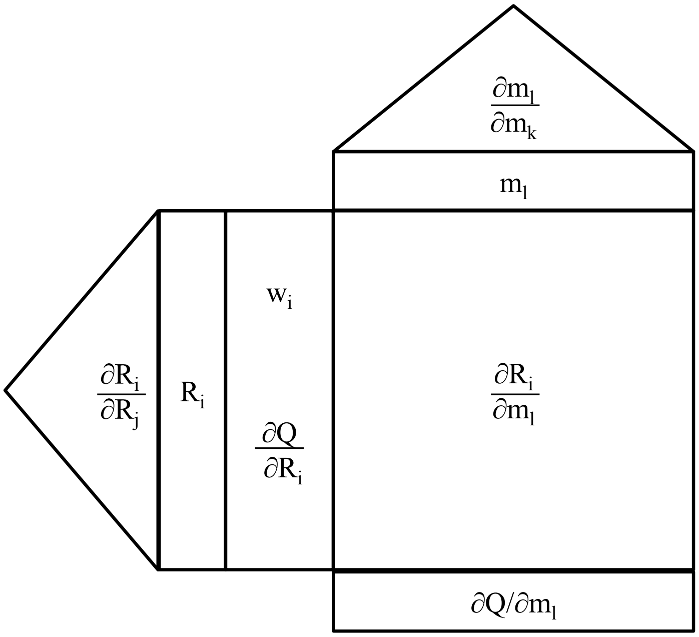


Figure 14: QFD Rooms Corresponding to Analytical ROSETTA Terms [60]

This depiction also includes a “greenhouse” on the left to capture the correlations between the requirements. The weightings on the requirements are listed as both w_i and the generalized differential equivalent $\frac{\partial Q}{\partial R_i}$.

2.3.5.1 Link Between ROSETTA and this Research

The result of both the graphical depiction and mathematical depiction of modeling within a QFD is a highly accurate representation of the problem in a structure that may be more familiar to those who have used QFD before. This is a way to translate from the language of the modeling and simulation community to a design process and management community. This serves as one half of the bridge connecting the two. The other half is what SP2 has started and this research is attempting to complete.

ROSETTA helps designers prioritize and understand the relationships and design variables for a particular problem better. The inverse relationship is a way for designers and engineers to influence the modeling and simulation of the problem. Designers understand the trends and can identify the strength of the relationships as they see them. Those areas with the strongest relationships or that have the greatest impact on the flow up or the flow down should be the highest priority for modeling. This may mean either that they require the highest fidelity or just that they should be modeled first.

An expert-based model can serve as the skeleton to develop these models incrementally, especially for problems where requirements and intermediate metrics each need to be calculated with individual models. As each model is created, it can be dropped in to the framework, replacing the expert-based version. When both the expert-based model and the traditional model are present for a particular set of relationships, this allows a new step that has not been done before. The two can be compared on the fly and both can be updated. The designers can use the traditional modeling to update their own understanding of the problem and perhaps adjust their own estimates for other relationships. The modelers can use the expert-based model as a form of validation. If the two do not agree, this presents the opportunity to identify the difference in understanding and the way the physics was modeled. It may be an error or it may be that the particular design problem and design space behaves contrary to common understanding.

If the expert-based model is used, it also allows for the ROSETTA equations to be used to ensure that the full functionality of the QFD process is maintained. This means that any expert-based modeling can be performed in place of the QFD step for some problems without a loss of information.

2.3.6 Linking QFD With Other Tools

The entire QFD offers insight into product development problems. However, the ability to quickly compute the importance of lower level product characteristics from higher level priorities has allowed others to incorporate QFD into interactive tools. These tools frequently allow users to change the priorities of the customer requirements and instantly view the changing product characteristics.

Biltgen incorporated QFD into several different types of analysis. His first use was using it to support concept selection for a tunable missile defense target using multiattribute decision making. Customers were able to adjust the desired characteristics of the program and QFD would translate these to characteristics of the individual missiles, which were then scored and ranked according to how well they met those engineering characteristics[14]. As part of his doctoral work, he later used a five-phase QFD to translate from very high-level customer requirements such as “Provide National Security” through Joint Services capabilities, operational activities, system functions, and physical systems to identify a database of architectures to analyze for a long range strike system. In that same research effort, he used a separate QFD to flow down the priorities of campaign goals down to generate a prioritized list of individual targets as part of a battle simulator. [15] Whereas the first approach was merely a selection tool of a single vehicle by means of prioritization, the second linked into an agent-based modeling environment. As the priorities changed, different vehicles or vehicle designs would be included in the modeling environment.

SP2 and SOAR, two of the most mature and developed examples, were discussed earlier in Section 2.2.5. For these methods, QFD is essential for translating a customer's goals into prioritizing particular technology programs that drive that goal. It goes beyond the simple ranked list of targets that Biltgen used to identify particular technologies that are must-haves and to include dependencies. [75, 107]

These advances are the closest to what is desired in this research since they use QFD as part of near-quantitative analysis. Unfortunately, the methods are intended not to produce a stand-alone model, but focus on providing priorities at a lower level.

2.4 Research Goals

The next chapter discusses the process of developing the method for producing expert-based models. In order to better focus that effort, the specific research goals for that method should be identified. Section 1.4 defines the type of model that was desired as a result of this research and creation of such a model is the primary goal of this research. The background information supports the choice to use QFD as a baseline method for gathering and using expert-based information. QFD is already frequently performed as part of engineering product development and is frequently limited to requirements development. It would be more appealing if that same QFD information could be reused rather than duplicated.

However, it is unlikely that QFD as it is currently will be a drop-in solution to producing an expert-based model. The advances identified are of limited use in changing the direction of flow and more specific problems associated with that change will be identified in the next chapter. As those problems are identified, further improvements or suggestions from literature can be presented. Any changes in the type of information needed to solve those problems should still allow the similar type of QFD flow-down as is currently used. This addition to the already existing model requirements produces the following research objective.

> **Research Objective:** Identify, implement, and test a method to rapidly collect expert-sourced information about a system and translate it into a predictive,

parametric model useful for early engineering design.

It is expected that any improvements that are necessary to meet a minimum amount of accuracy needed to be useful may also improve the accuracy of the QFD flow-down of priorities as well and aligning with ROSETTA for a common feedback loop between design and modeling throughout the design process.

CHAPTER III

DEVELOPMENT OF THE METHOD

To accomplish the research objective, a structured method is necessary to ensure that the process tested is the same one that a person using this research would use. In order to provide such a method, a generic modeling framework is developed by investigating and comparing existing methods. This framework defines broadly what the steps are to create a model: *Define the Problem*, *Gather Information*, *Create the Model*, *Test the Model*, and *Use the Model*. Each of these steps is established with specific tools and approaches based on available literature. Several preliminary experiments are performed to properly characterize the needs associated with a step and evaluate possible solutions. The alternatives that were investigated and those that require further testing are summarized in a methodological morphological matrix.

3.1 A Framework for Model-Building

The field of model development is not a new one. Various processes for model development have been proposed for more than half a century [94]. As it became a teachable skill, standardized processes for creating new models were developed. A new modeling approach should leverage past research and experience to minimize any reinvention. Using a standard process as a base will also help aid in understanding and explaining the differences to others. That said, each author tends to have a process that differs slightly from others in literature. The processes presented here are of particular interest because of how well they apply to a wide range of problems.

Many of the processes developed were based on the classical scientific method [55]. They include observation, formation of a hypothesis and predictions in the form of a model, and experiments to validate the predictions. Some processes appear to be more complex than others. They range from five simply defined steps to ten or more detailed steps. For clarity, they are each presented in order from the simplest processes to the most complex and specific. The explanations of each process is based on the authors' own explanations. Some processes exclude steps that may be obviously necessary, but the omissions remain in the descriptions.

3.1.1 Cross and Moscardini

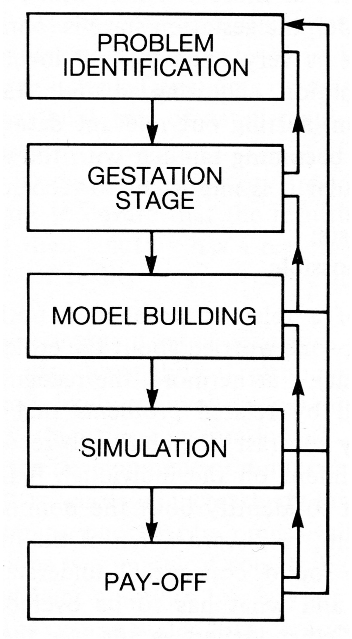


Figure 15: Cross and Moscardini's Stages for Problem Solving with Mathematical Models[38]

Cross and Moscardini present perhaps the cleanest process for creating and using a model, originally based on Peel's 1979 paper [105], and shown in Figure 15. It begins, as most of the processes do, with identifying the problem. They do not go into detail about this step, but make it clear that models should be built to solve a particular problem of interest. The next stage is a unique one known as the gestation stage. It includes gathering of background information and understanding the system under construction. It also includes the necessity for an individual to sift through the available knowledge on the way to a moment when they "see the light or experience Eureka" and identify a solution [38]. The ability to reach this depends on the individual and requires a certain amount of intuition. Such intuition is only gained through experience. This stage ends with a good conceptual understanding of the problem and a framework in mind for the model.

The model building stage includes both the development of equations and validation of those equations. Some other modeling processes fit within this one stage. The model is built by first creating a formulation of the mathematical model in equation form and then identifying a solution method. This solution method could be analytical, but Cross and Moscardini suggest that most realistic systems necessitate numerical evaluation. The model is validated against the available data and the physical constraints of the system.The final step of model building is refining the model to tune it to better fit the validation data and to improve its applicability and scaling. The simulation stage focuses on carefully selecting a set of simulation runs and performing them. Modelers should be skeptical of their results and continuously check that the results make sense. The final stage in the process is the pay-off stage. Where the simulation stages uses the model to generate results, this pay-off stage uses the results to make recommendations, defend the model to the decision, and appraising the model as whole [38].

The feedback loops within a modeling process are essential to allow for the modeler to reform previous decisions based on additional information and further understanding gained later in the process. A modeler may come to better understand the original identified problem during the gestation stage or any of the later stages. It may also be necessary to further specify the problem as the model grows too large to be feasible or additional input is needed. During the simulation and validation, errors may be found in the model that need to be corrected by rebuilding the model. If the results do not provide the answers or decision-support desired in the pay-off stage, this feedback must flow up to generate the answers that meet those requirements. Returning to a previous step builds upon previous decisions and refines them rather than starting over from scratch.

The simplicity of this process allows a modeler to quickly understand the general process. It also aids the modeler when presenting the development of the model to others. The simplicity is less helpful when a modeler sits down to create the necessary model. These processes are not intended to be modeling cookbooks, nor could they be, and remain applicable to a wide range of subjects. This model is presented here because it captures the full process from recognizing a need to satisfying that need and does so with a simple diagram. Its suggestion of a gestation phase is an important one that most other processes take for granted.

3.1.2 The Open University Seven-Block Process

The Open University modeling process is based on a diagram of seven blocks, as shown in Figure 16. This representation includes an additional differentiation of the real and mathematical realms based on where Galbraith and Haines see them [57]. Other depictions of this particular process use the same structure with different names and can include additional whitespace in each box for note-taking as part of an academic course[38]. This process starts with identifying a real world problem, as the previous process did. The Open University process puts most of the model development into the the model formulation step, but separates the mathematical manipulation into a separate step. This has a similar benefit as the previous method's choice to separate the gestation stage and the model building in that it allows one to focus on the structure of the model first and the best way to simplify that structure afterwards. The solving mathematics step also includes generating numerical results.

Interpreting outcomes is the step associated with understanding the numerical results generated in the previous step. This starts with translating raw data into more readable information, such as plots or summary tables. The interpretation also includes figuring out what the results mean to form conclusions. The evaluating solution questions the usefulness of the results and the model itself based on the original real world problem. This step

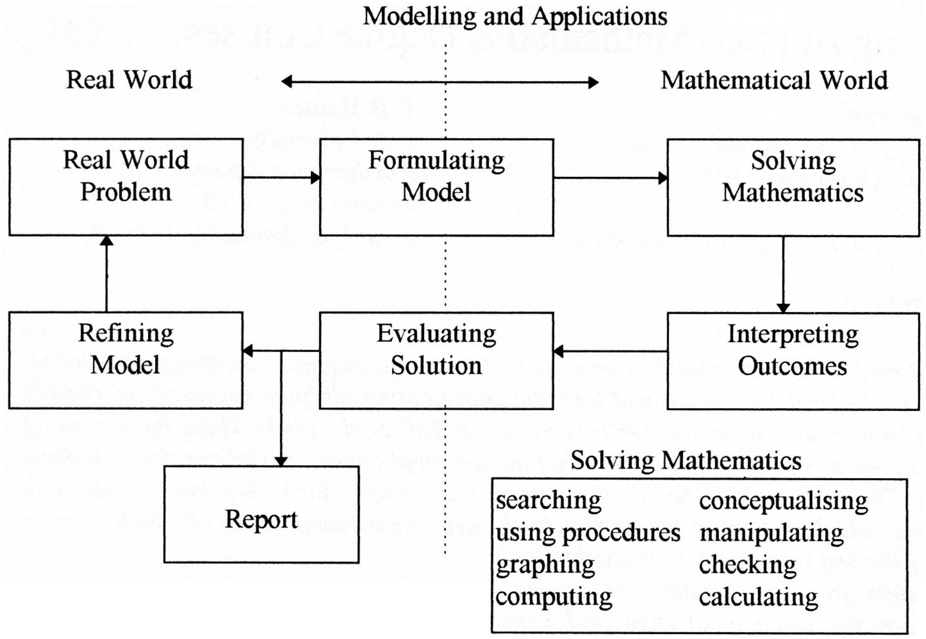


Figure 16: The Open University Seven-Block Process for Modeling[57]

determines whether the conclusions reached in the previous step actually answer or solve the problem. If they do not, an analysis of where the point of failure or misunderstanding is performed and changes are made as part of the refining model step. This analysis should give insight into the real world problem or may reveal a shortcoming in it or the way it was stated. The process loops to the beginning for the modeler to iterate until the real world problem is either solved or has exhausted the time and resources available. Outside of this loop, the reporting step is the way in which results and conclusions are presented to the outside. It is essential to prevent the problem of modeling for modeling's sake. This is similar to the pay-off stage in the previous process.

The Open University modeling process shares the benefit of simplicity with the previous process. It also covers the full scope of a modeling exercise from problem definition to final reporting. That said, it is more focused on the modeling aspects than the problem-solving aspects. The differentiation of the real world and the mathematical world of the problem is an important one. It is important to remember that the model is distinctly different from the system it is modeling. It must represent the behavior of the system, but not necessarily its inner workings.

3.1.3 Bellomo and Preziosi

Bellomo and Preziosi define their modeling and model analysis process using the diagram shown in Figure 17. It takes a formal mathematical approach to model definition where the previous processes refer to the model abstractly. The independent variables are pre-defined as time and a position vector in a fixed system of orthogonal axes. The state variable is a vector valued variable that is usually a function of the independent variables. Parameters are physical quantities that describe the physical system being modeled. Note that the parameters are different than the independent variables and that the state variable is not directly a function of them.

Their process is unique among those presented here in that it does not explicitly include any sort of problem identification or definition. It starts with a general observation and numerical measurements of the system of interest. The general observation is used to define the parameters and later formulate the model. The measurements are used to determine the values of the parameters and to set the ranges of the independent variables for which the model must be valid. The choice of the state variable is a result of the motivating problem, but the specific measure may be determined as part of the observation.

The formulation of the model uses the definitions of the variables and parameters as a starting point for the modeler to determine the form and structure of the model equation. The bulk of Bellomo and Preziosi's book is dedicated to different ways to formulate models. Model analysis refers to solving analytically or numerically in order to get sufficient information to perform validation to measurements of the real or representative system. The formal mathematical nature of the models allow for different specific measures to be presented based on the form of the equation and classification of the model. A validated model is then available for general use. [12]

Bellomo and Preziosi are able to give formal mathematical definitions of the inputs and

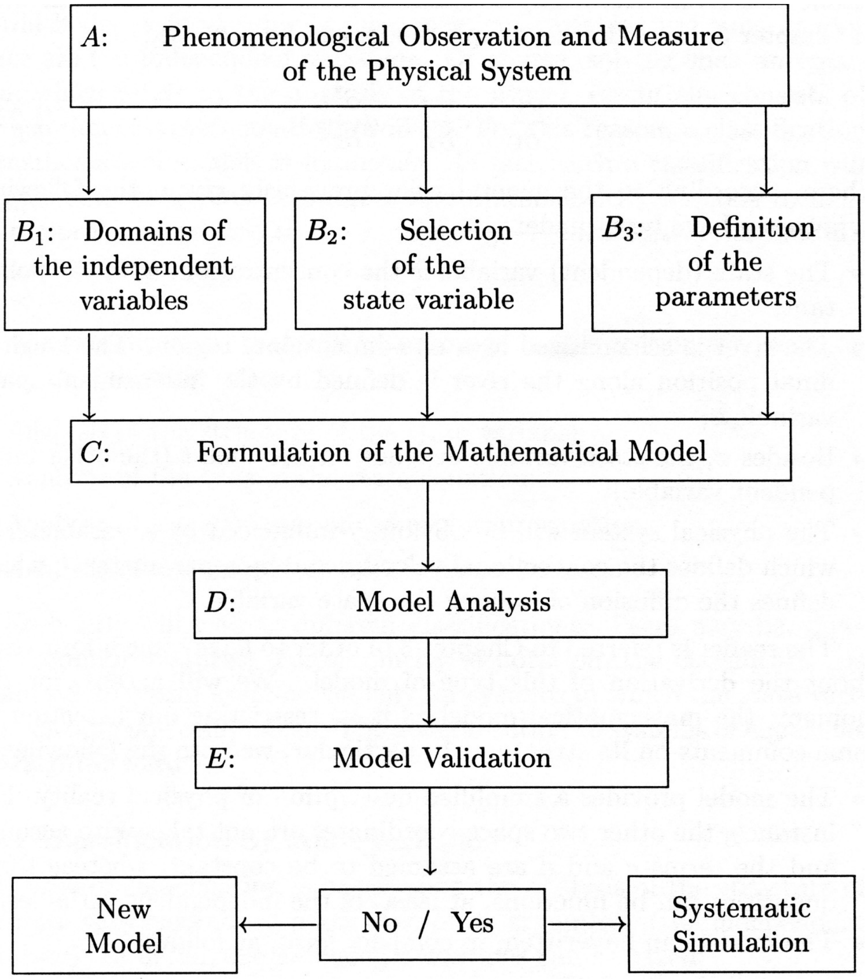

A: Phenomenological Observation and Measure of the Physical System
B1: Domains of the independent variables
B2: Selection of the state variable
B3: Definition of the parameters
C: Formulation of the Mathematical Model
D: Model Analysis
E: Model Validation
New Model
No / Yes
Systematic Simulation

Figure 17: Bellomo and Preziosi Modeling Process[12]

outputs of the model without limiting the process to a particular field. This is extremely helpful in understanding how to structure a model under development. This may also be restrictive if the system is not readily depicted primarily in time and space and defined by a state equation. Some systems, such as those that rely on information-based behavior over physical behavior, may not be modeled in this construct at all. The balance is between a process that works on every conceivable problem and a process which provides useful and structured guidance.

The omission of a problem definition step or a reporting or recommendation step makes it quite clear that the focus here is on the model. Where other processes could generally be considered problem solving methods, this one is intended to build a model whose need has already been determined. Problem definition still occurs when building a model here. It just happens outside the limited scope of creating a model. This is a useful lesson for the present research.

3.1.4 Fitzharris

The modeling process that Fitzharris presents, as shown in Figure 18, is perhaps the most applicable to the modeling approach planned here. Like several others, it starts with problem definition. However, unlike others, the problem definition step includes planning the necessary hardware, software and personnel resources. The second stage is defining the boundary of the system and the model. Identifying which parts of a system are being included helps to reduce the complexity and allows for reasonable assumptions about the interactions with portions outside of the boundary. This stage also determines which techniques for simplifying the system will be used, such as omission of unnecessary detail, aggregation, and substitution of complex portions with simpler portions.

Stage 3 first defines the specific data required, the available sources for it, and how the data will be represented or translated to be useful. Once the data needs are determined, the data is gathered according to the developed plan. The prepared data is used to formulate the model according to the problem at hand. The standard types of verification and validation follow the model formulation. As usual, if there are any problems at this point, the model is modified and refined to better represent the available information. The output of one or more simulations of the model is analyzed to better understand the behavior of the system. Fitzharris specifically mentions determining parameters of the system from this analysis. The parameters here are better described as performance metrics rather than the type of parameters used by Bellomo and Preziosi. The process ends with reporting the results and making recommendations based on those results. [51]

Fitzharris's decision to include planning details and determine sources of information as part of the process captures the pragmatic side of model creation that is often overlooked.

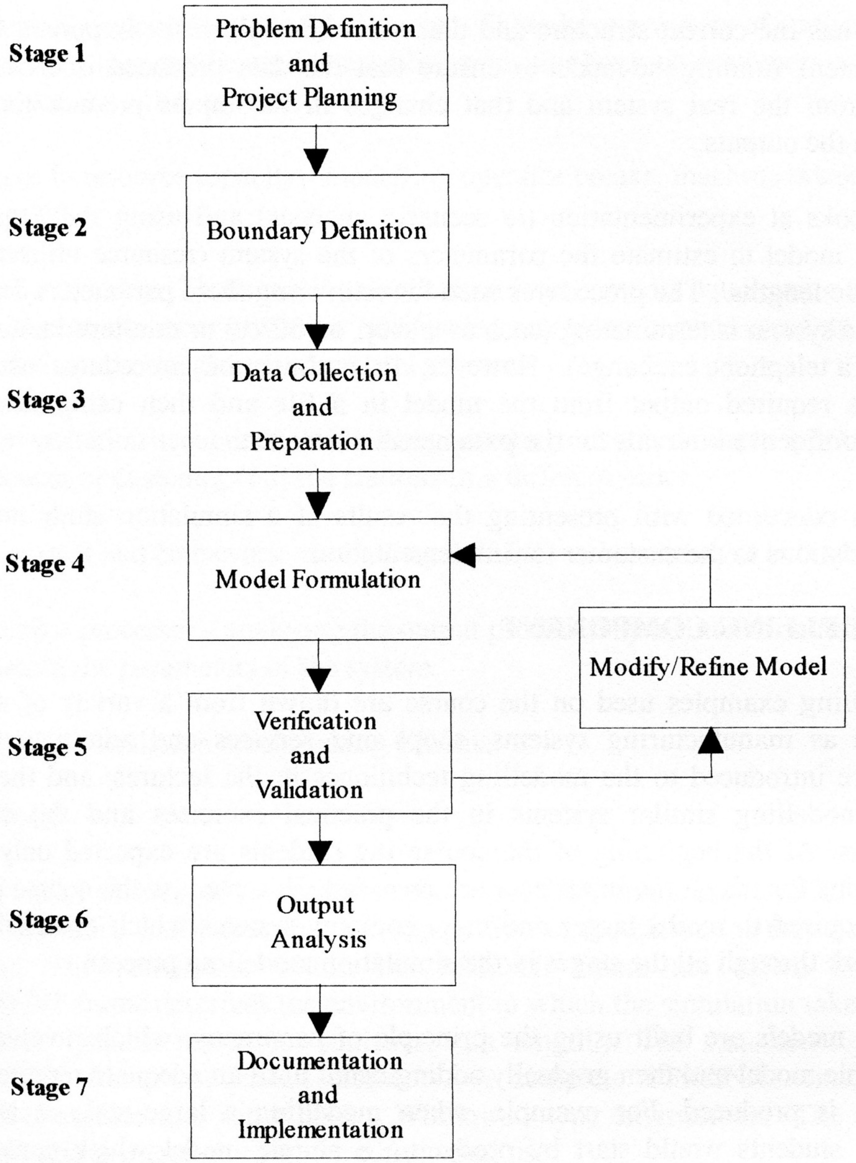

Stage 1
Problem Definition and Project Planning
Stage 2
Boundary Definition
Stage 3
Data Collection and Preparation
Stage 4
Model Formulation
Modify/Refine Model
Stage 5
Verification and Validation
Stage 6
Output Analysis
Stage 7
Documentation and Implementation

Figure 18: Fitzharris Simulation Modeling Process [51]

The specific mentions of the need to identify what parts of the system can be simplified in modeling and the awareness of where the modeling data will come from relate directly to this work as well. This process balances the level of detail nearly perfectly for the purposes here. The diagram and descriptions are simple enough to explain quickly and easily to someone who is not familiar with them. At the same time, there is a great deal of thought that has gone into the individual steps and what must be performed as part of them for the modeling exercise to be successful.

Chung's modeling process is very similar to Fitzharris's with a few small differences. Chung splits the problem definition and project planning into separate steps as well as splitting up verification and validation into two steps [34]. Because of the similarities, it is not discussed separately, but many of the benefits and detriments apply equally.

3.1.5 Blilie

Blilie's modeling process places much more focus on the decomposition of systems into subsystems than other processes presented here. His summary is shown in Figure 19. Though this is represented as a serial process, Blilie makes it clear that this is a spiral method rather than a waterfall method [16]. The first step is the same problem identification that most other processes start with. The next two steps focus on the ability to identify and isolate the system of interest from the world around it. Within a complicated set of entities, a single system must demonstrate coherence and internal structure to identify it as distinct enough to model. This is not always a simple matter and for some sets of entities, it may be impossible to model a subset of the entities. Once a system is identified, suitable boundaries and the corresponding boundary conditions must be found. These boundaries will apply to the model as well. The boundary conditions will correspond to the inputs and outputs of the model.

The next few steps decompose the system into subsystems. The subsystems are identified in a similar manner as the system was. The connections into and out-of a subsystem

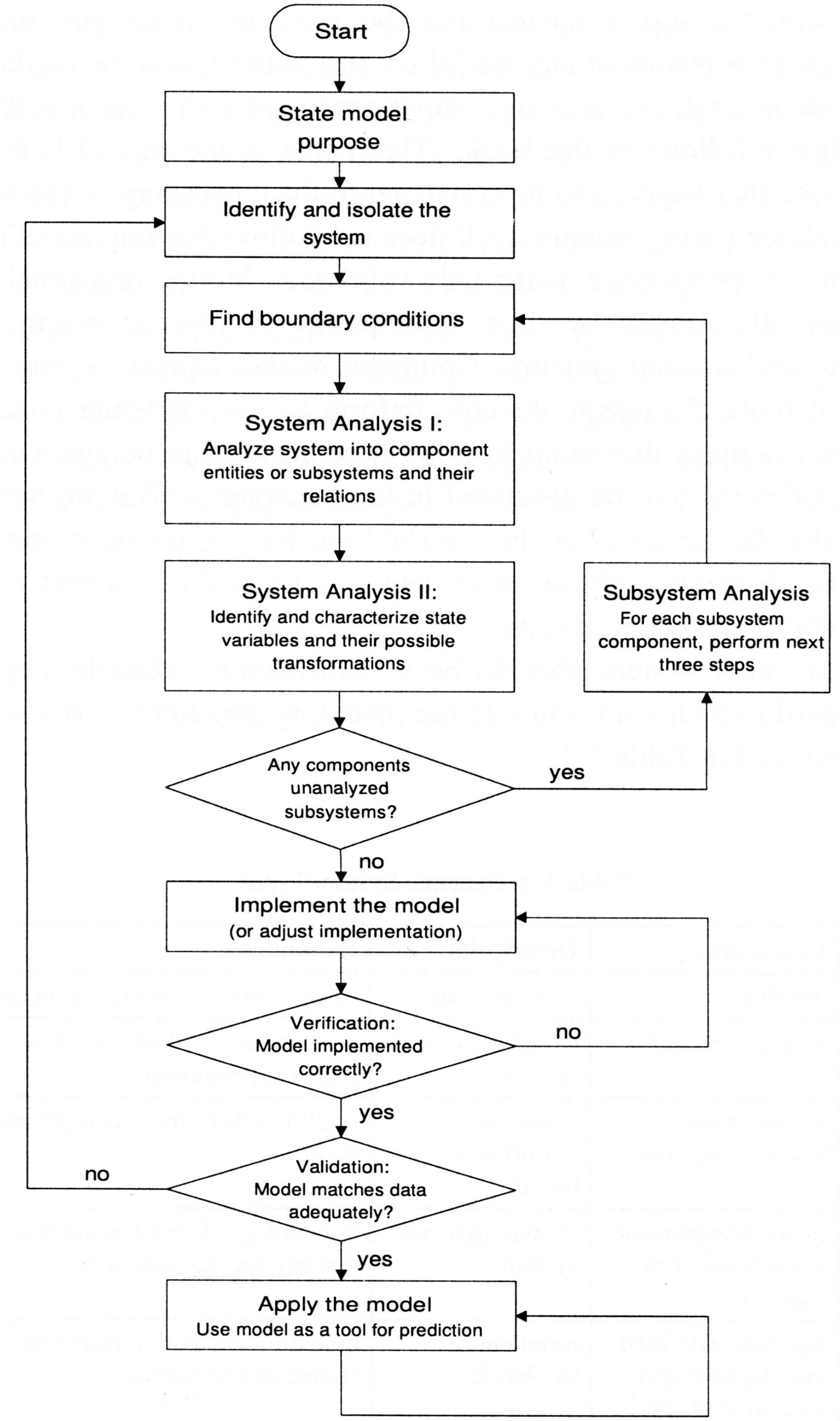

Start
State model purpose
Identify and isolate the system
Find boundary conditions
System Analysis I: Analyze system into component entities or subsystems and their relations
System Analysis II: Identify and characterize state variables and their possible transformations
Subsystem Analysis For each subsystem component, perform next three steps
Any components unanalyzed subsystems?
yes
no
Implement the model (or adjust implementation)
Verification: Model implemented correctly?
no
yes
Validation: Model matches data adequately?
no
yes
Apply the model Use model as a tool for prediction

Figure 19: Blilie Modeling Process [16]

are known as interfaces rather than boundary conditions. Similar subsystems can be represented with a single model to reduce the required modeling effort. Step five defines the numerical quantities that describe the system (or subsystem). Quantities which are strictly internal and do not affect the behavior of the system or other external-facing quantities can, and should, be neglected. A good identification of these quantities will help ensure that the model maintains matter and energy conservation. The behaviors identified are the core of the models for the individual components. These steps are repeated for any identified subsystems or subsubsystems until the smallest useful grouping is identified for analysis.

Once the problem has been decomposed as far as is desired, the models for these subsystems are combined into a model of the system as a whole. Blilie goes into less detail about the reconstruction process than the analysis necessary to fully decompose the system. Hopefully the choice of subsystems and interfaces makes the implementation and integration a relatively simple step. The resulting integrated model is verified and validated similar to the methods suggested in other processes. If either step fails, the modeler returns to the previous steps to correct the errors. The final step is to use the model for prediction of system behavior to answer questions of interest.

This method is included here because of its approach to the partitioning of systems and subsystems. If limited to a low fidelity model, decomposition can allow multiple low fidelity models to capture more detail than would be possible with a single model. On the other hand, the difficulty associated with decomposing a system several levels increases the time and information required to produce a valid model. Certainly aspects of this decomposition are helpful, but the process as described here is generally too complicated to be used in a rapid model development exercise.

3.1.6 Other Processes

There are too many other modeling processes by different authors to include them all here. Certain types of models have been specifically omitted because of their limited applicability. Some of them are developed primarily for education, such as Ikeda's process of five state and four stages between them [72]. These frequently include a classroom version of the model that is different than the "real" model or can include steps intended to make an instructor's evaluation of student performance easier. This type is useful for teaching but the additional steps mean that more effort must be performed to create and use the model outside of a classroom setting.

There are other process which are less of a process and more of a conceptual diagram of the modeling process than a set of steps or goals. Some of these, such as Sargent's three-box diagram, include bidirectional arrows from every step to every other step with no clear starting or ending point [113]. These types of representations can be helpful when discussing model theory, but are significantly less useful when structuring a modeling exercise.

Still others are primarily intended for specific fields or types of models such as system dynamics models or manufacturing processes [63]. These processes may pre-define specific types of inputs and outputs necessary to create the model or pre-define the form of the resulting model. Within their respective fields, such assumptions are useful to help focus the modeling exercise to what is necessary and better standardize the resulting models. While these types of processes are very useful to their particular fields, the additional assumptions or structure included make them inappropriate on problems outside of those fields.

3.1.7 Simplified Modeling Process

The quantity and variety of modeling processes present in literature makes it difficult to identify a single process to use as a framework for developing a new model. There is no wide-spread consensus in the modeling or systems engineering communities for a preferred process. For the work here, a process is desired that is as simple as possible while still identifying the essential steps in creating a model. On one end of the spectrum is a process with a single step that implicitly includes all other steps. The other end of the spectrum is a detailed cookbook recipe for creating a specific model. To find the proper location between the two extremes, the process should be built based on those tasks that are explicitly necessary.

The purpose of a modeling process is to produce a useful model. Thus, creating the model is the most essential step in the modeling process by definition. It is difficult to decompose this step further without specifying the type of system or the form of the model. Blilie defines the type of system by assuming the existence of subsystems[16]. Bellomo and Preziosi define the form of the model by specifying the need for certain types of variables and parameters[12]. Without making similar assumptions, the simplified process centers around "Create the Model".

Some quantity of information is necessary to create the model. In some rare cases, the individual creating the model may already possess sufficient knowledge to create the model from memory. In all other cases, the information must be collected from some other source such as other experts, theory, test data, or other trusted models. A model can only be as good as the information used to create it and the way information is gathered can be just as important as the information itself. It is important to explicitly include the information gathering step in the process, as Fitzharris and Bellomo and Preziosi do in their respective processes[51]. The details of how a modeler performs the "Gather Information" step depends on the modeling method used, the model required, and the information available.

The sections in systems engineering handbooks dealing with models frequently discuss the need to verify and validate models. A model is verified to confirm that it behaves as the modeler intended for it to and it is validated to confirm that it represents the true behavior of what it is modeling. In some communities, a model must also be further accredited by someone with the authority to give permission for it to be used for certain types of decisions. Modelers often verify a model by running sample cases, observing the behavior and comparing it with the modeler's expectations. It may also include analyzing the code or equations by another modeler. The ideal way to validate a model is to compare its results with the performance of the actual and physical system it is modeling. When performance data is not available, other models or expert opinion may be used[42]. Verification and validation can be split, as Blilie does, but they are so often discussed together, that there is no need to separate them. The simpler term "Test the Model" is used here since both verification and validation are testing whether the model is correct but against different standards.

Systems engineering and problem-solving processes start with identifying and understanding the problem. In the case of modeling, this step determines what model is needed in terms of inputs, outputs, accuracy, run-time and so on. It also defines the system being modeled and the conditions of how the system will be used. These determinations drive what information is required, the form of the model, and how it must be validated. It is possible to split up this step into multiple parts, such as problem identification and system definition, but the tight interconnections of the most modeling processes and problem solving approaches support keeping it unified in a high-level process. The wide-spread inclusion of a step to "Define the Problem" as well as the need to explicitly identify when planning and decisions about the modeling effort should be performed motivate it's inclusion.

The goal is to produce a useful model rather than just producing a model for its own sake. The distinction between the two can be captured in the process by including a step where the developers "Use the Model". All of the processes presented include this step in some fashion. As the model is employed for analysis and decision making, shortcomings or further requirements can be identified, leading to iteration to previous steps of the process. Including this step in the process also recognizes the need for the model creators to be

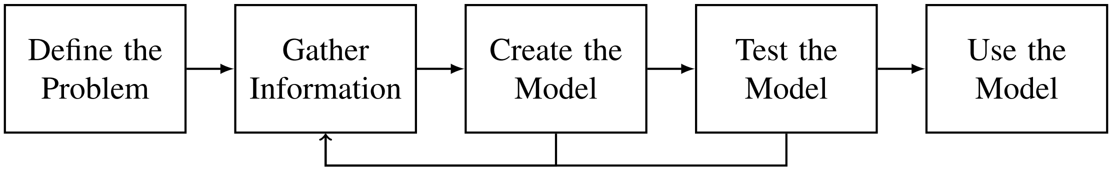


Figure 20: Simplified Modeling Process

involved in how it is used to ensure that it is used correctly and to explain or correct any variations from the expected results.

The process including all five of the discussed steps is shown in Figure 20. The entire process is iterative since any step can identify changes that need to be made in a previous step. The feedback loops are depicted for the three middle steps where iteration occurs most frequently. It is common to realize the need for additional information as the model is created. In some cases, it is preferable to gather all possibly relevant and useful information before beginning model creation. In other cases, it is more efficient to wait until the needs for additional information are narrowed to the minimum necessary. In other cases, the cost of creating a model is much smaller than the cost of gathering data such that a new model is created each time a new piece of information is available. Testing the model is often what identifies where additional data is needed, so this too can directly drive the need for more data to correct areas with unacceptable inaccuracy.

Several steps that were praised in individual processes have been omitted. Cross and Moscardini's gestation stage identifies the need to think about the problem, but readily identifies that it is really the chance for a modeler to use intuition[38]. The difficulty in defining how to perform this step as part of a method makes it difficult to include here. Obviously thinking is still required, but here it is necessary throughout the process. Fitzharris's boundary definition can help clearly identify the extent of the model. Here, that step is part of the system definition included in the problem definition step because boundary definition is an essential part of any system definition[51]. Blilie's choice to include steps related to subsystems is useful, but is only valuable when modeling a system that has subsystems that need to be modeled[16]. For other problems, the additional steps are unnecessary and can confuse the process.

The simplified modeling process is a high-level view of what should be done, but not how it should be done. The proposed modeling method defines how a useful model should be produced using this process as a framework to better explain and structure the steps required and the flow of information.

3.2 Define the Problem

Problem-solving processes start with a step that intends to capture the scope and requirements of the problem at hand. Engineering design processes, as a type of problem solving, start at the same step. For model development, defining the problem identifies the type of model that is needed, the scope and detail required, and the process or system that is being modeled. If one needs precise pressure distributions over the geometry of an aircraft, the first step is to identify the need for a computational fluid dynamics model. One would then need to define the geometry of the aircraft and flight conditions where it was operating as part of defining the system being modeled. It would also be necessary to define the fineness of the grid, the numeric solver method, which set of underlying equations to use (whether they be Navier-Stokes or a vortex-lattice formulation) and perhaps even the type of computer on which the modeling will be performed. Each of those may depend on the problem. The geometry and flight conditions in a particular model will change much more frequently than the other aspects. Each of those decisions may include a separate trade before the modeling is created.

For this research, the decision to use experts and the desire to produce a bottom-up model help to define many of the model-type concerns that will be true for any problem. When this method is used in practice, these decisions will not need to be redefined each time but may need to be verified that they are still applicable. For every problem, it will be necessary to define the system of interest for that particular problem.

3.2.1 Preliminary Experiment: Variation Between Design Points

A series of preliminary experiments have been performed to explore the needs of different steps of the process. They also provide initial analytical comparisons of potential options for the model development process. In planning these experiments, it was desired to separate the impact of using expert-based information from the impact of using a particular approach to interpret and utilize that information. This also separates the sources of error into faults from the data sampling and faults from the mathematical formulation. Therefore, there are two sources of information used in the preliminary experiments: expert-based information created by the author and data from a reliable model that has been reduced to simulate "perfect" expert-based data. If the experiments using the "perfect" data provide satisfactory results, then the mathematical formulation presented is valid. If experts are able to provide information that is very close to the "perfect" data, then the use of expert-based information is valid. If both premises are valid, the resulting methodology is valid.

The first experiment provides support and evidence for the need to define the problem. It also provides information about the sizes of differences due to differences in the problem. The test will be to compare two designs of similar vehicles of different sizes (and correspondingly different missions). The more similar the relationships between design variables and intermediate metrics and between the intermediate metrics and the system/-customer requirements are, the less concept definition is required. This experiment will only use the simulated "perfect" expert data.

The "perfect" expert data is created from a set of data points uniformly distributed over pre-determined ranges for the design variables and run through an accepted performance model. Values for the intermediate metrics and customer requirements are recorded. When filling out a QFD, experts are asked to give scores to answer "How strong is the relationship between X and Y?" or "How significant is changing X to the variability of Y?" where X is a lower level variable and Y is a higher level variable. A similar measure of this for a set of data is the correlation of X and Y, often measured by the Pearson product-moment correlation coefficient ρ_{xy}. Possible values for ρ_{xy} vary from -1 to 1.[1] These correlation values will serve as the ideally accurate scores experts for each relationship as a point of comparison.

The correlation is not a perfect measure of the existence of a relationship as it only measures the strength of a linear relationship and does not prove that the relationship is causal. The assumption of linearity is tested in Section 3.4.2 and found to be valid in most cases with only a slight deviation when it is not appropriate. The causality of the relationship may not be necessary for testing purposes, but is a shortcoming that can be accepted for the time being.

The systems studied were both passenger airliners. One was a 70-passenger class vehicle and the other was a 300-passenger class vehicle, both based on an intensive academic problem[89]. This problem traditionally involves optimizing a civil airliner over a set of design variables both with and without technologies to meet certain system requirement thresholds. The benefit of using a pre-existing problem is that the model has already been vetted with enough certainty that the results will be meaningful. This particular problem has enough variability in the design variables and requirements to be a useful assessment of a typical design range. The set of system requirements, intermediate metrics, and design variables used are listed in Table 2 while the input ranges for the design variables for both classes of vehicles are shown in Table 3.

These are the same variables and ranges that are used in the academic problem with only slight changes. In the academic problem the intermediate metrics are primarily used to measure the impact of technologies as technology *k*-factors rather than being calculated separately. [76] Since technologies will be held constant for this research, their inclusion made them ideal intermediate metrics. In the academic problem the *k*-factors were separate independent variables that multiplied these values to increase or decrease them from

[1] The Pearson product-moment correlation coefficient is discussed in greater detail in Section 3.5.3.1.

Table 2: List of System Requirements, Intermediate Metrics and Design Variables

System Requirements	Intermediate Metrics	Design Variables
Approach Velocity	Wing Weight	Thrust to Weight
Landing Field Length	Hor. Tail Weight	Wing Area
Takeoff Field Length	Vert. Tail Weight	Wing Aspect Ratio
Takeoff Gross Weight	Landing Gear Weight	Wing Taper Ratio
Operating Empty Weight	Hydraulics Weight	Wing Sweep
Block Fuel Weight	Engine Weight	Wing Thickness Ratio Root
NOx Emissions	TSFC at Start of Cruise	Wing Thickness Ratio Tip
Acquisition Cost	L/D at Start of Cruise	Hor. Tail Area
RDT&E Cost	Induced Drag	Hor. Tail Aspect Ratio
DOC+I	Zero Lift Drag Coeff	Hor. Tail Taper Ratio
Avg Required Yield per RPM		Hor. Tail Thickness Ratio
		Vert. Tail Area
		Vert. Tail Aspect Ratio
		Vert. Tail Taper Ratio
		Vert. Tail Thickness Ratio

Table 3: Design Variable Ranges for 70 and 300 Passenger Class Airliners

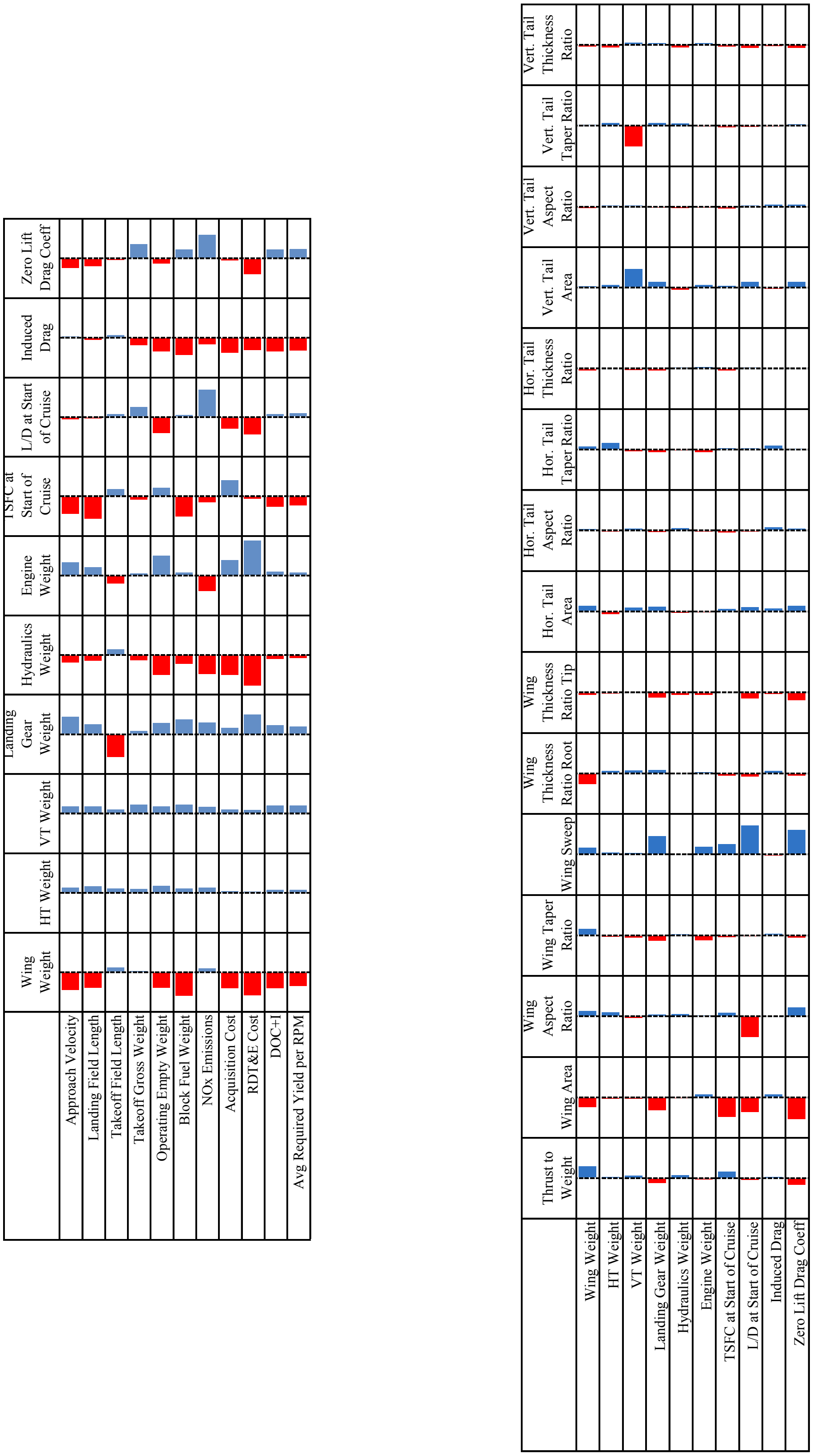
	70 pax		300 pax	
Design Variable	Min	Max	Min	Max
Thrust to Weight	0.33	0.38	0.26	0.31
Wing Area (ft^2)	750	830	4500	6500
Wing Aspect Ratio	7.35	9	8	10
Wing Taper Ratio	0.3	0.4	0.19	0.25
Quarter-chord Wing Sweep (deg)	26	35	27	37
Wing Thickness Ratio Root	0.11	0.13	0.1	0.13
Wing Thickness Ratio Tip	0.11	0.13	0.09	0.12
Hor. Tail Area (ft^2)	170	250	900	1100
Hor. Tail Aspect Ratio	3	5	3	5
Hor. Tail Taper Ratio	0.4	0.55	0.34	0.38
Hor. Tail Thickness Ratio	0.11	0.13	0.07	0.105
Vert. Tail Area (ft^2)	100	180	550	700
Vert. Tail Aspect Ratio	0.8	1.4	1.15	2.3
Vert. Tail Taper Ratio	0.5	0.7	0.2	0.5
Vert. Tail Thickness Ratio	0.11	0.13	0.08	0.11

whatever the baseline value was for the set of design variables. For this research the baseline values for a set of design variables are what is important and they are difficult to treat independently. On the contrary, the intermediate metrics only change value as a result of the design variables changing value.

The set of correlations for each design point were compared against each other by taking the difference. For ease of interpretation, these were tabulated graphically in Figure 21. For each box, a bar indicates the difference between the two design points. If the bar is red and to the left of the dotted line, the 300-passenger class vehicle had a higher correlation. If it is blue and to the right of the dotted line, the 70-passenger class vehicle had a higher correlation. The larger the bar is, the greater the difference. Ideally, all bars should have zero length, indicating that the two design points share similar physical trends and there is no difference in how different classes of vehicles behave.

Looking first at the differences for the intermediate metrics versus design variables in Figure 21b, the two design points line up fairly well. The largest differences are in the relationships for wing area (where the 300-passenger class design has higher correlations) and the wing sweep (where the 70-passenger class design has higher correlations). The other difference is the difference in vertical tail weight versus vertical tail area and taper ratio. These two relationships are the only two differences for the vertical tail weight and are roughly equal in opposite directions.

The differences for the system requirements versus performance requirements as shown in Figure 21a are much more pronounced. The larger relative wing size has greater impacts through the wing weight and hydraulics weight as well as the induced drag. On the other hand, the engine weight (a reasonable substitute for engine size) drives a much larger research, development, testing and evaluation (RDT&E) cost for for the smaller craft. Most of the difference continue to be aligned by column, as with the lower level relationships. The relationships based on engine characteristics are less consistent suggesting that the engine performance relationships are more complicated and may require additional measures



(a) System Requirements vs Intermediate Metrics

(b) Intermediate Metrics vs Design Variables

Figure 21: Comparison of Correlations for 70- and 300-Passenger Class Airliners

to be considered.

Some of these differences are due to differences in the ranges of the design variables. Even for variables that have been normalized in such a way that they would be comparable, the ranges are not equal. While this might be a concern, the truth is that these ranges have been tailored to the vehicle classes. A larger vehicle has sufficient size to support a more slender wing because the stiffness due to actual thickness (not thickness ratio) increases at a faster rate than the span or chord. Forcing the two designs to use similar ranges would force one or both to be in suboptimal design ranges which is a poor design problem to start with.

The point of the experiment was to show that an expert, if given only enough information to identify that the system was an airliner without specifying the size or mission, would have to assume aspects of the aircraft. It is unlikely that all experts would make the same assumptions and that those assumptions would line up with the final design. Furthermore, without recording the assumptions experts used, there is no way to identify when the model is valid for future problems.

Consider the relationship between an arbitrary design variable and an arbitrary performance metric. If you poll many experts about the relationship between these two without defining the system, there are many possible correct responses based on each individual's implicit assumptions about the input range, the usage scenario, values of other inputs, a baseline design and so on. The combination of these relationships appears as shown in Figure 22a. These undefined assumptions are each a degree of freedom in the design. While the current push is to keep as much design freedom as possible for as long as possible, the design process as a whole is intended to intelligently reduce degrees of freedom to a single design. Some of the decisions to fix those degrees of freedom have already been made but have not been identified explicitly. Once these degrees of freedom are taken into account by experts, their responses will move towards Figure 22b. There is still some variability, but they are in better agreement and the uncertainty around the true value (as

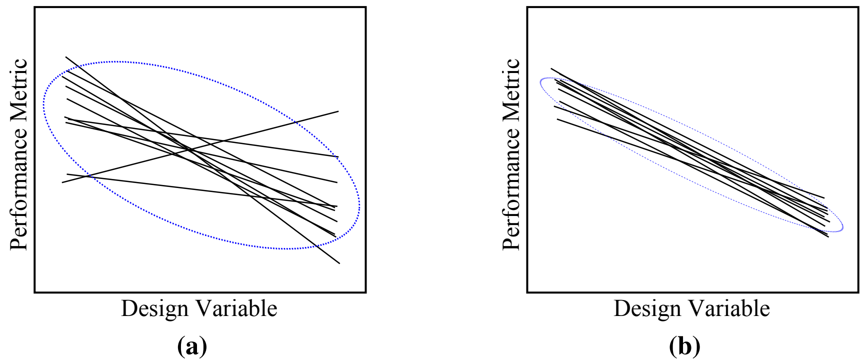


Figure 22: Reduced Variability as a Result of Reduced Degrees of Freedom in Design

represented by the blue ellipse in each image) is reduced.

3.2.2 Design and Mission Concept Definition Alternatives

One of the processes for reducing the degrees of freedom in a design is known as concept definition. Traditionally, and in many companies still, concepts are defined by a small group of people in an *ad hoc* fashion. Structured methods have been developed to aid in this process. Within these structured methods, there is frequently a distinction between concept generation and concept selection where concept generation produces a single concept without evaluating its performance and concept selection is the process of reducing the set of concepts to a small number or single concept. Liu et al. describe the concept generation process as a divergent step and the concept selection as a convergent step where the design is abstracted to a larger space and then evaluated and reduced to smaller set of designs [84]. The concept selection step cannot occur without concept generation. When building a model from a blank sheet, therefore, both are necessary.

Okudan and Tauhid performed a survey of concept definition methods developed over a twenty-eight year span [102]. The majority of these methods fall into two categories: those that operate over a small number of qualitatively- or minimally quantitatively-defined concepts and those that operate over a large number of quantitatively-defined concepts. The first category includes methods such as Pugh's evaluation matrix [106] or those based on the analytic hierarchy process (AHP) [111]. These methods can be performed using only expert-based information, but require a small set of concepts to be identified, something that should not be relied upon for solving a general problem. The second category includes the use of s-Pareto frontiers [86], joint probability decision making, grouping concepts into fuzzy sets to use AHP[127], and other approaches for grouping concepts or to reduce them *en masse*. Many of these methods require some form of modeling to generate the data necessary to reduce the results. Since this research is most appropriate when no such modeling exists, these methods are also inappropriate.

A well-accepted approach for structured concept generation without the need for quantitative information ahead of time is Fritz Zwicky's morphological method. While it is excellent at concept generation, it does not have a defined process for concept selection. An improvement to the morphological method known as an interactive reconfigurable matrix of alternatives (IRMA) provides a structured method for both concept generation and concept selection. The next sections discuss these approaches in detail.

3.2.2.1 Morphological Analysis

In 1948, Fritz Zwicky formalized the morphological method as a way of decomposing complex problems and looking at specific components and their interrelationships [135]. As an astrophysicist, Zwicky first demonstrated the method with a morphological matrix to define the possible types of telescopes one might build. Each row in the matrix represented a particular attribute of the telescope, such as the ratio of energy entering the aperture to the energy absorbed in the recording instrument, the choices of recording instruments or the type of interaction of the light with the optical parts with the telescope. For each row in the matrix, there were a number of options. For the recording instrument row, this included photographic plates, ionization chambers, and photocells. By selecting one item from each row, one created a particular class of telescope.

He later generalized the method further using geometric tessellations and ideas of an *n*-dimensional morphological box[136]. Others took it and developed a more engineering-friendly form of the matrix by using a tabular format and established structured processes for creating the morphological matrix, now sometimes referred to as a matrix of alternatives for the different alternatives that could be generated using it.

An example matrix of alternatives for the design of a burrito is shown in Figure 23. The example includes several different styles of rows as an example. The first two rows, tortilla type and protein type, are the traditional type of row where one must select one and only one option from each row. For the third row, type of beans, is slightly different since it includes an additional option for the choice of abstaining. The "none" option is typically listed when it is important to differentiate between making a decision not to include an option from that row and having not yet made a decision. The fourth row, rice, could have been represented by a row labeled "Starch" with the options of "rice" or "none", but the yes/no choices make it clearer. The tradeoff is that it is more difficult to change later to add another option if the restaurant decides to offer couscous or another type of rice as an option. The last two rows, vegetables and sauces, are significantly different in that they allow more than one option to be selected. This depiction is selected instead of having a separate yes/no row for each item as was done for rice to make the design cleaner and also because the options logically fall into categories. There is no standard way to denote the difference as of yet, so it is important that individuals using the matrix of alternatives are knowledgeable about the problem and the options.

Category	Options			
Tortilla	Corn	Wheat		
Protein	Beef	Steak	Chicken	Tofu
Beans	Black Beans	Pinto Beans	None	
Rice	Yes	No		
Vegetables	Lettuce	Onions	Red Peppers	Jalepeños
Sauces	Sour Cream	Guacamole	Salsa	Pico de Gallo

Figure 23: Matrix of Alternatives for a Burrito

It is common to calculate the number of alternatives represented within a matrix of alternatives. The total number of alternatives is the product of the number of possibilities for each row. Since the first four rows are made up of mutually exclusive options, the number of possibilities is the number of options listed (2, 4, 3, and 2 in order of row). The last two rows are different since each combination of options is a different possibility. For these rows the number of possibilities is $2^{n_{Options}}$, or 2^4 for each row in this case. The total number of alternatives is calculated to be

$$2 \times 4 \times 3 \times 2 \times 2^4 \times 2^4 = 12,288 \tag{4}$$

unique burritos. These options are just the discrete choices. Variations between the quantity of any particular ingredient on the burrito further increases the possible "designs". The traditional method for concept selection where a handful of options are generated by a group of individuals would be the equivalent of four types of pre-made burritos that might be available at a catered meal.

The use of any matrix of alternatives approach would provide a reliable, structured, and repeatable method for concept and scenario generation. The concept and scenario selection process when using the morphological method is frequently an *ad hoc* approach depending on the individuals present and their own experience. Recent developments have created interactive versions which include additional information about the choices to aid decision-makers and experts in down-selecting to a single concept or set of concepts.

Several methods have integrated them within the design process more directly. Ölvander demonstrates the applicability of applying an automated and iterative matrix of alternatives approach linked with a modeling and simulation environment[103]. In his research, he describes a series of alternatives for cooling, actuation and electric systems components in a "More Electric Aircraft" example. His integrated matrix of alternatives optimizes the selections for each component based on a series of vehicle level characteristics. For this research, integration with a modeling and simulation environment is neither required nor desired. No such modeling environment is available yet and creating one that is detailed enough to account for step changes at that level would be a significant undertaking outside the scope of the intended use. The important aspects to improve are the ones that make it easier to make decisions in the absence of detailed information.

A method that focuses on this is Interactive Reconfigurable Matrix of Alternatives (IRMA) originally developed between 2004 and 2006[49]. A matrix of alternatives created using IRMA has the additional benefits of capturing compatibilities between items within and between rows, applying interactive filters that affect all rows, multiple attribute decision making (MADM) to support selecting items within each row, automated calculation of the total number of alternatives remaining, updating the options available based on the previous decisions to support real-time and collaborative decision-making. The next section discusses the details of creating an IRMA and some of the theory behind it.

3.2.3 Development of an Interactive Reconfigurable Matrix of Alternatives

While there has been a great deal written on various aspects of morphological analysis, IRMA has only been introduced and the process and implications have not been well-discussed in literature. This section will attempt to present a coherent overview of the method, its purpose, and how to apply it to a generic problem before addressing how to apply it to the particular problem of expert-based model creation.

3.2.3.1 Standard Matrix of Alternatives Development

The first few steps of developing an IRMA are identical to those for any other matrix of alternatives. It begins with a decomposition of the system, object, or phenomena described within it. The specific type of decomposition may vary with the problem. For systems, the common suggestion is to perform a functional decomposition and identify the different things that a system must do in order to meet its requirements. This type of decomposition is especially important for revolutionary and unconventional designs or for problems that have not been solved before. For a simple aerospace vehicle, one would start with the different functions necessary for the vehicle to accomplish its objective. It would need to “Generate Lift” to get off the ground, “Provide Thrust” to move to its destination, “Carry

Cargo” to move whatever or whomever needed to be relocated, and “Land Safely” at its destination. For other systems it may help to trace the flow of energy, matter or information through the system to ensure that no functions are left out. For a flashlight, energy would start with “Store Energy”, and then travel through some activation method “Control Energy Flow”, to the point it is put to use to “Generate Light”. The flashlight might also need to “Focus Light” and would need a way to “Protect Workings”.

Other problems are better described through a physical decomposition. Zwicky himself used a physical decomposition in his own original example. Physical decomposition is generally better for problems which are already well-understood and partially-defined by other means or for problems that are primarily focused on classification rather than design. When the aerospace vehicle has been defined as a military strategic airlift vehicle, it is unlikely that any method other than fixed wings would be used to generate lift or that the cargo would be carried in anything other than a pressurized internal cabin. While it is possible to decompose these functionally further into “Provide Wing Strength” and “Provide Extra Lift” or “Provide Cargo Access” and “Provide Cargo Security”, it is more natural to use “Wing Material” and “High Lift Devices” or “Location of Cargo Doors” and “Cargo Tie-down System”. Some examples can be decomposed functionally, but the result is awkward and forced. The burrito example is a good demonstration of the failure of only functional decomposition. While the tortilla row could be represented as “Provide Flavor Containment”, the other rows would all fall under “Provide Flavor” or “Provide Filling”. Even so, the options fall into a logical physical decomposition that is also demonstrated by the order they are usually added. This functional decomposition might leave out important attributes such as seat material and color that are important to product development.

Regardless of the decomposition, the next step is to identify options for each function or physical attribute. These options may come from any number of sources. For well understood problems, these options may be included in open literature. If other groups or competitors have solved a similar problem, it is important to include the options they selected. Brainstorming is an excellent source of new options, especially when those involved are encouraged to be creative. It is always easier to eliminate unrealistic options later than to find alternative concepts when the original choices cannot meet the design requirements. The resulting IRMA will be a living document that can be reused in the future, so additional options should be added continuously throughout the design process.

3.2.3.2 Choice of Interactive Platform

The previous steps could be performed with a standard table-form matrix of alternatives on paper, but since the goal is to make this process interactive and make use of digital technology, it makes more sense to do this within the same framework that will be used to exercise it. This brings up the point of selecting and creating that framework.

The original versions of IRMA were developed within Microsoft Excel. Excel offers a rapid prototyping environment that allowed easy calculations, arrangement of the user interface and simple formatting while being widely available and readily accessible without library or runtime environment concerns. Other versions have been created as stand-alone applications using Java. There has recently been a move to create and use web-based versions so that the information is centrally stored and available to multiple individuals at different locations simultaneously.

One of the decisions that should be made is whether to present all of the functions or physical attributes and their respective options at the beginning or to build them up as the user makes a selection in the previous row. These are known as destructive and constructive IRMAs respectively. The benefit of a destructive IRMA is that, by showing the full realm of options available at the outset, it allows the designer to start anywhere within the structure and jump around quickly. This is frequently useful when certain options can be immediately eliminated and any other options that would have required them also be eliminated. A constructive IRMA only presents options in a particular order which can help guide a less experienced designer and reduce the number of options a designer is presented with at any given time. A constructive IRMA requires slightly more information than a destructive IRMA requires and is slightly harder to implement.

While IRMA has been recreated in many different formats, one of the largest motivations in its development was to provide a standard framework that could be used for many different problems without the need to recreate the interface and back-end logic. As new instantiations have moved away from Excel, it has become increasingly modular. Maintaining standardized formulae, formatting and data formats allow users to transition between platforms and reuse existing efforts. It also means that designers used to a particular format do not need to relearn the interface before contributing to the process. Whichever platform is used to develop an IRMA should be generic enough to allow information to quickly be added or changed.

The format is centered around the common layout of a typical matrix of alternatives with additional user interface elements built on to it. For the demonstrator, each row was limited to five options across for readability. As a particular function or physical attribute had more options in a single row, more than six or seven, it was difficult to compare options against each other due to the difficulty of seeing them all at once. Spanning such attributes over multiple rows allowed for more options while still being easy to read and comprehend. The example in Figure 24 shows three options per row in order to fit within a printed layout.

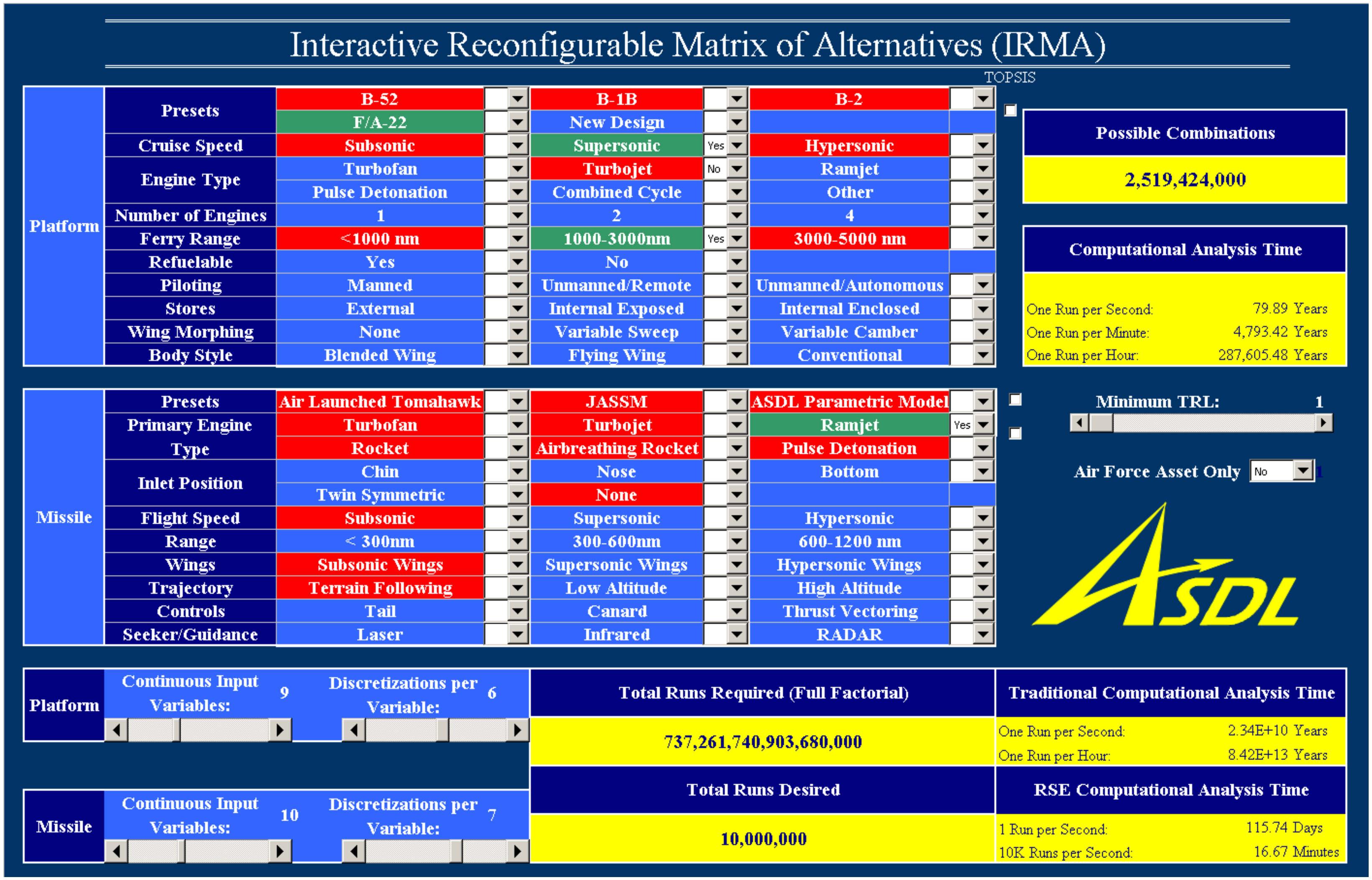
Each options has one of three colors: green for options that have been selected, red for options that have been eliminated and blue for options that have not had any decision made. These three possible states are mirrored by the three options in the drop-down boxes for each option: Yes, No and blank. Options with a red background may have been eliminated directly by selecting No next to them or by the filters and compatibility matrices discussed in the next section. Positive, negative, and neutral positions allow the designer to approach the problem from multiple directions depending on the particular situation. It is often easier to eliminate options that don't fit the current problem rather than to select a particular option.

Similar to the standardized format of the table, the data should also have a standardized format. The more modular the data is, the more it can be reused and expanded from project to project. In many decision-making tools, the tool is designed for a single problem and requires creation from scratch whenever a project shifts focus, moves to a more detailed design phase, or given a new goal. Using a modular tool enables the design process to continue with the same tool when increasing the fidelity of the data, changing MADM setups, or adding new options. This encourages a continual improvement of data to ensure that all decisions are made using the best information available. The reusable aspect means that new projects can spin up more quickly when they are able to build upon previous projects. These features are what allow any IRMA to be truly reconfigurable.

When using IRMA in an IPT, its speed enables collaboration on the decisions made. Tools that require users to wait for an answer are inherently difficult to use in groups of mixed backgrounds. When the team is waiting on the tool to calculate, focus shifts away from the task. Since much of the basic data and compatibilities are input beforehand, the team can focus on the decision-making without needing to focus on disciplinary basics. The data needed to support MADM would be input beforehand by the appropriate subject matter experts. As more data is brought to the table and integrated into IRMA, the collaboration process yields better, more informed design decisions.

3.2.3.3 Compatibility Matrix

Not all alternatives that can be selected in an IRMA are realistically feasible. There are certain combinations of options which are incompatible. For example, while there may be some way to create a retractible engine, it is generally not possible to build a hypersonic-class aircraft powered by a piston-driven propeller engine. Unless the problem mandates such unusual combinations, it is best to identify those incompatibilities early and eliminate options that are incompatible with previous decisions. At the same time, there may be some



Interactive Reconfigurable Matrix of Alternatives (IRMA)

TOPSIS ☐

Platform	Presets	B-52		B-1B		B-2	
		F/A-22		New Design			
	Cruise Speed	Subsonic		Supersonic	Yes	Hypersonic	
	Engine Type	Turbofan		Turbojet	No	Ramjet	
		Pulse Detonation		Combined Cycle		Other	
	Number of Engines	1		2		4	
	Ferry Range	<1000 nm		1000-3000nm	Yes	3000-5000 nm	
	Refuelable	Yes		No			
	Piloting	Manned		Unmanned/Remote		Unmanned/Autonomous	
	Stores	External		Internal Exposed		Internal Enclosed	
	Wing Morphing	None		Variable Sweep		Variable Camber	
	Body Style	Blended Wing		Flying Wing		Conventional	

Missile	Presets	Air Launched Tomahawk		JASSM		ASDL Parametric Model	
	Primary Engine Type	Turbofan		Turbojet		Ramjet	Yes
		Rocket		Airbreathing Rocket		Pulse Detonation	
	Inlet Position	Chin		Nose		Bottom	
		Twin Symmetric		None			
	Flight Speed	Subsonic		Supersonic		Hypersonic	
	Range	< 300nm		300-600nm		600-1200 nm	
	Wings	Subsonic Wings		Supersonic Wings		Hypersonic Wings	
	Trajectory	Terrain Following		Low Altitude		High Altitude	
	Controls	Tail		Canard		Thrust Vectoring	
	Seeker/Guidance	Laser		Infrared		RADAR	

Possible Combinations

2,519,424,000

Computational Analysis Time

One Run per Second:	79.89 Years
One Run per Minute:	4,793.42 Years
One Run per Hour:	287,605.48 Years

☐ ☐

Minimum TRL: 1

Air Force Asset Only: No

Platform	Continuous Input Variables: 9	Discretizations per Variable: 6	
Missile	Continuous Input Variables: 10	Discretizations per Variable: 7	

Total Runs Required (Full Factorial)

737,261,740,903,680,000

Traditional Computational Analysis Time

One Run per Second:	2.34E+10 Years
One Run per Hour:	8.42E+13 Years

Total Runs Desired

10,000,000

RSE Computational Analysis Time

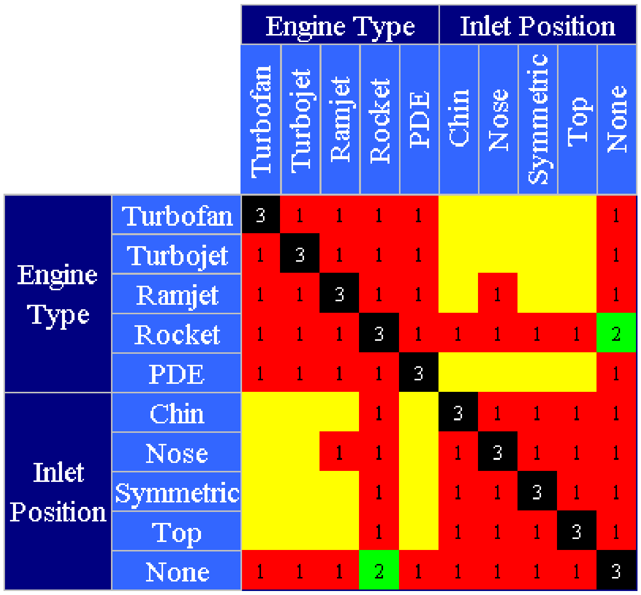
1 Run per Second:	115.74 Days
10K Runs per Second:	16.67 Minutes

Figure 24: Example IRMA Main Interface

decisions that mandate a particular option. In a generic missile matrix of alternatives, there might be a design option for engine inlet position that must be set to “None” if a rocket is selected as the engine type (assuming an air-breathing rocket is not considered).

One of the most significant characteristics of IRMA is its use of compatibility and dependencies between options to intelligently down-select or eliminate inappropriate options in other categories. This feature allows information to be brought forward in the design process and enables better decisions by making it instantly clear how one decision in one dimension affects the choices for all other appropriate dimensions. The example above may be obvious to many designers, but for more complex relationships, this compatibility information may not be known by all involved with the concept definition process. Being able to capture this information for later use helps bring forward tacit information and improve the speed of defining concepts and quality of the resulting concepts with less dependence on who may be present at the moment. When the system-level interactions are clearer to the decision-makers, the elimination process can focus more on the system effects rather than the individual components or categories. A list of compatibilities and dependencies for each option is essential to be able to eliminate or require other options based on the choice’s state. Note that even a partially completed list can narrow the number of possible combinations by orders of magnitude. For IRMA, a compatibility matrix is used to record and track the relationships. The compatibility matrix is also used to differentiate rows in which items are mutually exclusive (such as the first four rows in the burrito example) from the rows where multiple items can be selected.

Figure 25 shows a small portion of a compatibility matrix. For this particular case, all of the engine types and all of the inlet positions are mutually exclusive as indicated by square groups of red boxes containing the number 1. While it may be physically possible to have more than one engine type on a missile (and may occur for a cruise missile where there is both a booster engine and a cruise engine) or more than one engine inlet, for that



		Engine Type					Inlet Position				
		Turbofan	Turbojet	Ramjet	Rocket	PDE	Chin	Nose	Symmetric	Top	None
Engine Type	Turbofan	3	1	1	1	1					1
	Turbojet	1	3	1	1	1					1
	Ramjet	1	1	3	1	1		1			1
	Rocket	1	1	1	3	1	1	1	1	1	2
	PDE	1	1	1	1	3					1
Inlet Position	Chin				1		3	1	1	1	1
	Nose			1	1		1	3	1	1	1
	Symmetric				1		1	1	3	1	1
	Top				1		1	1	1	3	1
	None	1	1	1	2	1	1	1	1	1	3

Figure 25: Example IRMA Compatibility Matrix

particular design problem, it is extremely unlikely that such a combination would be realistic. Some incompatibilities are absolutely physically impossible, but many more are a result of the design problem. Other incompatibilities are also shown in a similar fashion. The relationship between engine inlet position and engine type for a rocket is shown as a dependency with a green box containing the number 2. The other two types of boxes are the yellow blank boxes which indicate that no relationship is identified between the two options. Along the diagonal is a line of black boxes containing the number 3. This type of compatibility enforce the reflexive property that ensure that, if a designer selects an item that has been eliminated due to an incompatibility, the IRMA assumes the designer has additional information and allows an override.

While this particular example is symmetric about the diagonal, it is not always the case. If within engine type, the rocket option was split into a solid fuel rocket and a liquid fuel rocket, selecting one of the two would require the “None” inlet position option. However, selecting no inlet would only eliminate the non-rocket options but leave both the solid and liquid options unselected.

The compatibility matrix is populated after the first iteration of the traditional matrix of alternatives has been filled out. While any small amount of information can make an impact, more information available in an intuitive format tends to lead to better decisions. From experience, the compatibility matrix is the most time-consuming steps of the process and benefits from distribution among several groups of experts focusing on separate areas of the problem or separate subsystems. In many cases, compatibility matrices from previous projects can be re-used in whole or in part, greatly reducing the time spent in this area.

3.2.3.4 Matrix-wide Filters

The compatibility effects for a given option are activated when another option has been explicitly selected or excluded. Filters are a method for applying a certain requirement across multiple rows independent of their relationships between each other. A classic example is to filter by technology readiness level (TRL). To design a system with low technical uncertainty, one would want to select only mature technologies and may eliminate any options which have not been demonstrated on a prototype. By setting a TRL filter to only include options of TRL 6 or greater, all technologically immature options would be eliminated. Another filter may eliminate any options which are not available in-house to minimize programmatic costs of subcontracting. Coarse Filters such as the TRL are useful early in the design process when the opportunity space is large, but little quantitative information is available for decision making. Later in the process, a fine filter such as life-cycle cost in dollars may be used. Fine filters typically have discrete values that are defined during the requirements development process. Specifying these specific thresholds is often much more difficult since those values depend on a variety of interacting factors. Relative scales are often much more appropriate for filters. Assigning each option a value on a scale of one to nine relative to the other options would allow the same comparisons without the need for extensive knowledge or pre-design work.

The data for the filters can be split up among subsystems also, but generally requires a system-of-systems level of interpretation to focus the decisions on maximizing the system's capabilities at that level. Referring again to the engine example, though specific ranges of thrust-specific fuel consumption are available for various cycles, it may be more useful to give them relative values based on those, but under the category of “Range” instead of “TSFC”. Filters that focus on either maximizing or minimizing particular characteristics of options have been the most effective in most cases, but in some situations Boolean filters (such as limiting options to those developed in-house) may be essential.

3.2.3.5 Multi-Attribute Decision Making

The use of compatibility information and filters can help narrow the scope of what we are looking at. Eventually, it will be necessary to select between the remaining options within a row. For some complicated rows, it is helpful to use data about each option to rank and eliminate the lowest performing options. Considering the point that IRMA is used within this research, it is unlikely that specific and accurate data is available for each option since this usually requires modeling or historical data to gather. Luckily, there are a number of MADM approaches that can utilize subjective data about options. Many of these methods also allow for a designer to vary the importance values for particular attributes depending on the problem and the customer’s preferences.

Population of the MADM tools is performed in the same manner as the filters, but may be more focused on category- or subsystem-specific characteristics. At the simplest level, a user would only differentiate between the options of a single category. Since the options within a category will have many common attributes among them, the attributes used for comparison are easier to determine and populate. An alternative to comparing a row at a time is to compare combinations of options. This presents several more problems since the combinatorial problem of a matrix of alternatives comes back into play. This method is generally better used when the number of options has already been reduced or there are several concepts identified. In this case, each combination could be treated as a separate option with a set of combined characteristics. Upon making a decision, all of the individual options in that combination are labeled appropriately in the main sheet. Having filled in the appropriate data and making whatever iterations necessary, the matrix of alternatives is now ready to perform the down-select exercise.

3.2.3.6 Calculation of Alternatives Remaining

A detailed matrix of alternatives for a complex problem can be quite large. The total number of possible combinations grows exponentially with the number of rows. This large number can sometimes be lost in the simplicity of the interface. In order to help provide a status check on how close to a defined concept the users are, many instantiations of IRMA include summary statistics such as “Number of Possible Combinations” and “Computational Time Required”. These also allow managers and supervisors to maintain an appreciation for the size of the design space and the difficulty in downselecting to a single concept. The primary statistic is the number of combinations remaining. This number is presented by finding the product of the options on each independent row or category, as described for the burrito example. The true number of combinations is less than this value when due to the incompatibilities. Unfortunately, calculating the difference in the actual and estimated numbers of combinations requires a full iterative search of the compatibility matrix and quickly becomes prohibitive. Since the precise value of the number of combinations remaining is rarely useful, the estimate is sufficient for most problems. Any mutually exclusive categories must have their options summed instead. For example, if one row describes options only available for a turbofan engine with five options while another describes three equivalent options specific to a turbojet, the total number of options between these two would only be eight rather than fifteen.

With a large and fully populated matrix of alternatives, the number of combinations quickly becomes too large for most people to comprehend. In order to make the magnitude easier to interpret, an equivalent time to calculate was included. In this case, the total amount of time required to evaluate each concept assuming several highly optimistic analysis times (one concept per second, per minute and per hour) were used to translate into days and years of computation time required. Since the development of a model for a single concept could take days, months or years if performed traditionally, these estimates are significantly smaller than the actual time required to evaluate all remaining combinations analytically.

3.2.3.7 Using the IRMA for Concept Selection

A populated matrix of alternatives or IRMA could be considered the end of the concept generation process. While it would be possible to create a list containing each of the possible concepts represented, such a list would likely be prohibitively long. The process to downselect from these options is shown in Figure 26. These will determine how well categories and individual options meet the system requirements. The first instinct for most users is to use the most stringent requirements and filter settings in an attempt to quickly reduce the number of combinations. Like many design methods, it is better to gradually eliminate options and focus on meeting the requirements first before looking for optimal concepts. Setting each of the filters to the maximums would be akin to maximizing each of the individual components without regard to the overall system. Throughout the process, users must perform sanity checks to ensure that any elimination makes sense and that remaining combinations are valid. Much of this work is performed by the compatibility matrix, but because of the size, errors may go unnoticed until the tool is exercised. The process also may lead to being able to eliminate specific options manually, allowing the compatibility matrix to eliminate any options marked as required for them.

Once the broad strokes have been made, the user or group should move to utilizing the MADM tools starting at a top level and moving to row-specific decision making tools. For many MADM tools, it is necessary to determine the appropriate weightings of attributes. The MADM can be applied at several levels: between options of a single row, between

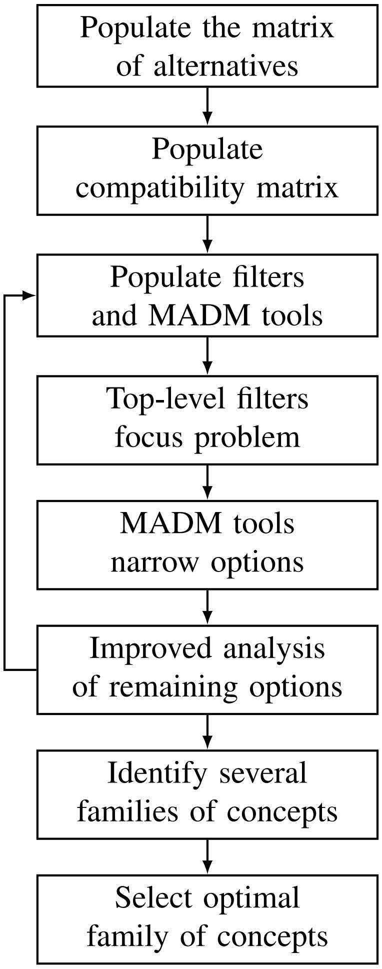

Populate the matrix of alternatives
Populate compatibility matrix
Populate filters and MADM tools
Top-level filters focus problem
MADM tools narrow options
Improved analysis of remaining options
Identify several families of concepts
Select optimal family of concepts

Figure 26: IRMA Concept Selection Process

families of concepts incorporating unique combinations of several rows, or between fully defined concept families incorporating options from each category or row. For each level the MADM is used at, there are two forms of reporting its results back to the main worksheet to display the results to the user. The first is to identify the best option and allow the user to confirm the choice by accepting it. The second is to mark all options below a certain threshold or rank as undesirable using the same method as those that are shown to be incompatible. The first reporting method produces faster results and speeds up the down selection processes. However, it may also prematurely eliminate options and combinations of options that are optimal at a higher level. Eliminating only the lower ranking options encourages a more gradual process which allows further discussion and analysis among the important few. MADM methods that report numerical scores for each option, allow a designer to identify concepts that are similar in performance. If those concepts also share similar options, the next iteration can select those options

Next, the team should go row-by-row through the entire IRMA to eliminate any options that do not meet the customer's requirements. The group should determine what criteria are used at this point, since much of the input may be increasingly based on tacit knowledge. Once the team or user has made a pass through the entire matrix, the next step is to focus on the remaining options. If the design space has been sufficiently reduced, it may be wise to incorporate higher fidelity modeling to generate more information about the remaining families of alternatives. Further analysis may be either for each individual category or for a combination of two or more categories. Practice has shown, however, that the benefit of the remaining options generally depends more on the interactions between the categories than the performances of the individual selections. Unfortunately, modeling combinations of several categories results in the same combinatorial problem of any matrix of alternatives. One solution to this is to pick a primary driving category and several driven categories. The goal should be to find combinations in the driven categories that are primarily a result of the decisions made in the driving category. As an example, if the driving category is the engine type, select the best inlet type for a ramjet and analyze that combination as one option and analyze the best inlet type for a turbojet as another option. Another MADM tool should be used to compare the different combinations in an objective and analytical manner when possible. Beyond the MADM tools, individuals engineering knowledge should be used to identify more promising combinations for further modeling. The results of this further modeling should then be again compared against the expected result and used for further down select.

The key to a successful down select using IRMA is the iteration associated with any design process. Each time a filter is used or a MADM tool eliminates poor choices, the number of combinations decreases and the amount of information about the system increases. Users should pay attention to the possible combinations remaining to note when it is feasible and appropriate to increase the detail of the categories or of the supporting data. As there are fewer options, it is reasonable to expect that the amount and quality of information for each individual option should also increase. With more information, it is possible to make further decisions and reduce the number of options again. After sufficient iteration, only one option is selected in each category, leaving a single unique family of concepts.

3.2.3.8 Differences in Using IRMA for this Research

Much of the process for creating, populating and using an IRMA is appropriate for any problem where an IRMA is used. In a traditional design process where the customer has given a great deal of time, the IRMA may be developed and used starting from nothing over the span of a week or more with several iterations and time to gather feedback from many individuals. The current research is intended to define a methodology that can be completed on the order of hours and cannot allow such a long period of time for concept definition. Instead, it must be assumed that a group intending to create an expert-based model is already familiar with morphological analysis and IRMA and already has an existing matrix of alternatives that has been populated as part of the group's standard field of business. It is also assumed that the larger decisions within the matrix of alternatives can be determined directly from the customer's requirements and a rapid scoping of the problem. It is not necessary to fully define a concept, so long as the definition is sufficient that the experts all start with the same assumptions. If an expert requests clarification on a particular aspect of a concept, it is important that any decision be added to the IRMA if it is not already present. In this way, IRMA serves as a single location to document assumptions about the system to be modeled. The test of the usefulness of an IRMA for this research is not whether or not the experts can use one, but rather whether or not the accuracy of the information they provide improves when given a completed IRMA as compared to the accuracy of experts who do not have access to one.

3.2.4 Model Definition

Once the system has been defined in broad terms using an IRMA, the attention turns to defining the properties of the model. These steps determine what will be needed in the future steps and lays out the specific form of the model.

When defining the problem for modeling, it is also necessary to define the needs and scope of the model. Some aspects of that scope have already been defined as part of the motivation and background investigation. For example, the desire to use hierarchical sets of system requirements, intermediate metrics, design variables and potentially subsystem design variables was based on the QFD-style format. Other decisions will be determined as part of this research, such as the form of the equation used in the resulting model (discussed in Section 3.4.1). There are several aspects that must be determined each time an expert-based model is created, such as which particular requirements, metrics and variables, should be included in the model. In the preliminary experiment, these were based in a large part on what was readily available for the model used in that experiment.

3.2.4.1 Required Fidelity

The required fidelity of the results drives what models are suitable to provide them. In the case of expert-based modeling, it should have already been determined that a low fidelity model is suitable. If not, this point is the proper place to have that discussion to determine whether this is a suitable solution. Using the proper fidelity to meet the requirements is important and if expert-based modeling isn't sufficient to answer the question with the proper amount of certainty, there may not be sufficient reason to continue with the process. During the testing of the model, the degree of variability and uncertainty present in the model can be assessed, but experiments here have shown that error of even a high quality model with good experts may be 10% or higher. If, at the end of the modeling exercise, the error is too large to be useful, the time lost may be insignificant in the larger scheme of design.

3.2.4.2 Selection of Variables and Ranges

When performed in the absence of a truth model for verification, as would be the expected usage, the choice of those values will have to be determined by some other means. The system requirements should be the easiest as they come directly from the explicit and implicit customer requirements. These may also include measures that the customer is less concerned with but affect the group performing the modeling.

Determining the appropriate intermediate metrics and design variables is more difficult. Those experts involved should have a good feel for what tend to be the most important and descriptive intermediate metrics and design variables to include. Methods as simple as brainstorming followed by discussion would be sufficient to identify the major contributors to satisfying the customer requirements. If experts are unable to identify the most important design variables and intermediate metrics, it is also unlikely that they would be able to provide relationships between them. If there is debate about whether or not to include a particular item in any of the three lists, it is best to include it and allow dissenting experts to give it a low or no impact on the next level. The balance to including measures is to realize that, as the number of relationships increases, the potential for error and the time required to provide information also increase. Experts are more likely to give a relationship a low score than a zero and as a large number of small impacts accumulate, so does the error.

There are some other constraints on what makes good design variables and intermediate metrics. One of the most important qualities is that the set should be as independent from each other as possible. An example of the effect of correlated intermediate metrics is shown in the preliminary experiment in Section 3.4.2. Identifying uncorrelated design variables is typically easier than for the intermediate metrics since, for most designs, any of the design variables can be changed on its own. Intermediate metrics are more difficult since they are all connected to the design variables and often the physics affects multiple metrics in the same way. Even using nondimensional measures is not always a solution due to the way most nondimensional measures are defined. For example, including both the lift coefficient and the parasite drag coefficient of an aircraft is troublesome since both are defined relative to the wing area. If the wing area increases, the values of both will change in the same direction at the same time for the same reason. For some of these variables, even though they would be linked in a traditional model, experts may be able to treat them separately to produce accurate results. But such a situation makes it difficult to compare against a traditional model that is unable to follow those same mental distinctions.

It is necessary to define ranges for the design variables during this phase. As the preliminary experiment shows, the size and location of the design space can have an effect on the relationships. Relationships may be mostly linear over one range and nonlinear over another range. In some cases the sign of the relationship may change. The ranges for each design variable should be determined either by the full group of participants or by a smaller group leveraging past designs and their own experience of what possible design ranges would be suitable. While larger ranges may be suitable for traditional models, if ranges are too large here, experts have difficulty providing relationships with certainty. Since this same problem is part of any design process, many modeling groups likely already have best practices in place for this. As part of a later step, experts can optionally provide estimates for the ranges of the intermediate metrics and system requirements and they do not need to be defined now. It is helpful to identify the units of these ahead of time for consistency.

3.3 Gather Information

Once the system, mission or scenario, and the measures of interest are defined, the next step is to gather the information necessary to create the model. QFD has already been identified as the starting point for the model, so it is reasonable to start with the information present in a QFD. Any gaps between this starting point of data and whatever is the minimum required to produce a useful model must be identified and closed. Some of the gaps can be identified based solely on the needs of a model. Others are better identified experimentally. This step may also be referred to as parameter estimation in other modeling methods. Vansteenkiste and Spiet note that parameter estimation has been the focus of far more research efforts than other step of most modeling efforts [126]. This is for good reason as this is often the most sensitive part of model development.

The previous preliminary experiment used correlations as a surrogate for "perfect" expert data. The next preliminary experiment will test how well a set of data from a human in the form of a typical QFD compares to the "perfect" data.

3.3.1 Preliminary Experiment: Accuracy of an Expert

Traditional QFD uses a qualitative scale to describe the relationships between levels. This typically includes choices for high, medium, low or strong, moderate, and weak as well as the option of leaving a position blank for no relationship. Using the same system requirements, intermediate metrics and design variables as the model used in the first preliminary experiment[2], the author filled out the central relationship matrix for the two levels of a QFD, despite not regularly working with civil airliners. This was performed prior to performing any modeling or analysis of the different designs. The author only identified the problem as a civil airliner, but did not identify a vehicle size class, mission or other concept definition information. One of the motivations to select an aircraft-based test problem rather than something more abstract where traditional models did not already exist was to ensure that the author would be able to give data as an expert for these initial tests.

To provide a consistent basis for comparison, the correlations from the previous experiment were reduced to qualitative scores according to the limits in Table 4. There are no universal values for what correlations are considered high, medium or low since they

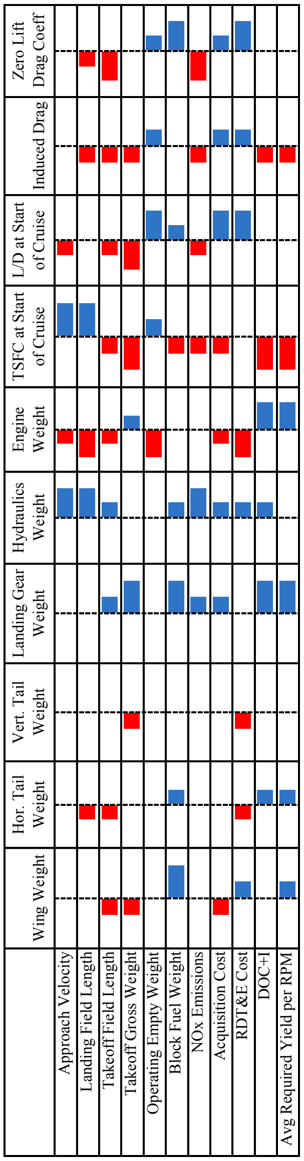
[2] Listed in Table 2 on page 80

vary significantly based on the problem. The values here were selected to give a typical distribution of high, medium and low scores as well as to minimize the difference between the expert-based values so that the comparison is the best possible. The two levels have different ranges due to the strong differences in typical correlations at each level. The differences were then compared in a similar fashion as before. In this case, the difference is measured between qualitative scores. If the expert-sourced score was high, but the model-based score was medium, the difference would be -1 with the expert overestimating the strength of the relationship. Both the 70-passenger class and 300-passenger class models were compared to the expert-sourced data and found that the 300-passenger class matched better. The graphical comparison between the expert and the 300-passenger airliner model is shown in Figure 27. Red bars to the left indicate expert overestimated the strength of the relationship, blue bars to the right indicate the expert underestimated.

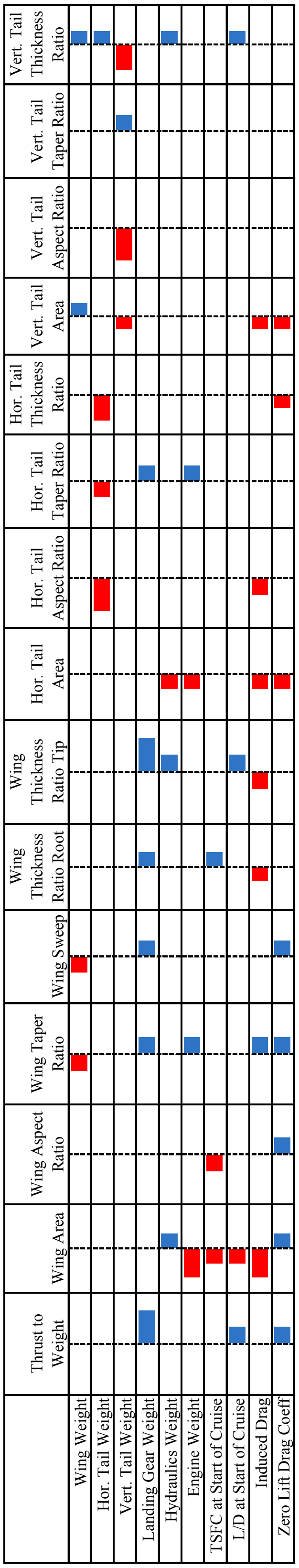
Table 4: Ranges of Correlations Corresponding to Qualitative Scores

Qualitative Score	System Requirements vs Intermediate Metrics	Intermediate Metrics vs Design Variables
High	> 0.85	> 0.70
Medium	0.50 – 0.85	0.30 – 0.70
Low	0.09 – 0.50	0.05 – 0.30
No Relationship	< 0.09	< 0.05

Even with the efforts to minimize the errors within the comparison, there are still noticeable differences between the expert-sourced information and the model-sourced information. The differences in the intermediate metrics versus design variables are small for the most part. There was a tendency for the expert to overestimate the effects of wing area and the sensitivity of the horizontal and vertical tail weights. On the other hand, the expert underestimated the sensitivity of the landing gear weight. Overall, the differences are small. The differences for the system requirements versus intermediate metrics are much larger. The expert significantly underestimated the relationships between the landing gear and hydraulics weights on the system requirements across the board. The effects from



(a) System Requirements vs Intermediate Metrics



(b) Intermediate Metrics vs Design Variables

Figure 27: Comparison of Expert-Sourced Relationship Scores and Truth Model for 300-Passenger Class Airliner

the engine metrics were overestimated. For the higher level, where the correlations were stronger overall, the errors were also larger.

There are several explanations for the differences. Some of the smaller differences may be due to the choice of score bins for the correlations. Slightly increasing or decreasing the bounds between the scores may reduce these individually. If this process were a part of the modeling methodology, it would be worth including fuzzy logic to mitigate the effects of correlations near the borders of two scores. Some of the relationships, such as those for the landing gear and hydraulics weights, may be the result of secondary effects that drive both the intermediate metrics and the system requirements. It is also possible, though less likely, that there are errors in the model when computing intermediate metrics that do not affect its accuracy calculating system requirements.

Admittedly, some errors may be due to the expert. Most expert-based methods combine the data from multiple experts to average out minor differences and to identify areas where a single expert may be incorrect. Some expert-based models take this further by having all data gathering occur as a group to promote discussion and consensus-building. In this case, the data from a single expert does not have access to such corrections. It may also be possible that the individual does not have sufficient expertise in the realm of civil airliner vehicles, as shown by the overestimates of certain groups of relationships. The next few sections identify ways to address these concerns.

3.3.2 Number of Experts

The use of multiple sources of data or multiple independent measurements from a single source in any type of model reduces measurement error with the goal of converging to the true values. Asking a single expert the same question multiple times would not produce independent measurements since experts have memory of previous responses. Asking multiple experts the same question does not provide independent measurements of a single data source, but still serves the same purpose: to reduce noise and converge toward what is hopefully the true values. If experts are able to see what others' responses, they can use that information to help make better decisions. This has the risk of the experts who respond first or who speak the loudest driving all others. Another benefit of data from multiple experts is the ability to validate one another. If four experts respond with one value and a fifth responds with a very different value, that fifth expert becomes an outlier that is worth investigating. While it is possible the outlying expert sees the problem in a correct light that the other four did not, it is more likely that the other four are correct. If an expert consistently provides very different values, it casts doubt onto the reliability of that expert.

In general, as more data from more experts is included, each of these three benefits improves. Including more experts has negative effects as well. Experts generally have to be paid, and the more experts that are involved, the more billable man-hours are required, raising the monetary cost of producing the resulting model. This also pulls those experts away from other work that may also be essential to the program, perhaps negatively impacting overall progress. Including more experts also increases the calendar time required, especially if all experts must meet together to provide data. There is always the possibility that additional experts just aren't available. It is important to determine a guideline for the minimum number of experts required to produce a useful model to minimize the cost and time required as well as to determine beforehand if a group possesses the necessary resources to create the model.

Typical guidance for providing information for a QFD is to have five and ten people involved. For customer-focused QFDs, the workshop should have an upper limit of forty people, half of whom would be members of the customer's organization [61]. Most of the guidance for QFD only specifies participants, without regard to their expertise and is focused on general business problems or nonspecific product development (rather than technical product design). It is suggested that a diverse group from multiple aspects of the business participate including sales, marketing, engineering, management, manufacturing and so on. For this research, general participants and diverse groups are not useful for creating a technical engineering model. Since the work is more specialized, it is believed that fewer experts will be necessary. Raczynski recommends inviting 15–20 technologists to the voting workshops as part of the SOAR process [107]. This value is based on Equation (5) where n_O is the number of observations required given the t-value of an associated α level, standard deviation s and acceptable margin of error d [10].

$$n_O = \frac{t^2 s^2}{d^2} \tag{5}$$

This equation and values is intended for use in selecting the number of samples necessary to represent the true behavior of a larger population. For this research, the goal is not to represent a larger number of experts with data from a subset of them, but rather to have enough experts to converge on true physical behavior. Still, the equation is valid for statistical analysis. It is easy to assume a value of $t = 1.65$ to correspond with an α level of 0.10. Raczynski used a standard deviation of $s = 1.167$, but this was based on a pre-workshop estimate rather than experimental data. The acceptable margin of error in this equation describes the margin of error of each response. In building a model, the margin of error associated with the output values of the model are more important. This is directly impacted by the number of terms in the model. With a constant margin of error for each term, as the number of terms increases, the margin of error of the output also increases. A margin of ±10% about a mean slope of zero has a much smaller impact on the overall error than the same margin around a slope of nine.

Since both the true variance of responses from experts and the margin of error of each term are not known, it is difficult to guarantee an accurate number of experts required. Instead, the necessary number can be identified experimentally. The experiment described in Chapter 5 will use a larger number of experts than would likely be available in order to be able to identify a standard deviation as well as account for additional sources of variability. These results will also be used to identify the “knee in the curve” of accuracy of the resulting model versus the number of experts included.

3.3.3 The Right Experts

The quality of any data is greatly dependent on the quality, accuracy and reliability of the source of that data. For expert-based modeling, that data source is the experts themselves. Therefore, the experts themselves must be of high quality, accuracy and reliability. This requires individuals who are knowledgeable on the subject area, motivated to give their best effort, able to express themselves, and diverse enough to capture all aspects of the problem.

3.3.3.1 Required Level of Knowledge

The first requirement of participants is that they have the knowledge and information needed to create the model. It’s easiest to discuss whether individuals possess this information in terms of their qualifications. Section 2.1 previously discussed the qualifications of an expert in detail. While education and experience are necessary qualifications, they are not always sufficient to identify the level expertise in a specific area. Having recognized competence is also essential. A facilitator responsible for direction model creation will identify and select experts rather than sending an email to all employees to ask for volunteers.

Several methods, including the Cooke method, test experts’ knowledge against known values for a similar problem to identify how accurate and reliable the information they provide would be. This need to identify and test seed information is a major contributor to studies using the Cooke method requiring between one and three man-months with elicitation sessions lasting a few hours each [37]. This testing is performed after identifying experts and is used primarily to apply different weightings to the information from different experts where those experts who are more accurate are given more influence on the final results.

Hammond splits the measurement and evaluation of expert judgment into two categories: coherence and correspondence. Coherence describes how well judgements ensure that all the parts of a particular belief or theory fit together logically. Correspondence describes how well that belief or theory matches up with the facts. He shows that expert judgments tend to be either "precisely correct or wildly incorrect" when measured by coherence. When measured by correspondence, judgments are accurate within the realm of physical perception, but lose accuracy as the conceptual context and uncertainty of the system increases. Hammond identifies education as the primary driver for good coherence of human judgment, but that particular (rather than general) experience drives correspondence. [62]

Though imperfect, education, amount of experience, and reputation are often the only measures available to identify which experts should be invited in the first place. These factors are continuous, but for the purpose of discussion they can be discretized into groups. To provide a frame of reference, these groups will be equated to classes of individuals at a typical research-focused educational institution.

A first-year graduate student would be the equivalent of an entry-level new hire at a company. Such an individual would have general knowledge of the topic area, but would be unlikely to have specific experience or detailed understanding of theory in a particular area. The upside is that they are plentiful and typically easy to schedule time with. Within a company, their time also tends to have a lower cost.

A doctoral-level student has gained two years of additional specialized education and experience. They are expected to have a good understanding of the theory in their area of expertise and have developed some intuition.

Research engineers and post-doctoral researchers have spent multiple years performing research and have proven their knowledge in their particular specialization.

Professors have the greatest experience and sufficient understanding of the corresponding theory to be authoritative in their fields. These would be the equivalents of technical fellows within a company. Professors have the least free time, are the most difficult to schedule and also the most expensive.

If selecting solely on level of expertise, one would always pick only professors to provide information. However, there are far fewer professors available in a particular specialty than research engineers and fewer research engineers than doctoral students. The individuals further down the list are more expensive and scheduling is more difficult. The goal is to balance availability and cost against the minimum level of expertise that is available to find the least restricted data who produces a useful model.

Even within and between those classifications there are different levels of experience in specific problem areas. Some first-year graduate students may have more experience with civil airliners than a professor who works exclusively with hypersonic vehicles. And while there is a certain level of knowledge a more experienced individual is expected to have, if it has not been exercised recently, any data that individual gives is dubious. In testing the method, individuals will be classified based both on their experience with aircraft as a whole, as well as their specific experience with the class of vehicle being modeled and how long it has been since they actively worked with that class of vehicle.

3.3.3.2 Ensuring Motivated Experts

Just because an expert has the information necessary to provide good information does not guarantee that they wish to do so. In some cases, an expert may view their participation in an expert-based modeling activity as an attempt to make their own position redundant by eliminating the need for them to make a full model. In these situations, an expert has a vested interest to ensure that the resulting model (and the method used to produce it) cannot be trusted and may intentionally provide inaccurate information. An individual may also feel ambivalent about the exercise if he or she sees no direct benefit. While this individual may not actively sabotage the information they give, they will also not put in the mental effort required to provide accurate information for complex relationships.

The more motivated an individual is to provide good information, the more likely it is that they will provide the best information they are capable of providing. Several approaches are useful here. The first is to identify individuals beforehand who are interested in actively sabotaging the results or simply disinterested and not including them. These individuals are frequently outside of the organization sponsoring the model creation effort. Contractors who are paid to provide a similar type of modeling product should certainly not be included.

Individuals who are working with the facilitator on a day-to-day basis outside of this particular will be more likely to at least appear to put more effort into their participation. Close-knit groups who frequently work together will be more likely to support each others efforts in the spirit of camaraderie or a common vision and goals. However, even when being paid for their time, not all employees put forth their best effort unless there is some measure of their performance. Since the models being created are frequently stopgaps, it is possible to score how accurate each individual is. This can be used to promote an informal friendly competition with a running set of rankings among peers or a more formal measure of knowledge associated with a bonus or material reward of some sort.

Perhaps the most effective method to motivate enthusiastic participation is to clearly show how the effort will have a positive impact on their own work or the final product. This means that the resulting model is put to use and is also an improvement over the status quo. Such proof would typically require using it several times within an organization so that individuals are convinced that it works for “their problem”. Individuals who are convinced help to convince their colleagues and develop an environment that places high value on providing good information.

3.3.3.3 Suitability of Crowdsourcing

With the growth of the internet, it has become possible to very easily interact with a large number of people with a very low cost. Crowdsourcing is the process for using such interactions to collect information or resources by distributing the load over a large number of sources. Some examples for collecting information would include Wikipedia or product review sites. Examples of collecting resources include Kickstarter or massive gift exchanges. One of the primary characteristics of these approaches is that individuals are not pre-screened or pre-selected to participate. Anyone who wishes to join in can do so easily.

The crowdsourcing approach for gathering information more applicable here than gathering resources. Since individuals giving information are generally not pre-screened or selected specifically, it is necessary to identify another method for identifying which information is good and which should not be trusted or used. Some efforts, such as the pre-internet effort of the Oxford English Dictionary to collect new words in common usage, rely on an organizing body to filter through and judge the collected information. Other efforts, such as Wikipedia or the content-aggregator Reddit, rely on the community of users and information providers to self-police themselves to remove incorrect or bad information and retain good information. And still other efforts, such as predicting the outcome of a basketball tournament based on combined individuals' predictions, use all data that is provided with the belief that either bad information will be in the minority or that bad information is still useful information.

Collecting good information from individuals reduces to a problem of ensuring that volunteers are motivated to provide such information. While there may be a small payment associated with providing information, such as Amazon's Mechanical Turk, most methods are unpaid and rely on individuals' altruism or some form of public recognition such providing individuals with scores or ranks. Relying purely on altruism tends to significantly reduce the percentage of qualified individuals able to participate who choose to. Methods using scores or ranks to reward individuals require either a testable result, such as a basketball tournament, or repeated scoring, such as Yahoo! Answers' points system.

Most methods require little-to-no knowledge or expertise to participate. The method presented in this research requires sufficiently technical knowledge about the system and field being modeled. For an aircraft example, a participant would need to know what the thickness-to-chord ratio of a wing and the drag coefficient mean in order to provide intelligent information about any relationships involving them. It is unlikely that such information would be present in a general population or even within a group of employees at a company sufficiently large to be considered crowdsourcing.

It is unlikely that crowdsourcing would be effective for this process because of the knowledge required. It is possible to attempt to use crowdsourcing from a likely knowledgeable populations such as at a large conference for the field being modeled. In such an environment, it is difficult to motivate individuals to give an hour or more of their time to be trained to give information and then perform the exercise, and may create a conflict of interests. It is also difficult to motivate individuals to give correct information if they do participate. Many might be willing to participate out of curiosity, but the results would be dubious. Performing this with a set of paid employees would get expensive in terms of man-hours very quickly and would not likely be cost-effective when compared to traditional model development.

The experimental results discussed in Sections 6.3.2 and 6.7 show the results of using individuals who already meet a certain level of knowledge and motivation. As these individuals approach the minimums for both knowledge and motivation, the accuracy of the results decreases. The use of crowdsourcing would result in a larger number of individuals with less knowledge and potentially less motivation, further decreasing the average accuracy.

3.3.4 Collecting Information from Experts in Groups Versus Individually

No matter what number of experts or who they are, the environment in which they give information can impact the information they give. There are tradeoffs for whether the experts meet together in a group or provide information individually.

Most QFD exercises take place in a group setting with all stakeholders present where experts discuss and negotiate to agree on a single value. This is commonly known as the committee or BOGSAT (bunch of guys sitting around a table) method. This method has the upside of getting a single value that all those present at the meeting have agreed on, at least in theory. The problem with using the committee method is that generally the most well-respected, the most charismatic or the most stubborn expert drives the result towards his or her opinion. This significantly reduces the benefits of multiple experts. The committee approach also requires scheduling all experts for the same time and preferably in the same location. For some groups, especially those with other demands on their time, it can take weeks to find a time that everyone is available at the same time. Experts may also be required to travel to a common location that further increases the cost and time requirements for gathering information. The meeting could take place with only the experts that are available, but this reduces the number of individuals who can participate and may reduce the average level of expertise of those individuals since those with more experience also tend to be more difficult to schedule.

SP2 and SOAR solve part of the problem of a single expert or small number of experts driving the results by incorporating anonymous electronic voting for each item of data. This also provides a set of data for each relationship that can be used to show the level of agreement. Discussion among the group takes place whenever the votes are significantly divergent. This method still requires experts to meet together physically. Since all experts vote simultaneously, the process can only go as fast as the slowest individual and as fast as the facilitator can read the information prompts. [75, 107]

The Delphi method is an iterative approach that aims to reduce the logistics required for meetings as well as minimize the effects of charismatic and stubborn experts. This approach asks all participating experts to give a position statement or an initial set of data based on common information. These initial responses are distributed to the other experts who then revise their own statements based on the information of the rest of the community. This process is continued iteratively until the experts converge upon a solution. Depending on the problem and facilitator, the responses may be distributed anonymously or with the full identification of the expert who provided each response. Identifying the source gives other experts additional information to determine how much to trust that particular response and how much to use it to influence their next iteration. Unfortunately, this may also lead to a small number of senior, well-known or best-respected experts to drive the direction of the results. Distributing the responses anonymously eliminates this bias, but experts are less likely to incorporate information from unknown, and therefore untrusted, sources. While this method does not require the logistics of a group meeting, it requires time for experts to give information, exchange information, read and interpret the new information, give revised information, and repeat. While some of this process could certainly be sped up with automated and interactive tools, there is a limit to how fast this process can move. [39]

If one removes the iterative review process from the Delphi method or the requirement to co-locate from SP2/SOAR, the result is a simple single data collection from multiple, possible geographically distributed experts with a reduction in the total time required. Unfortunately, it also gives experts only a single chance to get it right. Cooke and Goossens make the argument that expert elicitation exercises are not intended to serve as opportunities for experts learning and that experts were selected for their pre-existing knowledge [37]. This is especially true when experts have been selected for their ability to contribute to model development. For practical reasons of this research, the limitations on human experimentation make safeguarding the identities of participants a priority in reducing their risk. Having group data gathering activities significantly increases the difficulty of getting approval from the appropriate bureaucracies. This limitation would not be present outside of a research activity. Therefore, for all these reasons, all data will be collected from individuals separately for this research.

3.4 Create the Model

Model creation is the process of taking the known and believed information about a system or physical behavior and identifying the correct form of an equation, constants within the model, and method for calculation. For engineering design models, a parametric model where a set of inputs produce one or more outputs is preferred. In fundamental model development this includes using Buckingham's Pi theorem for dimensional analysis and scaling of results using universal constants or sizing factors [45]. For historical and surrogate modeling, it involves regression of the available data and adding terms or performing transformations to improve the fit of the regression [97]. For structural or fluid finite element analysis, creating the model is rendering the geometry of the object, generating a mesh on the object, and applying boundary conditions and material properties [81].

For the ALTER methodology, creating the model is the process of combining the information gathered from experts using a pre-determined model form into a parametric equation. This research must identify the appropriate way to combine that information and which model form to use. The following sections discuss the choice of model form and the implications on the type of information that should be collected from experts in the future. Preliminary experiments are used as a source of observations to help support and drive the choice and identify possible shortcomings.

3.4.1 Identifying a Baseline Model Form

There are several goals to keep in mind when selecting a method to flow up information. The method should minimize the information that is necessary to create the model. The motivation for this research was the case where a low-fidelity model did not exist and creating one would take more time than was available. As the amount of information required to build the model increases, the savings of time and effort over a traditional low-fidelity model decreases. Since one of the secondary motivations for this research is to extend the usefulness of QFD, the model form should make use of as much information pre-existing in the QFD as possible. This also means minimizing the information that is discarded. Modifying the type of information that is gathered during the QFD process is preferred over additional information that would be collected outside that process. Whatever the general form, the resulting model should strive to be correct enough to be useful. George E. P. Box famously wrote that "all models are wrong; the practical question is how wrong do they have to be to not be useful" [21]. This research intends to test that practical question by examining the trade between accuracy and correctness against speed and ease of creation, focusing on the latter.

3.4.1.1 Inverse QFD

Considering the reliance on QFD, the math used to flow down information within a QFD is a good place to start. A QFD is limited to translating the priorities of the customers' requirements into the importance of engineering characteristics, parts characteristics, or manufacturing characteristics. For the highest level of a QFD, the priorities of the engineering characteristics v_k are calculated based on the multiplication of the vector of weightings on customer requirements w_i and the matrix of relationships between the two M_{rel}.

$$v_k = w_i M_{rel} \tag{6}$$

The same process is used for lower levels. While this is very helpful for prioritizing characteristics, it is not useful for identifying particular values of those characteristics. If the reverse was true, the priorities of the engineering characteristics would be known and the priorities of the customer requirements would be calculated using the pseudo-inverse M_{rel}^{+}.

$$w_i = v_k M_{rel}^{+} \tag{7}$$

Unfortunately, the priorities of the system requirements are not the values of interest for predictive modeling and the values for the weightings of the lower levels are not known. This makes this particular formulation inappropriate, so another approach is needed.

3.4.1.2 ROSETTA

ROSETTA translates between QFD and modeling, making it another reasonable starting point. ROSETTA generalizes the problem beyond priorities and specific levels with a different notation. ROSETTA uses the following generalized equation to calculate the priorities of engineering characteristics from modeling and simulation.

$$\frac{\partial Q}{\partial m_l} = \sum_{i=1}^{n} \sum_{k=1}^{p} w_i \frac{\partial R_i}{\partial m_k} \frac{\partial m_k}{\partial m_l} \tag{8}$$

This research already has the QFD representation and wants to be able to calculate values of the requirements *R* from it, which is assumed as part of the model in ROSETTA. This value is found by integrating Equation (8) which produces an equation of a series of nested functions over the set of all *R*, *m* and *x*.

$$Q = \sum_{i=1}^{n} w_i R_i \left(m_k \left(m_l, x_h\right)\right) \tag{9}$$

Knowing *Q* is not particularly interesting, since the values of *R* could be fed into any of a number of multiattribute decisions making methods other than a simple weighted sum. This further reduces the equation to show that *R* is a function of all *m* and *x*.

$$R = f \left(m_k \left(m_l, x_h\right)\right) \tag{10}$$

While correct, this is far too general to be useful for building a model since it offers no guidance to a useful model form. This does not help us predict the values of the customer requirements, so another approach is needed.

3.4.1.3 SP2 and SOAR

The SP2 and SOAR processes discussed earlier flow up some information to estimate the impact of funding different programs or technologies. They start with the same QFD-style matrix but instead of inverting it, they simply reverse the equation with the same relationship matrix.[3]

$$R = M_{rel}m \tag{11}$$

This can easily be expanded to show the simple underlying linear model for each R_i.

$$R_i(x) = M_{1i}m_1 + M_{2i}m_2 + M_{3i}m_3 + \quad + M_{ni}m_n \tag{12}$$

SOAR also includes a normalization step to account for differences in the magnitude of values for M_{ij} so that the OEC-representation of R_j has consistent values. [107]

The linear terms with constant values for M_{ij} mean that the change of each input variable is the same no matter what the current value is and does not depend on any other settings (under the assumptions of the design space identified in the Define the Problem step). If the input is the amount of funding for a technology or the number of personnel assigned to a project, this means that the increase is the same for the first \$100K or first person as the increase from \$800K to \$900K or from nine researchers to ten. This may not be true for all cases, but for an initial down-select it is sufficient. Some instances of SP2 include a nonlinear funding profile to capture the difference of impact, but that may not be necessary here.

[3] SOAR actually multiplies all the hierarchical matrices together to create a single matrix and avoid calculating intermediate steps. The math is equivalent since matrix multiplication is associative, so it is separated out for the sake of clarity here.

The linearly additive combination of terms means that there are no interactions or cross-terms between the independent variables of each equation. For technology planning, this means that the effects of each program are independent, similar to using additive *k*-factors. This assumption is not necessarily valid for technologies since some technologies improve the impact of others, while others may reduce the benefit of others. This effect is small enough or infrequent enough that the additive nature is still a valid assumption. A similar assumption is made for the modeling here.

SOAR and SP2 limit each of the values of the relationship matrix to one of seven qualitative discrete choices which are assigned a numerical value *a posteriori*. This means that the values of R, m and x are all unitless. The OEC that combines the values is a normalized measure of goodness, but does not correspond to a measurable quantity. Any analysis performed with this approach can only be comparative, and not strictly predictive. Since there are only seven discrete scores possible, even if predictive values were possible, there would be no guarantee of a perfect model. This may be sufficient for engineering design, but a predictive output with correct units would be more useful if it is possible.

3.4.1.4 Other Forms

A slight upgrade from the pure linear approach would be higher-order Taylor-series polynomial. As the order increases, the model would be able to capture increasingly complex relationships. With each increase in order, the number of parameters and thus the amount of data needed from experts increases exponentially. This would also require a significant departure from using baseline QFD-type formatting. While this is certainly possible, it is wise to test the applicability of the simpler form first and only increase the order if shown to be necessary.

At an extreme would be feed-forward multilayer neural networks, which have been shown to be universal approximators[71]. This means that it is possible to represent absolutely any relationship using a sufficiently complex neural network. While this would still qualify as a standard form, the model parameters are nonintuitive and extremely sensitive to noise. There are also significantly more parameters than would be necessary for the linear model form or the Taylor-series polynomial form. Neural networks are best applied to problems where a large number of observations are available to regress or train the model. This is not the case here.

The most accurate model would be a custom form developed for each response using dimensional analysis or knowledge of the true physics[45]. Since each expert might make slightly different assumptions beyond what could be captured in an IRMA, each model form might be slightly different and very difficult to combine into a single model. Even if they were able to be combined, the result would be the low-fidelity model that was eliminated as part of the motivation for this research. If developing the low-fidelity model can be done quickly enough, then that approach is preferable to a generic form with expert-based information.

Very few aerospace systems, if any, truly behave according to a linear model. However, any smooth and continuous curve can be approximated with a line if the range of the independent variable is small enough. For a given design range, if the linear effects are dominant, a linear model may be sufficient to capture the trends most important to making very early design trades even if the trend is not entirely linear. Another preliminary experiment will provide evidence to support or reject this hypothesis.

3.4.2 Preliminary Experiment: Model Form

This preliminary experiment will test two separate issues: "How accurately could a linear model of the form above represent the behavior of the true relationships?" and "How well could QFD-style data serve as the slopes of such a model?"

Testing the first issue is performed by creating a linear regression of data produced by a higher fidelity "truth model" and compare the accuracy of the fit. The process used for this is the same as a typical surrogate modeling exercise with a simplistic model form. The requirements are regressed against the metrics and the metrics against the design variables. Each resulting regression is tested using the coefficient of determination R^2, the root mean squared error (RMSE) and plots of actual value against predicted value and the residual error against predicted value. Normally goodness of fit tests would include some measure of model representation error, but in this case the number of points is significantly larger than the degrees of freedom and the test of interest is whether or not a linear model could fit the data, not the predictive capabilities of the particular resulting regression. If the regressions pass the goodness of fit tests, it is conceivable that a linear model would be sufficient in some cases.

The second issue is tested by comparing the slopes of the accepted model from the first test against existing expert-sourced data and the simulated "perfect" data representing that information. The expert-sourced data are the QFD values used in the second preliminary experiment. The "perfect" data is the correlations used in both the previous preliminary experiments. Since the magnitude of the slopes depend on the magnitude of both the inputs and outputs of the particular relationship, they are first scaled by multiplying the difference between the maximum and minimum values of the input range. This changes the physical meaning from a slope to the total vertical change in the output due to each term. This value is consistent and comparable for all terms in a given regression. The t-ratio for each term is also included to show the statistical certainty that each term drives the variability of the output.

Each of these four is normalized using the infinity norm ($||x||_\infty$) to allow ease of comparison. Since the magnitude of the values in a QFD are selected primarily for ease of interpretation, normalizing in this way does not impact their meaning so long as the ratio between them is maintained. Since this it the point of comparison, the same normalization must also be valid for the other measures, allowing all four measures to be compared simultaneously on the same axes.

For this experiment the data created for the 300-passenger class airliner for the previous preliminary experiments was used since the expert-sourced QFD best matched that class.

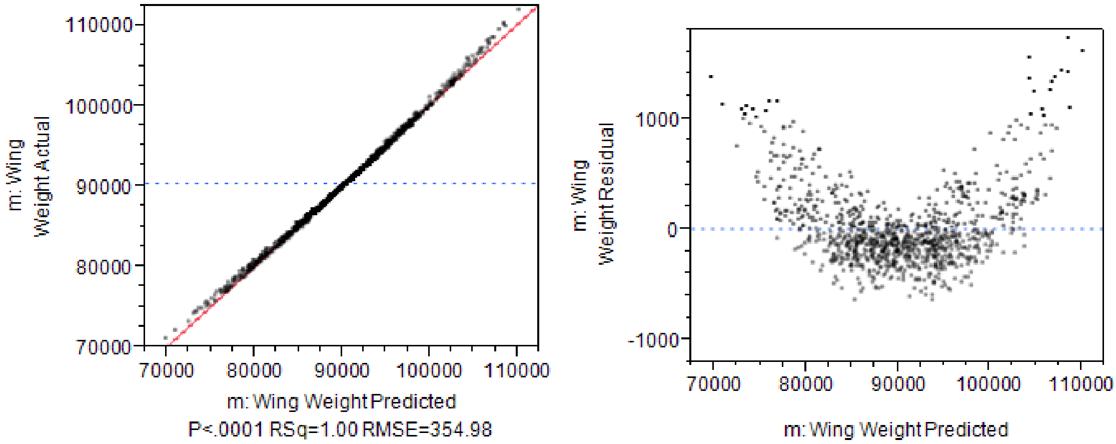
The preliminary models fell into three groups. The first, referred to as type A, had good model fits and the correlations between higher- and lower-level variables were very similar to the scaled coefficients. The second type, type B, had poor fits that could be solved with a slightly more complex model than the pure linear form. The final type, type C, had good model fits, but the correlations between higher- and lower-level variables did not correspond to the values of the scaled coefficients due to confounding. The next sections discuss each of these three types with results from a representative example.

3.4.2.1 Results Type A: Good Fits and Good Matching

Most of the models that were fit with linear regressions passed both tests easily. As an example, the results from the fit of the engineering metric wing weight are shown here. Wing weight was regressed against all of the design variables in Table 2 and a constant intercept term. The resulting model had a coefficient of determination (R^2) of 0.998 and a root mean squared error (RMSE) of 354.98 or 0.3% of the mean value. The actual-by-predicted and residual-by-predicted plots are shown in Figure 28. The actual-by-predicted shows that the points are very tight around the diagonal. The residual-by-predicted shows a clear concave-up trend, but the magnitude of the error is approximately two orders of magnitude smaller than the predicted (and actual) values.

For the example of wing weight, it appears that the regressed model has a good fit and would serve as an acceptable surrogate if needed. Therefore, for this particular case, it is clear that a pure linear model is sufficient to capture the relationships and the assumption of a pure linear model is a good one. This supports the choice of a linear form.

The second test evaluates the correspondence between data correlations, expert-sourced QFD values, model coefficients, and the statistical significance of those coefficients. The terms of the linear model for wing weight engineering metric are shown in Figure 29 in order of decreasing absolute value of the t ratio. Of the four lines plotted, the values for



(a) Actual by predicted

(b) Residual by predicted

Figure 28: Summary of Goodness of Fit Measures for Wing Weight

the QFD are the most divergent. The other three values are identical for the first six terms of the model. After that, the correlation coefficient does not follow the same trend as the scaled model coefficients (here labeled as Y Range) and the t ratio.

For the correlation coefficients and scaled model coefficients to be so similar is very promising. This suggests that, if the correlation coefficients are accurate surrogates for the expert-based values, that the scaled model coefficients can be adequately described from this same source of data. That the expert-sourced information stands out is the result of several effects. Some of these, specifically the accuracy issues, were discussed in Section 3.3 as part of the Gather Information step. However, even if the accuracy were perfect, this plot shows clearly that the limitations of positive-only values and a discrete numeric scale consisting of four values is too coarse to capture the detail necessary. Solutions for these shortcomings are discussed in Section 3.4.3.

With the exception of thrust-specific fuel consumption (TSFC), all of the engineering characteristics could be adequately modeled with just a pure linear model and had similar trends for the correspondence of the four relationship measures.

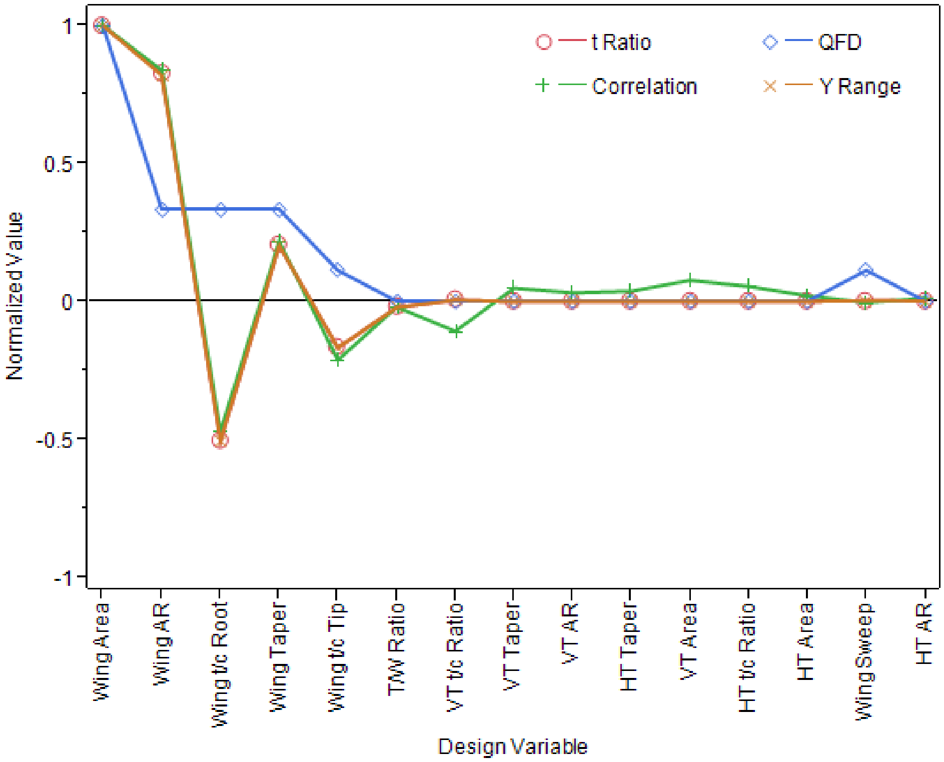
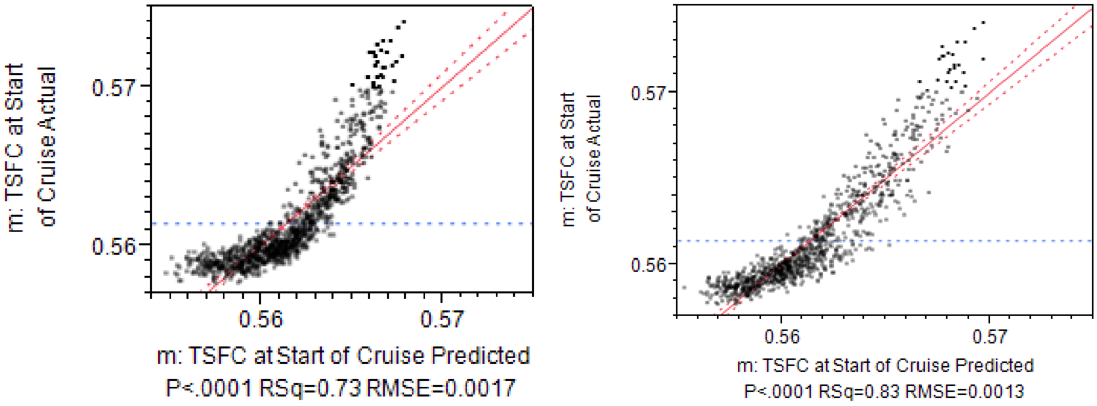


Figure 29: Summary of Normalized Model Coefficients for Wing Weight

3.4.2.2 Results Type B: Poor Fits

One of the intermediate metrics (TSFC at start of cruise) and one of the customer requirements (takeoff field length) had poor fits from the linear model. These two models are classified as Type B. Since TSFC at start of cruise had the poorer fit, it is examined in more detail here. The pure linear model has an R^2 of 0.727 and a RMSE of 0.0017 or approximately 0.3% of the mean value. While the RMSE is quite good and the R^2 is good enough to establish that a relationships exists, the actual-by-predicted plot in Figure 30a shows a clear pattern off of the diagonal significant error. Note also that the points are not evenly distributed along the diagonal. Since the behavior has a curve to it, it is clear that there is a major nonlinear effect present. Since these two regressions failed the first test, the second test was not performed.

As a first attempt to improve the fit, a single quadratic term was added. A single term is preferred to only require additional information where necessary (rather than upgrading



(a) Pure linear model

(b) With one quadratic term

Figure 30: Comparison of Actual by Predicted Plots for TSFC With and Without a Quadratic Term

every term to the maximum available). A quadratic term is used as the non-linear term for a similar purpose as well as the ease of averaging quadratic and linear effects if not all experts agree on the type of relationship. In this case, after testing all of the inputs, using a quadratic term for wing area had the greatest effect. This single additional term improved the R^2 to 0.828 and reduced the RMSE to 0.0013 or 0.2% of mean value. The actual by predicted plot in Figure 30b is the biggest demonstrator of the improvement. While there is still a fair amount of noise about the diagonal, the severity of the curve is reduced significantly.

Including additional quadratic terms or higher order terms had no significant effect on the quality of the fit. The model for takeoff field length was similarly improved with a single quadratic term and no additional benefit of further quadratic or higher order terms. While it is simple to change the form of the model when regressing from existing data, the inclusion of terms other than linear terms requires a different form of data collection. Other methods for including nonlinear terms are discussed in Section 3.4.4.

3.4.2.3 Results Type C: Good Fits but Poor Matching

With the exception of takeoff field length, which fell into Type B, the linear models for all of the customer requirements fell into a third class. As an exemplar, the results for the operating empty weight are shown here. The fit was excellent, having an R^2 of 0.999 and a RMSE of 41.29 pounds (0.012% of the mean value). The actual by predicted and residual by predicted were similarly near-perfect, passing the first test easily. The plot of correlation coefficients, scaled model coefficients, expert-sourced values, and t ratios for the second test are shown in Figure 31.

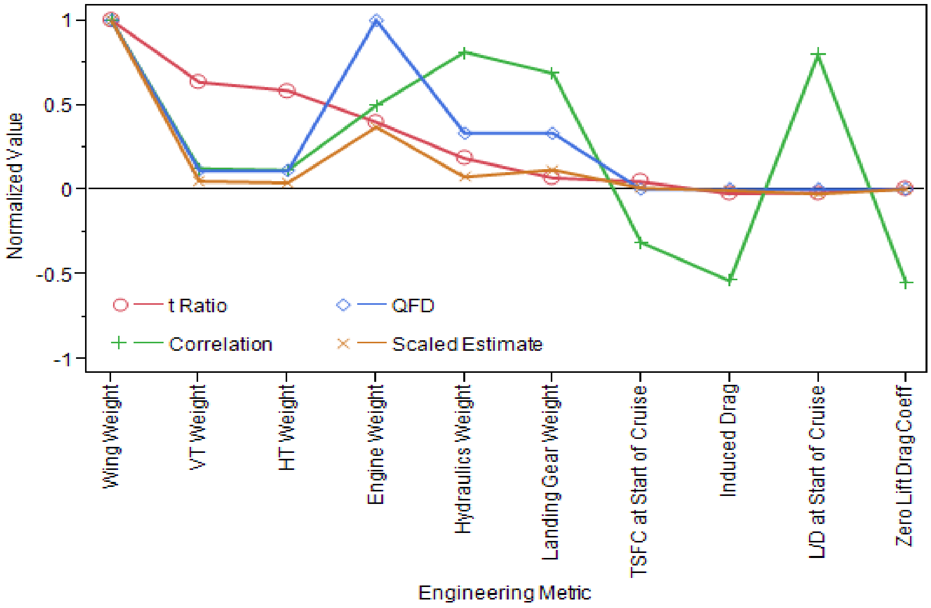


Figure 31: Summary of Normalized Model Coefficients for Operating Empty Weight

The QFD values (indicated with a blue diamond) line up very well with scaled estimates and correlations coefficients for the first three most dominant terms, which is very promising. Unfortunately, the three lines that overlapped so well in the Type A results do not line up as well here. By passing the first test, a linear model is sufficient, but further investigation is needed to explain why the values not lining up as expected.

One of the areas of examination was the correlations between metrics. Normally when regressing equations, great care is taken to make sure that the independent variables are uncorrelated and orthogonal. Even when they are not truly independent and uncorrelated, most design of experiments try to force low correlation between inputs. The values used as an input for this model were observed, not set explicitly and did not have the benefit of a design of experiments setting their ranges or orthogonality.

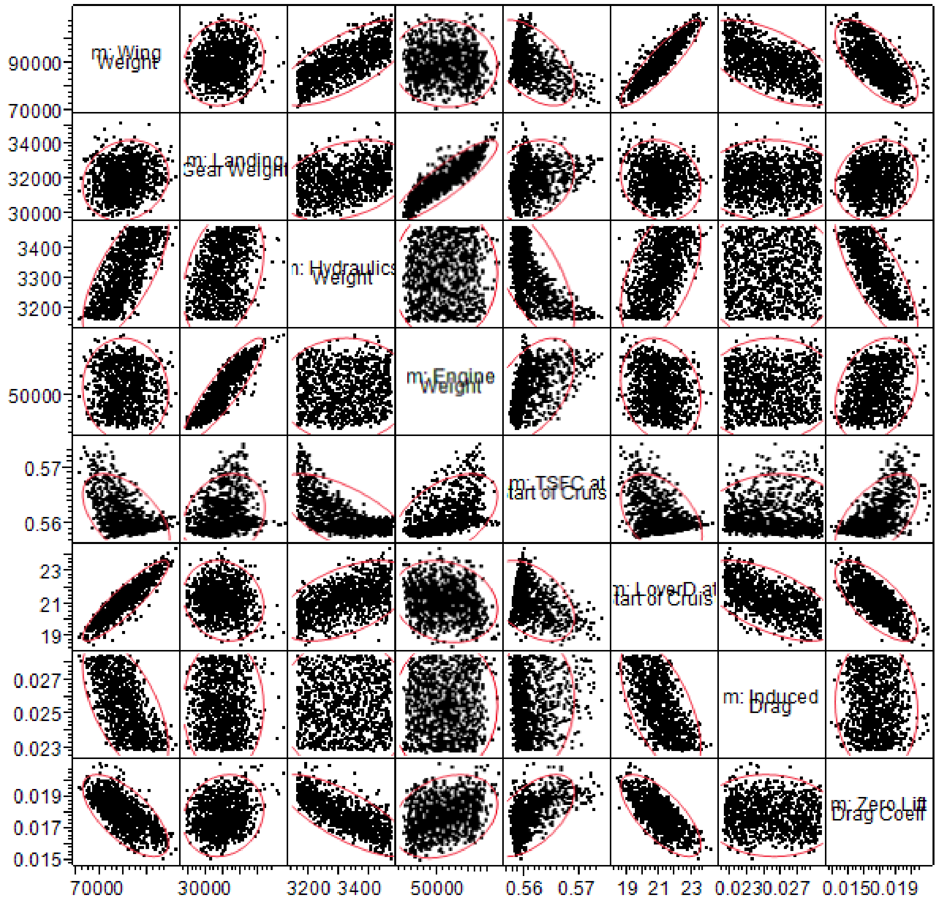


Figure 32: Correlations Between a Subset of Intermediate Metrics

Figure 32 shows a scatterplot matrix of a subset of the intermediate metrics. For entirely orthogonal distributions, each scatter plot would have no clear pattern to them and the density ellipses (red lines) would be more or less circular. This is true for the plots between hydraulics weight and induced drag (row three, column seven). For many of the other plots, this is not the case. This means that there are very strong correlations between the inputs of this model. The cases when the correlations between metrics and requirements were the worst in Figure 31 are for TSFC at start of cruise, induced drag, lift to drag ratio at start of cruise and zero lift drag coefficient. Lift to drag ratio at start of cruise and zero lift drag coefficient are both very tightly correlated with each other and with wing weight. Induced drag and TSFC at start of cruise are also correlated with wing weight, though slightly less so. Since wing weight is the dominating factor in this relationship, these other contributions are confounded and cannot be distinguished from the effect of wing weight.

The implication on the model is smaller than it may seem. This effect is present when fitting intermediate and observed data points, making it difficult to verify hierarchical models individually. Since the model coefficients are not set by regression or by data points, the only driver of concern is whether experts would be confounded in their responses. For the standard QFD approach, the question is usually about the correlation between higher and lower levels. A simple change in language from "What is the correlation?" to "What are the driving factors and how much do they drive the output?" changes that mindset. Thus, though hydraulics weight and wing weight are correlated, an expert would be perfectly able to identify that they both contribute to the operating empty weight individually, and do so in different amounts.

3.4.3 Precision of Relationship Scale

The scale used in a typical QFD to describe the relationships is typically a four-level qualitative scale that may be symbolic or word-based (strong, moderate, weak, none). These scales do not indicate the direction of the relationship. It is intuitive that, for an arbitrary system, some relationships will be positive while others will be negative. Holding all other design factors constant, increasing payload capacity of an aircraft will have a negative relationship with range and a positive relationship on acquisition cost. Keeping the items on the scale constant while adding the direction of the relationship would replace strong, moderate, weak, and none with strong positive, moderate positive, weak positive, none, weak negative, moderate negative, and strong negative. While this is atypical of most QFDs as taught now, direction of relationships has been included in some teachings, such as those from Hauser and Clausing [66]. It has also been used and demonstrated more extensively as part of SP2 and SOAR [75, 107].

Both non-directional and directional scales are associated with a set of numerical values. Since the results of the calculations are generally normalized in some fashion, the specific values are less important than the relationship between them. For convenience, scales can be discussed in terms of integers. The most common scale for QFD uses values of [0, 1, 3, 9]. The values in this scale are a logarithmic interval scale with 3 as the basis. Linear scales are also relatively common, such as [0, 1, 3, 5], [0, 1, 2, 3] and [0, 1, 5, 9]. SOAR uses a scale that focuses on the extremes for some relationships with its [0, 1, 7, 9] scale. The choice of which scale is "correct" is a frequently researched and discussed topic. The most common conclusion in this research is that scale matters, but depends on the problem and is difficult to determine ahead of time. [26, 54, 117]

In most cases, the numeric scales are applied to the qualitative scales *a posteriori*, meaning the participants were not explicitly considering numeric relationships when providing data. In many cases when non-technical individuals are providing information, numeric values may be more of a hinderance than a help. Some introductory engineering design texts, such as Dieter's, prefer numeric values when teaching engineering students to use QFD [43]. For this particular problem, the participants are expected to be technically knowledgeable and skilled. It is expected that they would be very comfortable around numeric ratings. In fact, to be able to accurately produce models, experts need to know if a particular score is 10% or 100% greater than the score below it. It is beneficial for these experts to work entirely with the numerical values.

One attempt to work around the problem of selecting a particular scale is to include a superset of the different scales as Chan and Wu demonstrate[32]. Such an approach essentially produces a linear scale including all integers between zero and nine, but allows experts to self-limit to whichever set of values they feel is most appropriate for the problem at hand.

Therefore, one of the possible improvements, combined with the use of directional relationships, would be to use an integer scale including [-9, -8, -7, . . . , -1, 0, +1, . . . , +7, +8, +9]. As mentioned before, the specific values of the scale are not as important as the relationship and ratios between them. The range is important only in that it also defines the possible resolution of the scale.

With nineteen possible values, the scale is significantly more discretized than the original four-level scales. Any further discretization would begin to be indistinguishable from a continuous range of values. The use of a continuous scale may be more intuitive for individuals with a technical background rather than trying to correlate scores with verbal meanings. This may also allow for smoother transitions and is worth investigating further.

With each increase in discretization and precision of the scale, the implied precision of the resulting model also increases. However, the increase in precision of the experts does not increase at the same rate and may reach its limit before the scale does.

A greater number of possible discrete scores or the possibility of continuous scores makes existing symbolic or clicker-based data collection approaches cumbersome. Using a slide bar along with a graphical representation of the associated slope of each value could aid experts in providing data and will be explored further as a method for the user interface of data collection. A mock-up of this interface is shown in Figure 33. The vertical scaling of the individual plots could be uniform and predetermined. On the other hand, setting all the values to 9 is equivalent to setting all of them to 1 (since it is the ratios that matter and not the specific values). It would be more informative to set the scale of all plots such that the vertical range of each of the lines over all plots sums to the vertical range of the plot area. This would produce identical plots for all values set to 1 and all values set to 9.

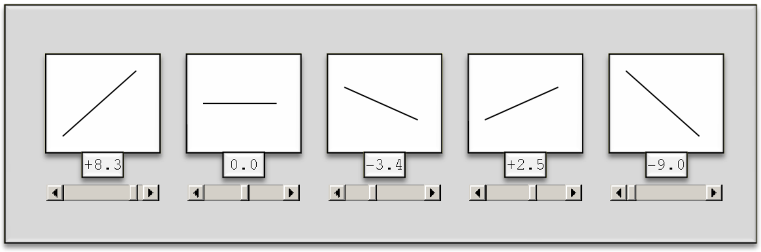


Figure 33: Mockup of Interface for Continuous Variables with Graphical Feedback

3.4.4 Nonlinear Relationships

The regressions of Type B showed the need to occasionally include nonlinear effects and relationships. For demonstration purposes, a quadratic term was used, but this relationship could be any nonlinear form, including higher order polynomial, a piece-wise spline, a neural-network or an arbitrary user-defined form. Offering such a variety of disparate choices has its benefits and its shortcomings. As a benefit, it allows an expert to select whichever form they feel is most appropriate and would likely produce a more accurate model. Unfortunately, this makes it less likely for experts to agree on a particular form and the more complex the form, the more information is required to define it.

A quadratic term was used in the preliminary experiment because it was the simplest nonlinear form and was a one-step upgrade from a linear term. It is also easy to understand intuitively and requires only three points to define it. Since it is just the next term in a Taylor-series polynomial, it is also easy to combine or average multiple quadratic terms with each other or with linear terms into a form that still is logical and easy to manipulate.

The use of a quadratic term should be the exception rather than the rule. As such, a data-collection interface including the possibility of quadratic information should default to a linear form and allow an expert to identify those terms that require a nonlinear term. Once enabled, the nonlinear relationship could be defined graphically with three points as shown in Figure 34.

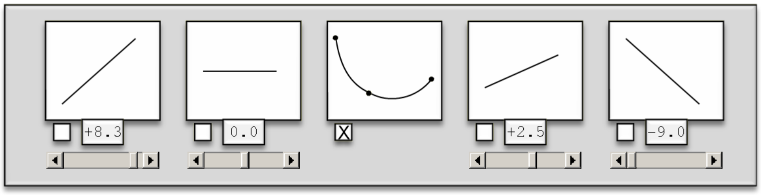


Figure 34: Mockup of Interface with a Quadratic Relationship

3.4.5 Normalization

All of the data collected describing the relationships so far is quantitative, but unitless. The numbers have meaning, but are not the actual slopes of the physical trends. The scores themselves are relative to each other. If an expert says that every input has a strong effect on the output, it is impossible to differentiate the effects of one input from another. The same is true if all relationships are weak. For each row in the data collection interface, the total variability is constant. With the existing information, all comparisons between design points are on a scale relative to the other design points and defined by the points at the extremes of each system requirement and intermediate metric. This means that the relative range of an output will be the same whether all the relationships are given a score of 1 or all are given a score of 9. The points that will be at the extreme in one case will be the same points for the other. And since the model is linear, the distribution of points along that slope will be identical as well.

When translating the outputs of the models for the intermediate metrics to the inputs of the models for the system requirements, differences in the numerical ranges of each term can have a large detrimental effect. Experts shouldn't need to worry about using a proper scaling factor in a subjective opinion-based interface such as the ones used here. Therefore it is necessary to normalize the output of each row to maintain a consistent numerical range for all inputs and all outputs. These ranges can be translated to ranges with physical meaning in a later step as discussed in Section 3.6.1.

The question is how to normalize these scores. In a flow-down QFD, the scores must be normalized across the entire relationship matrix since the values of the prioritizations between outputs is important. The same is not true here and such a normalization would not produce the similar scales desired. Instead it must be done row-by-row. The first step is to code the design variables according to a -1 to 1 scale so that the minimum possible value of each design variable -1 and the maximum is 1. This method is commonly used in building designs of experiments. If all the relationships are purely linear, if the sum of the slopes is equal to 1, the combination of all possible extremes will produce an output that is coded on the same scale from -1 to 1. This makes translating between levels easy and any calculations are independent of which level is currently being investigated. Thus, to normalize the relationships M_{ij} in the ith row which has only linear relationships, the following equation is used.

$$M_{ij,norm} = \frac{M_{ij,raw}}{\sum_{j=1}^{m} M_{ij,raw}} \tag{13}$$

Quadratic relationships do not follow the same trend for two reasons. First, the minimum and maximum output values of a particular relationship are not necessarily at the far edges of the range of inputs. Second, since experts were allowed more control over the size and shape of the curve, the curves may no longer be centered at the same location vertically. In this case, the row is normalized by finding the maximum and minimum values for each individual relationship. Since each relationship can only either be linear or quadratic, there are closed-form solutions to find the minimum and maximum values in all situations. The minimums and maximums are summed respectively and an intercept term is added such that the maximum and minimum values are equidistant from the center point (defined at zero in this coding system). The intercept, slopes, and quadratic terms are then divided by the value of the maximum above zero to force it back to the -1 to 1 range.

3.4.6 Combining Responses

The model-creation discussion so far has only assumed a single curve. This curve is the result of a combination of responses from multiple experts. Section 3.3.4 eliminated the choice of a group developing a single result by means of collaboration and general agreement. Other common approaches of combining expert information are some form of weighted average of the individual pieces of information collected. For this research that information consists of the values for the relationships within the QFD or the coefficients of the linear model. Since these relationships have been normalized according to the previous section, there is no need to maintain different units during the combinations.

The question then arises as to how to determine the weightings. Cooke identifies four qualities necessary to build a rational consensus: reproducibility, accountability, empirical control, and fairness [35]. The weightings might be based on years of experience, pay grade, self-assessment, peer assessment or number of citations. Experiments using such measures have shown that they are not consistently and reliably valid [36]. It is also possible to test each expert's knowledge by using a set of baseline questions with known answers. The more accurate an expert is for the known data, the higher their information is weighted for the unknown data [6].

Since the values or coefficients have geometric meaning in describing a line or curve, methods for combining responses based on geometric combinations developed for averaging arbitrary surfaces and curves are also available. These methods discretize and parameterize the curves so that the form of the curve is irrelevant to combining them. They then average the displacement angles from point to point, take the geometric mean between a series of curves, or employ one of a number of other geometric combination methods. [116] These methods can be extremely powerful, but are also computationally intensive and would not make use of the knowledge of the form of the relationships that we already know (that they are limited to linear or perhaps quadratic equations over a given range and magnitude).

Following the trend set previously for this research, the simplest approach is preferred. Since experience, pay-grade and publications have been shown to be unreliable metrics of expertise or knowledge, there is little reason to consider them as sources for weightings.

Self-assessments or experts identifying their own certainty of their information may be used to capture the certainty of the results, but could not be used both for certainty and for weightings of averaging. Of the two, a measure of certainty of results is more useful. There is no guarantee that outside baseline information exists or that it is a similar enough problem to aid in testing the quality of the experts. Even if it is, if sufficient information exists to establish a baseline, expert-based modeling may no longer be necessary. The use of geometric methods is overkill and may contribute additional uncertainty to the results as a result of increase complexity.

This would seem to eliminate all of the available options, but the simplest approach is to use an unweighted average and treat all experts equally. If an expert is qualified enough to give information according to the discussion in Section 3.3.3, their information should be trusted until there is reason not to. Such a case may occur if the information from all experts but one agree within a very tight band. The information from that one expert is either the correct information or, more likely, less trustworthy.

3.5 Test the Model

When developing a system, the system is tested in some fashion before put into production use. A model must also be tested before it is put into use. The steps in testing a model include verification, validation, and in some communities, accreditation. Verification describes the process of ensuring that the model behaves logically and according to the designer's intent. It tests that the outputs appear reasonable for the problem and for the inputs given. Validation is the "process of building an acceptable level of confidence that an inference about a [similar] process is a correct or valid inference for the actual process" [93]. This is typically done by testing a model against a reliable outside source that serves as "truth data". Accreditation, when performed, is the process of a particular authority certifying that the model can be used for certain types of problems. Accreditation is discussed in Section 3.6. The steps are commonly abbreviated at V&V or VV&A. Balci argues that, in order to ensure that the model is consistent and reliable throughout, VV&A must occur throughout the modeling life cycle [8].

3.5.1 Verification

For the expert-based model created with this research, verification happens at the same time the experts are giving information. This is especially true when the data collection interface includes the graphical display of the slopes of each term. Each expert has an instantaneous visualization of the behavior and relationships he or she is describing. If the expert disagrees with what is displayed, any corrections or adjustments can be made immediately.

Experts can perform an additional verification of the model after the relationships have been combined as well. There are two approaches here. One shows each expert their own relationships relative to the consensus relationships and allows an opportunity to adjust it. The danger with this approach is that many experts will adjust their curves to join in with the consensus whether or not the group answer is more correct than their own. This artificially reduces the uncertainty associated with the model without necessarily improving it's quality.

A second approach is to allow each expert to look at only the consensus answer and offer an adjustment from that to a corrected value. If each expert does this alone, each expert has the benefit of information from others without the pressure to agree with it. It also has the benefit of allowing an expert to provide data a second time (and faster) since it is unlikely that individuals will remember the values they supplied for every relationship. The combination of the adjustments provides an additional set of data to analyze. If most experts adjust the relationships in the same direction, the final result has greater confidence. If experts instead each adjust it back to their original positions, then the resulting model is no worse off than it was prior to the adjustments.

3.5.2 Validation

Validation is typically performed against existing, trusted data. This data should span as much of the range of the model as possible. If sufficient data exists to test the ranges of the model, then it would be better to use that data to create a low-fidelity model rather than an expert-based approach. The motivation for this research assumed that historical data or higher fidelity models are not available and that assumption does not change when it comes to validation. Referring back to the definition of Mihram's definition of validation, the important aspect is the building of confidence in the resulting product. Comparing against existing data is only one method for that. Banks et al. state that the basis of model validation falls to face validity, historical methods, and sensitivity analysis [9].

Another option is to validate a model by expert inspection. The Department of Defense identifies this as an acceptable approach, though it prefers other methods if available [42]. Davis includes expert opinion in his hierarchy of VV&A and identifies that validation establishes that the model is "doing the best we can — or, at least, something that is 'good enough"'[40]. He goes on to detail how experts can perform many of the same tests that would be performed by using historical data, using field-test data, testing analytic rigor, and evaluating the economy of the model. For each of these cases, he implies that an outside expert is providing the validation. For this model development problem, having an independent party validate the model would require additional time that may not be available. Also, if the individual is qualified to validate the model, it is likely that he or she is also qualified to contribute to it and should be included in the group of experts giving information.

When creating surrogate models as part of the response surface methodology, a portion of the source data is held back to calculate the model representation error. In cases where there is limited data, different portions of the data may be withheld from the data used to fit the model and each resulting model is compared. This is a form of iterative co-validation. Wright and Bauer defend a similar approach for using two different types of unvalidated models to validate each other [132]. The full set of information from a single expert creates a model. Combining these approaches results in iterative co-validation of each expert's contribution by comparing it to the combination of the remaining experts.

The more each of the models agrees with the others, the more certain the result is. Some terms may agree better than others and it is useful to identify the terms with the greatest agreement and those with the worst. It is also important to evaluate the total amount of uncertainty resulting from combining terms, each with their own uncertainty. Measuring and displaying the certainty of the resulting model can be performed in several ways. The easiest is not to do so. This may correspond to a pass/fail grade and the model is only made available if it passes. Obviously this does not generate much confidence in the model or the process and may also result in the model being used beyond its applicability.

Another option is to show the standard deviation of each of the terms, or a combined standard deviation of all the terms for a given output. This has the benefit of providing a single summary statistic that can be referenced quickly. A third alternative is to graphically and numerically show a confidence interval for each model based on the distribution of the information from the experts.

Since all three choices are a matter of post-processing and displaying the result, all three choices can be included with minimal extra effort. The limitation would be on the portability of the results to other programs and formats.

Recall the feedback loop from Test the Model to Gather Information in Figure 20. If the resulting model has standard deviations or confidence intervals that are too large for the particular problem, additional information must be used to further define the problem or by including more experts. If no further experts are available, the problem may be too complex or vague for expert modeling.

3.5.3 Measuring Agreement

In order to perform the co-validation, one needs a way to measure how well the group of experts agrees with each other. The behavioral sciences fields have a similar need to compare how individuals think based on their preferences and scorings of things they observe. For the most part these fields use the term *raters* for individuals providing a score, or rating, of an observed event or opinion. The class of measures are known as inter-rater reliability metrics. Unfortunately, many of these measures have been developed for specific purposes that makes them ill-suited outside of behavioral sciences and this problem in particular.

Two of the most promising measures for measuring agreement for this research are Pearson's product-moment correlation coefficient and Krippendorff's alpha.

3.5.3.1 Pearson's Product-Moment Correlation Coefficient

Pearson's product-moment correlation coefficient ρ_{xy} is by far the best known measure of inter-rater reliability and similarity. To this end, most times when an individual uses the term *correlation*, they are referring to this value. Intuitively, it refers to how well the relationship between two variables can be described as linear. This makes it a poor measure when trying to identify the existence of other trends, but for this research the ideal relationship would be a perfect 1:1 matching in values. Unfortunately, it is also very sensitive to individual outliers which may have less of an effect on the final result, especially if individual scores are removed. ρ_{xy} is calculated as shown in Equation (14) where X_i and Y_i represent individual values, $\overline{X}$ and $\overline{Y}$ are the sample means, and s_X and s_Y are the sample standard deviations.

$$\rho_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_X} \right) \left(\frac{Y_i - \overline{Y}}{s_Y} \right) \tag{14}$$

Like most other measures of inter-rater reliability, it is only able to compare two individuals or two data sets at a time. For this research, it is important to be able to identify the degree of agreement within a group of individuals. When more than two sets of data are included, the average of all the pairwise comparisons will be used. While not a perfect measure, it is fast to calculate and intuitive to understand. This pairwise comparison is useful in identifying clusterings of agreement or disagreement among certain groups within the sample population.

3.5.3.2 Krippendorff's Alpha

Thus far, the only metric that can handle more than a pair of raters and continuous interval-based metrics is Krippendorff's alpha α_k. That said, it is not an ideal measure. Like most others measures of inter-rater reliability, it performs's best when working with coded integer or binary data rather than continuous scores. It describes the deviation from perfect reliability by the ratio of observed disagreement D_o to the expected disagreement D_e according to Equation (15) [77].

$$\alpha_k = 1 - \frac{D_o}{D_e} \tag{15}$$

The actual definitions of how the observed and expected disagreements are calculated are quite involved and the reader is encouraged to refer to the references for the full explanation. This complex definition has a number of disadvantages. The first is that the true physical meaning of the value is far from intuitive. Krippendorff tries to explain it as "the degree to which independent and interchangeable observers respond to given, but unknown phenomena identically rather than randomly" [80].

The other disadvantage of the complex definition is that it is not widely implemented. In fact, it was little used at all between 1970 when it was first published until Krippendorff published a series of implementations for various software packages over the last decade [78, 79]. All of these implementations are heavily iterative and do not scale well with large numbers of raters or larger numbers of scores for each rater. The speed is further hampered when values are dissimilar. It shares the same problem that it does not handle outliers well, so if an individual disagrees significantly on one score, the reliability of the entire group drops more than one would expect.

3.6 Use the Model

Some organizations, most notably the Department of Defense, require a model to be accredited for a particular use. An accreditation is a certification that the model has been tested and proven to be valid for a particular problem with certain ranges and for certain types of outputs and that its results are reliable. When working with multiple groups and sharing models between them, models should be accredited or "blessed" by the group most qualified for that particular model. For example, if an integrated design environment includes an engine model, the propulsion group should be responsible for accrediting that model.

The expert-based models created here are more general and less-specific. They already have some small amount of blessing from those experts who provided the data to create them. However, there is no evidence to clearly identify how the models can be used. While a truth model would not exist when creating an expert-based model independent of this research, having one available now makes it possible to test the accuracy of the results. One way to test the accuracy of the expert-based model would be similar to the preliminary experiment in Section 3.4.2 by reducing the truth data to the same form and comparing the slopes. While this is a helpful measurement, it does not identify what types of results are valid.

Instead, consider three levels of reliability of results for "accreditation". Start by sampling the truth model across its ranges with a moderate number of points (say around 50-100). Sample these same points using the expert-based model. Using an arbitrary overall evaluation criterion, rank these points in some order. To meet the lowest level of accreditation, the points from the expert-based model should match the order of the points from the truth model. Being accredited at this level means that the model could be used to correctly identify regions of space that are of interest or that are likely to give the optimal performance. The model can be used to prioritize design concepts and the design space. This level is useful since it would allow a model to show that some design points are clearly better than others, even if it is unknown how much better one design point is than another.

This first level is best calculated using Spearman's rank correlation coefficient r_s. This measure, originally developed by Spearman in 1904, focuses on comparing how similar the monotonic ordering of two sets of data are [120]. This is different than Pearson's correlation coefficient which also takes into account the ratios of the values. As an example to show the difference, refer to Figure 35. There is a clear trend between the two arbitrary variables, but it does not fall along the linear diagonal as one desires for the Pearson coefficient. The value of the Spearman's rank correlation coefficient is 0.9618 compared to a lower value of 0.9110 for Pearson's correlation coefficient.

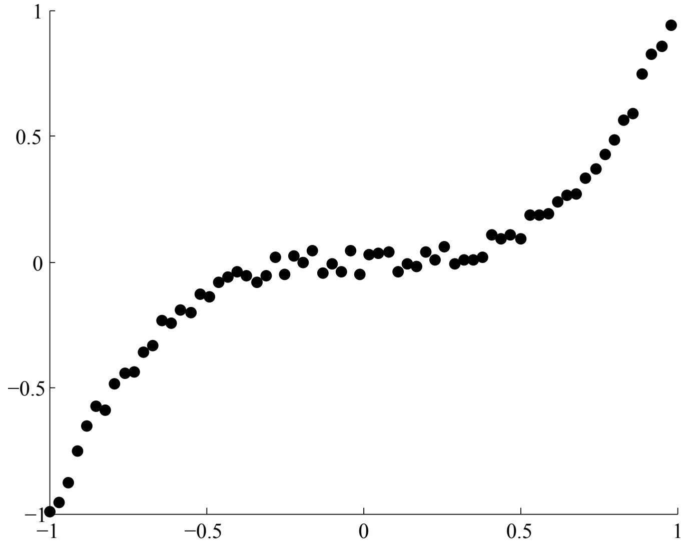


Figure 35: Example Data for Spearman Correlation

Like Pearson's coefficient, Spearman's correlation coefficient is limited to comparisons of two variables at a time. It is calculated by sorting both sets of data according to one variable and finding the integer position in ascending order. When two values are identical, they are both given the average of the ranks they split. For example if, out of four ordered values, the two in the middle are identical, they would both be assigned to rank 2.5. r_s is defined by Equation (16).

$$r_s = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \tag{16}$$

By defining d_i as the difference in ranks between the two variables of a single point, this simplifies down into Equation (17).

$$r_s = 1 - \frac{6 \sum d_i^2}{n\,(n^2 - 1)} \tag{17}$$

For the second level, normalize the values of the outputs of both models (again using the same inputs for both). To meet the second level of accreditation the normalized values for both sets of data should be equivalent or very close to it. If the expert-based model qualifies at this level, it can be used to directly compare results and identify the margin of difference between specific concepts and designs. This is more useful than a ranking because it can help to identify the relative improvement between designs and trade those changes with the effort required to meet the design point.

For this, the distance between the points in both variables should be the same after being normalized. An ideal relationship would fall along a linear diagonal in an actual-by-predicted plot. As discussed previously, this makes Pearson's product-moment correlation coefficient an ideal solution. Another possible measure would be the coefficient of determination (R^2). Pearson's is more intuitive and easier to calculate quickly, so it is the preferred choice here.

The third level of accreditation compares the non-normalized outputs of both models against each other. If the outputs are very similar, then the expert-based model can be used to predict the true performance of design points and identify how well they meet hard requirements. The previous two levels can be achieved with the existing information that has been collected, but the third level requires additional information about the intercept and ranges of both inputs and outputs at each level.

3.6.1 The Need for Units and Ranges

As part of the Define the Problem step, some information about the ranges of the design space were identified. That step defined rough scale of the system and defined the concept sufficient to have all experts on the same page. At the lowest levels of inputs, the system design variables, experts certainly have a rough idea of the ranges of values they are providing relationships for and the corresponding units.

The ranges and associated units of the next level up, the intermediate metrics are known with less certainty. Experts should still have an idea of possible values that may result from the design variables, but the complexity makes it more difficult to be precise. The ranges of the customer requirements have the most uncertainty.

Additionally, the ranges at the higher levels depend on the ranges of the lower levels. If the ranges for the design variables are too far off, there is little chance of the intermediate metrics or customer requirements being correct. The reverse is not true. Even if all the ranges are incorrect, there is a chance that the relationships may still be correct and allow for accreditation at the first two levels.

To capture the ranges and units, experts will be given the opportunity to set a minimum and maximum and identify the corresponding units. If an expert is not able to give these values or is not confident about them for any level, the rest of the relationship data will still be used.

3.6.2 The Resulting Model

The ultimate output of this method is a set of models, one for each intermediate metric in terms of the design variables and each system requirement in terms of the intermediate metrics. These can be combined to produce a set of models for each system requirement in terms of the design variables. By default, these will be non-dimensional models that give scores for each output over the range from -1 to $+1$ and may either use non-dimensional normalized or dimensional inputs. With the addition a consensus on the ranges for the units of the outputs, the models can produce dimensional results. Since the mathematical form of each relationship is a simple linear or quadratic, the resulting model will also be linear or quadratic. If the two levels of models are combined together, the result will be a polynomial network, which can be simplified to a polynomial for portability.

On top of the resulting models, the information used to create it is available to provide insight into the level of certainty and consensus among the experts. While consensus does not guarantee accuracy, a perfect consensus would tend to be more accurate by means of co-validation. Depending on the tool used to collect the information from experts, this may be included along with the model itself or reported and distributed separately. The simplicity of the resulting model form encourages distribution in plain-text or other readily translatable forms without metadata.

3.7 Summary of Method Development

Figure 36 shows a methodological morphological matrix describing the different degrees of freedom that have been identified. This structure will be used to summarize the proposed method for development of a model based solely on expert-sourced information. The far left column lists the phase of model development. The next column describes a particular aspect or decision of the method that has been discussed and the third column is the section or sections where it is discussed. The remaining columns list the alternatives that were identified for that particular aspect. In cases where alternatives were eliminated, these alternatives have been crossed out. The total number of possible combinations identified in this approach is approximately

$$5 \times 307,200 \times 240 \times 5 \times 8 = 2,949,120,000 \tag{18}$$

possible combinations. The values on the left hand side of Equation (18) are the number of options in each larger step of the process. A great deal of other alternatives were eliminated earlier prior to the point of including them here.

For the *Define the Problem* step, concept definition is the primary step associated with it. Several options were investigated. The first preliminary experiment shows that the concept definition is essential to providing accurate information, but the option of not performing it is included as a baseline worst case scenario. Either of the *ad hoc* approaches have the potential to define the concept, but no guarantee since these methods would not be consistent from one effort to another or from one group to another. If the effort of performing morphological analysis is going to be included, the benefits associated with an IRMA outweigh the slight improvement in speed with vanilla morphological analysis. Especially since it is assumed that either the morphological matrix or the IRMA exist prior to beginning the modeling exercise.

The gather information step includes those improvements necessary from the second and third preliminary experiments. The first decision to use expert-based information as the source was one of the driving factors for this research, but that decision is still important to document here. The number of experts included could range from around 5 to around 40. It is highly unlikely that any group would have 40 experts available on a moments notice, so that option is eliminated. The levels of experience correlate to the job-levels. The lack of availability of technical fellows (equivalent of professors in an educational institution) in the numbers identified in the previous row prevent them from being a viable minimum requirement.

For the data collection setting, a group setting would take far too long to schedule and complicates the experimentation while also allowing individual experts to dominate the discussion. Thus, individual data collection is the remaining option. For the scale, literature has already shown that engineers are capable of giving numerical rather than linguistic scores. The use of continuous scales and graphical slopes will be combined into

Step	Aspect	Sec.	Options				
Define the Problem	Concept Defintion	3.2.2	~~None~~	~~*Ad Hoc* Discussion~~	~~*Ad Hoc* Document~~	~~Morphological Analysis~~	IRMA
Gather Information	Source of Information	1.2 1.3	~~Historical Data~~	~~Physical Experiments~~	~~Finite Element Analysis~~	Expert Information	~~Theory~~
	Number of Experts	3.3.2	~5	~10	~20	~~>40~~	
	Level of Expertise	3.3.3	New Hire	Post-Grad	Specialist	Experienced	~~Technical Fellow~~
	Data Collection Setting	3.3.4	~~Group Meeting~~	Individually			
	Signed Values	3.4.3	~~Positive Only~~	Positive and Negative			
	Scale	3.4.3	~~Low, Med, High~~	0,1,3,9	0,1,2,3,...9	0-9 Continuous	Graphical Slopes
	Nonlinear Type	3.4.4	None	Quadratic	~~Higher Order Polynomial~~	~~Piecewise Spline~~	~~Arbitrary Curve~~
	Bounds and Units	3.6.1	~~None~~	Design Variables	Design Variables & Performance Metrics	Design Variables & Perf. Metrics & System Requirements	
Create the Model	Baseline Model	2.2.4 2.2.5 2.2.6	Quality Function Deployment	~~Strategic Portfolio Planning~~	~~Portfolio Analysis Tool~~		
	Model Equation Form	3.4.1	Additive Function	~~Multiplicitive Function~~	~~Differential~~	~~Arbitrary Expert-Defined~~	~~Feed-forward Neural Network~~
	Types of Terms	3.4.1 3.4.4	Linear	Quadratic	~~Higher Order Polynomial~~	~~Expert-defined~~	
	Combine Individual Responses	3.4.6	Unweighted Average	~~Weighted Average~~	~~Cooke Method~~	~~Geometric Method~~	
Test the Model	Validation Data	3.5	~~No Validation~~	~~Historical Data~~	~~Physical Experiments~~	~~Higher Fidelity Models~~	Expert Co-Validation
Use the Model	Output Applicability	3.6	Correct Rankings	Correct Relative Values	Correct Absolute Values		

Figure 36: Methodological Matrix of Alternatives

a single option for testing purposes. The remaining options are present to identify the minimum level of granularity required. The third preliminary experiment identified the need for nonlinear data on occasion. While there are several options, the higher level the option, the more difficult it is to define and the less likely experts would agree upon the form or the trend. Lastly in this section, the levels where the bounds and units are applied could start from none at all to adding each level to those available. Since it is unknown what level experts will feel comfortable answering information for, all levels are included and the maximum level will be determined by experts during experiments.

The create the model step starts with the selection of the baseline model form. QFD, SP2/SOAR and PAT were all considered, but QFD was the most promising due to simplicity and extensibility. The next row discusses the model form using information from that layout. The use of a simple linearly additive form was preferred again due to simplicity and because of the past use in QFD and the demonstration of its applicability in SP2/SOAR. The third row in this step lists several types of terms that could be used within an additive model. The previous decision to consider quadratic terms is mirrored here in addition to the standard linear terms. Higher order and arbitrary expert-defined terms are not included due to the difficulty of matching those with multiple experts and the increased level of complexity associated with them. Finally, the method for combining data from different experts examined several options. While the weighted average, Cooke, or geometric methods would certainly give better combinations of data, they all require information that may not be available in a short timeframe.

Once a model is created, it needs to be validated in some fashion. No validation is an option, but is not a well-trusted one. The other methods require access to information that would also likely supplant the need to create an expert-based method. The final option of expert co-validation provides sufficient assurance that the information is correct without a significant burden to gather additional information. The certainty of the resulting model could be displayed any of three ways. Since these are all relatively simple calculations, all three can be offered and left up to the user of the model.

The last step is to use the model. The particular use of the model is not for this research to determine. However, this research can identify what detail/precision of information a typical model can produce. Experimentation is necessary to determine the highest level of accreditation.

CHAPTER IV

RESEARCH QUESTIONS AND HYPOTHESES

As part of developing the method, a number of problems and possible approaches as alternatives were identified for each step. Each of these problems and set of alternatives was one in a chain of research questions that started with “How can available expertise be leveraged to quickly provide analytical models?” through “How should this step be performed?” to the specific needs of each of those steps. Each of those problems was answered by referring to existing literature, by analysis, or by a preliminary experiment. Several options remain and require a full-scale experiment to identify which is appropriate and optimal.

Each of these remaining options can be formalized as a research question. These questions mirror those that were asked during the development of the method for each step.

> **Research Question 1:** Which type of interface and scale should be used to collect useful information?
>
> **Research Question 2:** What level of expertise is required or sufficient to provide useful information?
>
> **Research Question 3:** How many experts should participate when using this method?

As with many problems in engineering, there are multiple measures that could drive each of the possible solutions. For example, including more experts would likely increase the diversity of opinions but also increase the cost and effort required. The best solution depends on which of those attributes is most important at present. Some intuition about the possible correct answers to each of these is based on the literature review and preliminary experiments presented in Chapter 3.

The fundamental question associated with the development of any new method or process is “Does the proposed method work?” or “How well does the proposed method meet its goals?”. These are not what would be expected of research questions but rather they are tests of the goodness of the method. In order to test these fundamental questions experimentally, the must be adjusted into measurable tests. To test the goodness of this particular method, three tests are used to evaluate different aspects of the method and the experts.

> **Test 1:** How well do experts agree with each other in their estimation of the relationships?

The first test is whether or not experts agree with each other enough to instill confidence in the results. If all, or most, experts agree with each other in their estimates, then the co-validation discussed in Section 3.5 can be completed. On the other hand, if they do not agree, it is necessary to identify which experts are correct and which are not. This process would require additional information about the experts or the system being modeled.

> **Test 2:** How well do the relationships provided by experts match the true relationships?

The next test is whether or not the values of the relationships match the partial derivatives of the true relationships. Two different models compared against each other should have similar partial derivatives if they are capturing the same behavior. This test mirrors the analysis performed in the second preliminary experiment in Section 3.3.1 on page 106 with more detailed and adds further analysis. It is possible for individual experts to be incorrect and their combined values to match up if some overestimate and others underestimate. Because of this, it is possible for a group of experts to perform poorly in the first test and still do well in this test.

> **Test 3:** How well do the outputs of the expert-based model match true behavior?

The final test compares the set of outputs of the expert-based models for a given set of inputs against the outputs from a trusted analysis for the same set of inputs. It is this test that is typically used for model validation. In this case, there are multiple levels of passing, as discussed in Section 3.6. As with the previous test, it is possible, though unlikely, for a set of experts to perform poorly in the previous tests while the resulting model is still capable of matching the output of the truth model and still produce a model capable of matching outputs.

These questions test whether or not the method works, providing a score for each of the three measures of success identified. The options explored by the research questions are each As identified at the end of the last chapter, there are several methodological options which have not yet been determined. It is expected that changing those has an impact on the answers to those three questions.

How do these tests vary with

1. the interface and scale that experts use when providing data?
2. the level of expertise of the individuals?
3. the number of experts included in the model?

The following chapter describes the experiment performed to generate the data necessary to answer the tests with those variables in mind. Rather than vary each of those options separately and individually, the first two will be varied together and distribute experts between them. The third will be varied by including different numbers of experts during the analysis process.

4.1 Scale and Interface Effects

There are three interfaces and scales that have been identified. The first is the standard reduced integer QFD-type scale [-9 , -3, -1, 0, 1, 3, 9] serving as the baseline approach. The second expands this to all integers between -9 and 9 to allow experts more flexibility in the scores they are able to provide. The third uses the same set of integers but adds a graphical display of the relationships and also allows experts to quickly and intuitively view the relative values of the data they are giving. Moving from the first scale to the third adds precision and the ability to add additional types of information but at the cost of additional choices and requiring more thought, as discussed in Section 3.4.3.

> **Hypothesis 1a:** Because of the reduced set of possible values, there will be a greater degree of agreement among experts who use the QFD scale than among experts who use either of the other scales.
>
> **Hypothesis 1b:** The greater degree of discretization and information presented to the participants provided by the graphical display will allow experts to more accurately capture their intended relationships as they believe.

4.2 Education and Experience Effects

Greater experience and expertise should lead individuals to a more correct understanding of the problem. If each individual is more correct and closer to the true relationships, then they should each share the same correct assessment and agree with each other as well. At some point, variation in opinion and between individuals will have more of an effect than the incremental improvement in knowledge and experience. Some minimum level of expertise is necessary to even understand the terms and variables included in the model but many of the relationships are a matter of fundamental knowledge in the field. Having a firm grasp of the fundamental knowledge and sufficient experience to understand the design space would seem to be the key drivers to being able to answer correctly. There are other ways to measure the level of expertise discussed in Sections 2.1 and 3.3.3, but to manage privacy limitations, research ethics, and testability, years of experience will be the primary measure used here.

> **Hypothesis 2:** Individuals with more than two years of experience with the relevant vehicle type who have demonstrated a general knowledge of aircraft design concepts by passing the doctoral qualifiers will perform noticeably better than those who have not. Individuals beyond this mark will perform similarly.

4.3 Number of Experts

The resources required to perform a modeling exercise can have a great impact on how likely it is to be used. For this research, the single greatest cost is the time of the experts who provide information. The number included must be large enough to include enough viewpoints raw data that the results can be trusted. If it is too large, the cost of the model would prohibit the method from ever being used. There is also an upper bound where increasing the number of participants no longer improves the accuracy of the model. Considering the minimum number of individuals that other expert-based methods suggest, this number for a problem of the size of a performance model is likely to be around five to six. For larger and more complex problems, such as those commonly addressed with SP2/SOAR, this number would have to be larger. The details of the implications of the number of experts are presented in Section 3.3.2.

> **Hypothesis 3:** Increasing the number of experts whose data is included in a model past five to six will not have a significant impact on the accuracy of the combined model.

There is one large caveat with this hypothesis in that the skill and expertise of the individuals included will have a larger impact than the number. A model based on data from 200 untrained and inexperienced people will not be better than a model based on only two highly skilled individuals.

CHAPTER V

DEMONSTRATION OF THE METHOD

In order to test the method developed in Chapter 3 and examine the questions and hypotheses presented in Chapter 4, the method must be performed as an experiment with individuals providing the relationships and data necessary to create a model. This chapter describes the specific implementation of the method that was used to perform this experiment. In some situations, new information was gained about the method itself during the experiment, while other conclusions were reached upon analysis of the results, as discussed in Chapter 6.

The experiment described here deviates slightly from how this method would be applied in a production setting due to the research-nature. Analogously, a prototype aircraft will have additional sensors over a production version and would be flown differently than it would be flown in commercial service. The differences are necessary to gather more information than would be needed or available typically as well as to account for the differences in resources that are available within an educational institution. Some variations are due to the ethical and legal limitations associated with any research involving human subjects.

This chapter will follow the same order of steps as the modeling process itself, starting with *Define the Problem*.

5.1 Define the Problem

Typically, the motivation for creating a model is to provide analysis to support a problem that has been given from an outside source. That source may be a customer or another group within the organization. In this case, the motivation for creating a model is to test the method. Therefore, the design problem and scenario are selected to support that goal.

And the required fidelity of the resulting model is to match an existing truth model as best as possible.

5.1.1 Selection of a Test Problem

Typically a problem would exist that required using this modeling process. The process is a just tool to solve the problem at hand. For this research, the problem is that the process itself needs to be tested and a suitable design problem is needed to test it. This problem must meet several requirements. The general form of the model, as described in Section 3.4.1, is a linear equation with the potential for a small number of quadratic terms. A suitable test problem should be adequately described with such a model. While it may appear to be rigging the solution, this research is testing the ability of experts to create the model more than the ability for a problem to be modeled with a particular form. Selecting a problem that would not be possible to model in this form would be a poor test of the method.

An ideal test problem is also well-understood by or easily explained to a large proportion of the aerospace engineering community. Readers should be able to focus on understanding the method being applied to the sample problem rather than the problem itself. As a corollary, the problem should not have any “interesting” regions where significantly deeper understanding is necessary to explain the true behavior of the problem. This eliminates abstract problems or hypothetical scenarios that are unlikely to be seen in actuality.

This method depends on having experts available to provide information. It should be relatively easy to find and gain access to experts who are qualified to participate. Since participants should also have a basic understanding of QFD, experts are most likely to be found in specialties that incorporate it as part of the education and research. While the method is intended to be used with limited interaction and would even be possible without physical co-location, additional contact is helpful during the development and testing aspects. This suggests that an aerospace problem should be well-known and well-used enough that experts are present within the Institute.

Last, since this research tests the accuracy of the expert-based model produced as part of this experiment, a good sample problem should have trusted higher fidelity models (or surrogates of them) readily available and easy to use. A difficult to use truth model increases the chance of error in the portion of the experiments that should be error-free. If a new model needs to be created and validated, there is a greater chance of errors coming from the truth model than with one that has been used and matured previously. Any truth model used should also be deterministic since additional error inherent in the source information would confound the uncertainty of the expert-based model. It should also be capable of narrowing the inputs required to be varied to a small number, hopefully similar to the independent variables used in the expert-based model. This last requirement eliminates any problems that involve defining complex geometry.

The civil airliner problems used in the preliminary experiments meet all of these qualifications. While several problems were identified with the choice of variables in the preliminary experiments, an intelligent selection of the variables used here will reduce the impact of those. There are several truth models available, the most accessible being NASA Langley's Flight Optimization System (FLOPS) aircraft performance and mission modeling code [92]. FLOPS is frequently used as part of academic designs problems as well professional research as an accurate and reliable method for predicting aircraft performance [89, 76]. On top of available models, there are numerous production aircraft with well-documented design and performance metrics available for comparison and baselining. Over small design ranges, the physics and trends are well-behaved and well-understood.

There are two disadvantages of using a civil airliner test problem. First, it is a perfect example of a problem where sufficient models already exist. It is unlikely that a civil airliner would ever need an expert-based model to be built for it. While this doesn't affect its applicability, it does take away from the obvious motivation of performing this method. The second is that civil airliners, like most aircraft classes, are highly interdependent and not easily decomposed into multiple levels and independent variables. This issue would be true of most aerospace design problems.

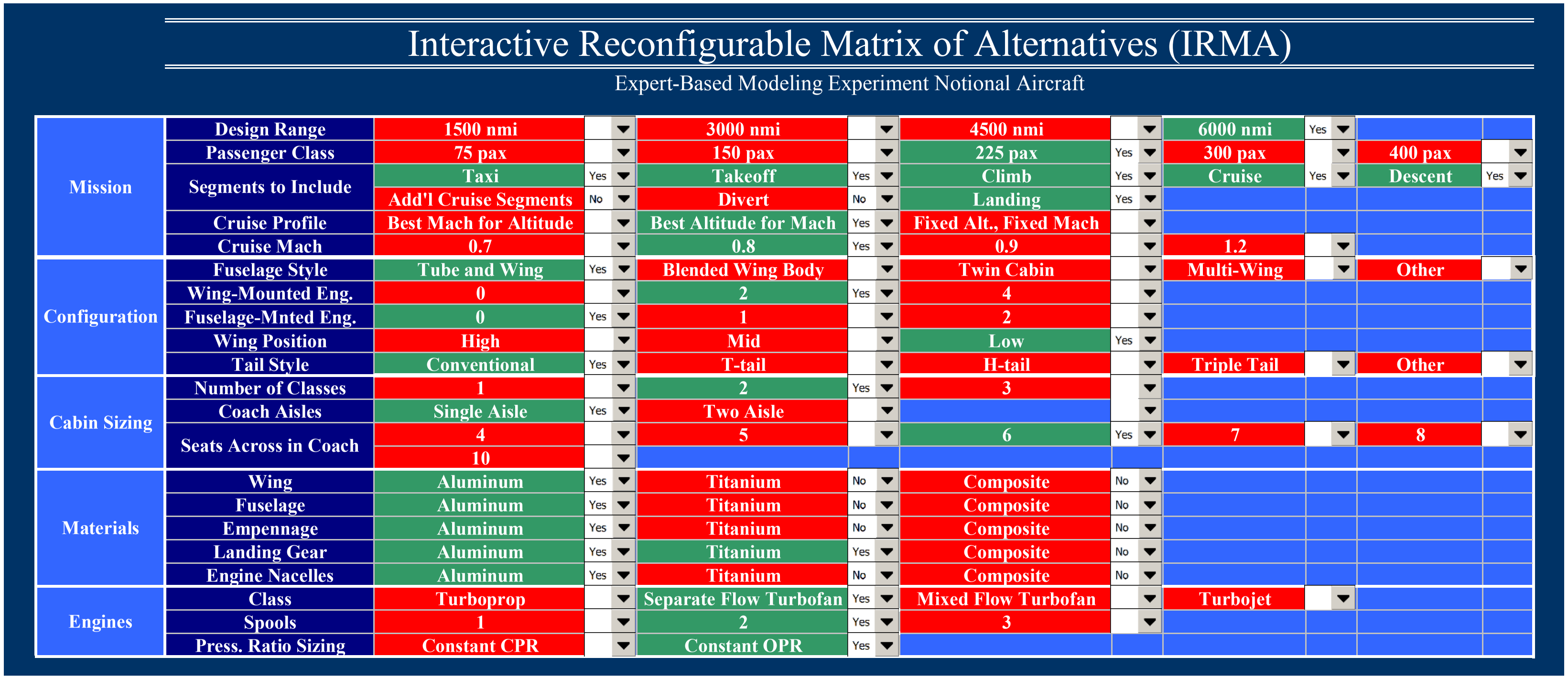
5.1.2 Definition of a Notional Aircraft

The preliminary experiments used 70-passenger class aircraft and 300-passenger class aircraft. While there are civil aircraft at further extremes of size, these are on the larger and smaller ends of the spectrum for a typical airliner. Considering the difference in the trends from the first preliminary experiment, being at any extreme might impose an additional source of error since experts would be less likely to correctly estimate uncommon relationships. For this purpose a notional 225-passenger class aircraft was selected. Several models were available for pre-existing aircraft, but using a model for a well-known vehicle may cause experts to provide memorized statistics or use reference data that would not be available for a new design.

As part of the aircraft definition, a mission was also selected. For civil aircraft the primary mission characteristic is the cruise range. For this example a range of 6000 nmi was selected. Other aircraft in the 225-passenger class with a range of approximately 6000 nmi include the Boeing 767-200ER and the Airbus A310-300 [1, 18]. The actual passenger capacity of these may vary slightly depending on the carrier's layout and number of cabin classes.

These two characteristics were the first two characteristics added to the IRMA shown in Figure 37. The remaining rows were populated based on information about an existing notional baseline that had been shown to have good behavior. Typically this would be created based on possible design decisions and filled in by a smaller subset of designers. This approach follows the path that designers have already narrowed down the design to what is available and have turned it over to the modeling and simulation group to create an initial model. Thus this IRMA served primarily as documentation rather than a test of the applicability of this particular aspect of the method.

The total number of possible design combinations in this IRMA, if no decisions have



Interactive Reconfigurable Matrix of Alternatives (IRMA)

Expert-Based Modeling Experiment Notional Aircraft

Mission	Design Range	1500 nmi	▼	3000 nmi	▼	4500 nmi	▼	6000 nmi	Yes ▼		
	Passenger Class	75 pax	▼	150 pax	▼	225 pax	Yes ▼	300 pax	▼	400 pax	▼
	Segments to Include	Taxi	Yes ▼	Takeoff	Yes ▼	Climb	Yes ▼	Cruise	Yes ▼	Descent	Yes ▼
		Add'l Cruise Segments	No ▼	Divert	No ▼	Landing	Yes ▼				
	Cruise Profile	Best Mach for Altitude	▼	Best Altitude for Mach	Yes ▼	Fixed Alt., Fixed Mach	▼				
	Cruise Mach	0.7	▼	0.8	Yes ▼	0.9	▼	1.2	▼		
Configuration	Fuselage Style	Tube and Wing	Yes ▼	Blended Wing Body	▼	Twin Cabin	▼	Multi-Wing	▼	Other	▼
	Wing-Mounted Eng.	0	▼	2	Yes ▼	4	▼				
	Fuselage-Mnted Eng.	0	Yes ▼	1	▼	2	▼				
	Wing Position	High	▼	Mid	▼	Low	Yes ▼				
	Tail Style	Conventional	Yes ▼	T-tail	▼	H-tail	▼	Triple Tail	▼	Other	▼
Cabin Sizing	Number of Classes	1	▼	2	Yes ▼	3	▼				
	Coach Aisles	Single Aisle	Yes ▼	Two Aisle	▼		▼				
	Seats Across in Coach	4	▼	5	▼	6	Yes ▼	7	▼	8	▼
		10	▼								
Materials	Wing	Aluminum	Yes ▼	Titanium	No ▼	Composite	No ▼				
	Fuselage	Aluminum	Yes ▼	Titanium	No ▼	Composite	No ▼				
	Empennage	Aluminum	Yes ▼	Titanium	No ▼	Composite	No ▼				
	Landing Gear	Aluminum	Yes ▼	Titanium	Yes ▼	Composite	No ▼				
	Engine Nacelles	Aluminum	Yes ▼	Titanium	No ▼	Composite	No ▼				
Engines	Class	Turboprop	▼	Separate Flow Turbofan	Yes ▼	Mixed Flow Turbofan	▼	Turbojet	▼		
	Spools	1	▼	2	Yes ▼	3	▼				
	Press. Ratio Sizing	Constant CPR	▼	Constant OPR	Yes ▼						

Figure 37: Reference IRMA Given to Participants

been made, is calculated as

$$(4 \times 5 \times 2^8 \times 3 \times 4) \times (5 \times 3 \times 3 \times 3 \times 5) \times (3 \times 2 \times 6) \times (2^3 \times 2^3 \times 2^3 \times 2^3 \times 2^3) \times \dots$$
$$(4 \times 3 \times 2) = 1,174,136,684,544,000$$

or approximately 1.17×10^{15} (1.17 quintillion) possible combinations before taking incompatibilities into account. The actual feasible number is much smaller due to the incompatibilities but is difficult to account for accurately. What is not captured in this IRMA are the direct effects this has on the design. The number of aisles, passengers, and seats across defines the approximate size of the passenger compartment to be 136 ft long and and 16 ft respectively. This gives an idea of the size of the fuselage. The number of passengers also suggests that total payload will be around 50,000 lb, assuming there is no other cargo. Other metrics were defined as part of the design variables or were left up to the assumptions of the participants.

When participants were giving data, each was given a printed copy of this IRMA to establish an initial design class. The same approach would be taken for any expert who was giving data who was not part of the initial design decisions. The situation would be similar if a modeler was asked to create a performance model for a particular vehicle and so it does not significantly change the workflow at this point. The benefit of using a hard copy was to make it easier for participants to refer to it without having to rearrange windows on their monitor or depend on all participants having access to a dual-monitor computer. Somewhat worrisome was that several participants asked why they needed the IRMA when it was given to them. No data was collected on how frequently the IRMA was referenced or if it was referenced at all by participants.

The technologies are fixed in this case (contrary to the academic problem this information was based on) since many of the decisions that would vary technologies are included here. Several participants asked to confirm that technologies (and corresponding technology *k*-factors) were not changing. Note that engine technologies and characteristics are not well-defined. Detailed information was not available about the engine model used. During the process of collecting information, only a single expert requested more information about the engine.

5.1.3 Selection of Variables and Ranges

In most aircraft design problems, the choice of what variables and metrics are independent depends greatly on the particular problem. This has been taken to an extreme in methods such as *cost as an independent variable*(CAIV) where the cost of a vehicle is determined upfront and other design considerations variables are made subject to that hard constraint. Even outside of such measures, independent variables are rarely truly independent once manufacturing and performance constraints are applied.

5.1.3.1 Design Variables

To select the design variables, the list of possible design variables from the preliminary experiments were first considered.[1] This list was quite large and too large to fit on a single screen or to expect experts to be able to accurately assess. In addition, many of the variables had only very minor effects on the system requirements over their design ranges. Since the tail design parameters were both the largest set of variables as well as the least impactful, these were removed first.

Wing taper ratio was also removed because it had little effect over the design range associated with it. It may have been interesting to include several variables that had little impact to see whether or not participants realized this or if they attempted to assign stronger relationships than were true just to make sure that all variables were used. This approach was considered, but rejected in order to further focus the efforts on getting correct relationships rather than “trick questions”.

The original list included two variables for the thickness-to-chord ratio: one at the root of the wing and one at the tip. While it is possible to vary both independently, in common

[1] These are listed in Table 2 on page 80.

design trends, they are somewhat correlated. Both have slightly different effects on higher level metrics, but it was determined that most participants would likely give them similar scores. Since this would have added more data to be collected and additional effort for the participants without a significant improvement in the results, these were combined into a single variable as the average thickness-to-chord ratio. The range for this value was based on the calculated ranges of the root and tip ranges of the original source file.

One additional design variable was added. With the exception of the thrust-to-weight ratio, all of the metrics were wing design metrics. In a tube-and-wing design with the passenger cabin already defined, the largest design decisions involve the wing and the engine. Fan pressure ratio was an attempt to include some aspect of engine cycle design. In hindsight, this was a poor choice. Including an additional variable not present in the original model required additional complexity in the truth model and also introduced an untested aspect. Furthermore, it was very difficult for experts to make trades between two significantly different parts of the system. The engine cycle and the airframe are rarely designed by the same corporation, let alone the same design group.

The thrust-to-weight ratio, on the other hand, is an excellent parameter to include since it is, by definition, a result of the interaction between the engine and airframe. Since the weight of the aircraft is included as part of the intermediate metrics and system requirements, changing the thrust-to-weight primarily was an engine sizing metric. A higher thrust to weight had a much larger impact on the thrust required than on the weight. The biggest impact on weight for this model would be in how it would affect fuel consumption.

Of the remaining wing design variables, wing planform area is the most fundamentally important aspect of a wing. The aspect ratio turns out to be a driver for the efficiency of the wing. The quarter-chord sweep angle is usually an output of the cruise mach since it has an optimal value. It was included here as another design parameter because it does have some effect if the aircraft is slightly out of the on-design conditions. The full list of design variables is shown in Table 5.

Table 5: List of Design Variables and Associated Design Ranges

Design Variable	**Symbol**	**Units**	**Minimum**	**Maximum**
Fan Pressure Ratio	FPR	-	1.5	1.7
Thrust to Weight Ratio	T/W	-	0.3	0.34
Wing Planform Area	S	ft^2	2500	3800
Wing Aspect Ratio	AR	-	7	10
Wing Average Thickness to Chord ratio	t/c_{avg}	-	0.1	0.15
Wing Quarter-Chord Sweep	$\Lambda_{c/4}$	deg	25	38

The ranges associated with the design variables were based on the established ranges in the academic design problem for the wing design variables and thrust-to-weight ratio. The range for the fan pressure ratio was determined by the range that the engine truth model had been created for.

5.1.3.2 Intermediate Metrics

The choices of intermediate metrics were more constrained than the choices for design variables or system requirements. Any metric included here had to vary based on the design variables and also drive the system requirements. The two-level hierarchical structure of the models means that the only way for information to move from the design variables to the system requirements is through the intermediate metrics. The truth model does not have this limitation. So if a design variable impacts a requirement in truth but the effect is not captured in an intermediate metric, there is no way for the expert-based model to correctly capture that relationship.

The third preliminary experiment also showed that the intermediate metrics needed to be as uncorrelated as possible. Since aircraft are highly integrated and since the effect of certain design variables (wing planform area in particular) have a large effect on many aspects of the design, this requirement was the most difficult. In truth, this requirement was not met as well as was hoped.

The choice of metrics started with those used in the preliminary experiments. Most of the weights in the original list did a poor job of transferring information and only affected the overall weight of the aircraft. Even then, their effects were dwarfed by the effects of the wing and engine. Since the tail design variables had already been removed, the horizontal and vertical tail weights had nothing that would have a significant impact on them. The landing gear and hydraulics weights were primarily driven by technologies in the previous case and, as before, had smaller effects on the variance of the overall vehicle performance within this type of model.

The wing and engine weights were included both because they have a significant impact on the overall weight of the vehicle, but also because they each serve as a way to translate wing size and engine size up to the higher levels. The thrust-specific fuel consumption (TSFC) was included as well. Since it changes throughout the mission as the aircraft weight changes as fuel weight decreases, it is necessary to define it at a specific point in the mission. The cruise segment is the longest and TSFC has the greatest impact on overall fuel weight from that point. For convenience, the point at the start of cruise was defined since there was less uncertainty about the aircraft's condition than at the end of cruise.

The aerodynamic parameters were modified only slightly from the original list. The lift-to-drag ratio was retained since it was independent of the other variables more than most other aerodynamic parameters and also had the benefit of capturing the efficiency of the wing. Like the TSFC, the lift-to-drag ratio was defined at start of cruise since it also varies over the duration of a flight. Induced drag is highly correlated with the lift-to-drag ratio since the induced drag is generally calculated by a multiplier on lift. Since these two tended to be redundant, induced drag was removed. The zero lift drag coefficient was kept. Unfortunately, there is a strong relationship between wing weight and zero lift drag coefficient as a result of both being highly dependent on wing area. Wing area is the primary driver of wing weight and part of how the normalization of drag into a drag coefficient is defined. It was included as a way to test the impact of some correlation remaining in the model. The zero lift drag coefficient is usually constant for a particular configuration, but it does vary between flight regimes.

Unlike the design variables, the ranges for the intermediate metrics did not need to be determined ahead of time. The final list of intermediate metrics is shown in Table 6.

Table 6: List of Intermediate Metrics

Metric	**Symbol**	**Units**
Wing Weight	$\mathrm{W}_{\mathrm{wing}}$	lb
Engine Weight	$\mathrm{W}_{\mathrm{eng}}$	lb
Thrust-Specific Fuel Consumption at Start of Cruise	TSFC	$\frac{lbm}{lbf-hr}$
Lift to Drag Ratio at Start of Cruise	L/D	-
Zero Lift Drag Coefficient at Cruise	C_{D0}	-

5.1.3.3 System Requirements

Of the three sets, the system requirements were the easiest to determine. The goal here was primarily to minimize the total number of relationships experts would have to give by minimizing the number of outputs. Of those included in the original list, several were primarily dependent on technology changes: nitrous oxide emissions and the research, development, testing and evaluation cost. The aspects of emissions that are not based on technology changes are a result of the total fuel burn which is captured separately in the block fuel weight. Since technologies would not be changing for this exercise, these were eliminated.

Other requirements were redundant since they often measured the same thing or were highly related. The landing field length and takeoff field length are both limited to the length of the same airstrip. Since takeoff field length is usually the limiting factor and landing field length is also heavily dependent on the approach velocity, it was removed. The takeoff gross weight is the sum of the operating empty weight, block fuel weight and passenger/cargo weight. The passenger/cargo weight was defined once the number of passengers were identified. The other two are included in ouputs. Since it would be trivial to calculate takeoff gross weight after the fact based on the other outputs, it was also removed from the list.

Direct operating cost of an aircraft at current fuel prices is primarily driven by the amount of fuel an aircraft uses. The other aspects of direct operating cost are based on crew and maintenance costs, neither of which is included in this model. Since the total fuel used for the mission is captured by the block fuel weight, direct operating cost is also relatively simple to calculate after the fact depending on the current fuel cost. The average required yield per revenue passenger mile is a combination of the direct operating cost of the aircraft and other aspect of a business including marketing, customer service payroll, and facilities costs. While this is an important aspect of airline management, these aspects are most certainly not included in this model.

Table 7: List of System Requirements

Metric	Symbol	Units
Approach Velocity	V_{app}	kts
Takeoff Field Length	TOFL	ft
Operating Empty Weight	OEW	lb
Block Fuel Weight	BFW	lb
Acquisition Cost	Acq$	$ Mil

The remaining system requirements are listed in Table 7. The takeoff field length, operating empty weight, and block fuel weight are relatively straightforward and intuitive. Several participants asked for a clarification on what was included in the operating empty weight. The acquisition cost is, intentionally missing a piece of information that is typically included when it is defined: the year it is measured. The materials used in the design as defined in the IRMA suggest that this is not an aircraft being designed and built in present-day. Even if it were, there would be an effect on the values for the acquisition cost from year-to-year. None of the participants asked for clarification on this topic.

Approach velocity is interesting since it is usually defined and determined as 1.2 times the stall velocity of an aircraft. The stall velocity is, in turn, defined by the maximum lift coefficient, atmospheric conditions, and the wing area. The maximum lift coefficient is usually a result of the design of flaps and slats and is often just an assumed value for many early modeling efforts. The atmospheric conditions are assumed at either a standard day or at the most constraining condition (such as Denver during the summer). The remaining variable is a design variable. This means that it would be easier to calculate the approach velocity without using the intermediate metrics than with them. Several experts commented on this and asked if there was any way to jump that step. It was included in this setup as a way to demonstrate the effects of imperfect model design. If this were being done on a less-developed class of problems, those relationships would not be known.

5.2 Gather Information

With the problem and model requirements defined, the next step is to gather the information to populate the relationships in the model. This section focuses on the interfaces and interactions with the participants in the study. Some of these interfaces and interactions were changed from the ideal case in order to meet Institute and federal regulations and ethics requirements.

5.2.1 Considerations Due to Research Including Human Subjects

The nature of this research necessitates including human subjects as part of experiments. Without the participation of such subjects, any investigation would be limited to comparing simplified models against full-fidelity models. While this supports the possibility of future expert-based models, it only demonstrates that simplified models have value. This claim has been well-acknowledged elsewhere with many examples of reduced order models. Such models are common in introductory texts to explain general trends of physical relationships.

The inclusion of humans as a component of a research experiment requires approval of an Institutional Review Board (IRB). The IRB is responsible for judging the risk for harm to the experts against the benefits of the study. Many of the methods included in this research are standard business practices and can be considered complicated surveys. The same oversight would not be required when using this methodology outside of a research activity.

The Georgia Institute of Technology IRB approved the protocol associated with this research as exempt status since the information gathered is a form of electronic survey or interview. It was required not to include any information that would identify subjects or disclose information that would affect employability or reputation of the subjects. This research qualifies for a waiver of documentation of consent according to Title 45, Section 46.117 of the Code of Federal Regulations since a consent document would not normally be used outside of research[125]. Note that receiving exempt status does not mean that the project is exempt from oversight. So long as a research project includes human subjects it is still required to submit for oversight. The approval notification is shown in Appendix A. The thesis advisor serves as principal investigator for purposes of responsibility after the conclusion of this research.

In order to meet the requirements to get IRB approval and minimize the impact that such oversight would have, several concessions were made that reduce the usefulness of the method as demonstrated compared to what would be expected in a use outside of research. The key measure of acceptability is the ratio between the benefit of the research against the risk to the human subject participants. While this research had virtually no danger of any physical harm, the process does test the ability for an individual to provide accurate relationships for a civil aircraft. Such information may indicate an individuals knowledge, or lack thereof. It could be argued that this information, if released, would pose a threat to their employability. Even if that were not the case, regulations only recognize minimal risk as there is never absolutely no risk. In order to protect individuals, the experiments had to be designed in such a way that the data collected would not use any personally identifiable information and protect the identities of those who participate.

Many expert-based methods rely heavily on group meetings. Others use the reputation of individuals to weight how their answers are incorporated. The experiments in this research could not take either of those approaches without risking a significant increase in red tape and bureaucracy. It also limited what demographic information would be collected to ensure that none of the information was so specific that it would be easy to identify an individual was from their information. The vast majority of the participants in this experiment had no concerns about their privacy or identity being associated with the data. In future efforts where there is more flexibility in time, it would be worthwhile to add back in some of the aspects above to test an improved version of this method.

Another aspect of any study involving human subjects revolves around how experiments themselves must be performed. With computer-based experimentation or even laboratory-based experimentation, if an error or deficiency is found in an experiment or experimental setup, it can be corrected and the experiments can be restarted. There may be a time or material cost penalty, but in theory the slate can be wiped clean and started over. With human subjects, once an individual has participated in part of an experiment, they no longer have the same experience they did before. If an error is found in the way information is presented or data is collected, restarting the experiment means that all past participants may no longer be eligible. If an error or point of improvement is found early in the process, the loss of a small number of participants may be accepted. If it's later in the process, the researcher is presented with the choice of continuing with the error so that all participants are consistent or correct it and try to account for the difference between the groups. Due to the limited number of individuals who might qualify as experts and be available for this research, it is unlikely that the experiment could be restarted or performed an alternative way after it starts.

5.2.2 Demographics of Participants

Additional information about the profiles of individuals provides additional opportunities for data analysis and insight into the performance of difference classes of experts. On the other hand, too much information about individuals may make them uncomfortable and less likely to participate in future phases. There is also the question as to what information is relevant to the research and what would be appropriate to use when making decisions about who to include in an expert-based modeling exercise.

The selection of experts is a critical part of this methodology. Experts are often nominated by someone else or selected by a supervisor based on a reputation of expertise, as discussed in Section 2.1. For this research, participants must volunteer. This means that ensuring that there are sufficient participants, each of whom are the most expert in their fields is unlikely. Participants were invited to volunteer so long as they had any experience with aircraft design, preferably with civil passenger aircraft. Participants were not filtered or rated by experience before they gave data. Therefore, in order to assess the expertise of participants, each individual was asked to provide information about their education and amount of experience.

The education questions focused on the highest level of education. Each expert was asked to provide their highest level of education completed and the field of that level of education. The choices for these two questions are shown in Table 8. Since the experts were being drawn from among graduate students and research faculty within the School of Aerospace Engineering, it was expected that all participants would have at least a bachelor's degree, so that was the lowest level of education. Also, traditionally students and faculty come from an aerospace engineering or mechanical engineering background so only those fields are called out by name. Note that the education level includes the option "Passed PhD Qualifiers" as a level of education. This was included as recognition that many doctoral students have the training of an individual with a PhD prior to completing the dissertation associated with that degree.

The other demographics questions were specifically about their experience. Originally the only two experience questions were for individuals to identify how many years of experience they had with aircraft design and how much experience they had with civil airliner

Table 8: Possible Responses for Education Demographic Questions

Highest Level of Education	Field of Highest Level of Education
Prefer not to answer	Prefer not to answer
Bachelor of Science	Aerospace Engineering
Master of Science	Mechanical Engineering
Passed PhD Qualifiers	Other Engineering
Doctorate	Other Field

performance. The choices for these ranged from no experience to more than 8 years, with the option to not answer. The specific choices of bins of experience are shown in Table 9. During the instructions, participants were asked to include any coursework that focused on that field as well as any academic or contract research and work experience outside of the Institute. It was up to each individual to determine whether a particular project counted towards this count of experience.

Table 9: Possible Responses for Experience Demographic Questions

Experience with Aircraft Design	Experience with Civil Airliner Performance
Prefer not to answer	Prefer not to answer
None	None
0–1 year	0–1 year
1–2 years	1–2 years
2–4 years	2–4 years
4–8 years	4–8 years
More than 8 years	More than 8 years

During a period of unscientific interface testing to identify potential bugs in the collection tool, one individual remarked that they considered themselves to have a great deal of experience, but had been working in a different specialty for the past few years and felt “rusty”. It was still easy to add an additional demographic question and there were not so many that it felt burdensome, so it was included as an additional qualifier for expertise. The options for this choices are shown in Table 10 but mostly mirror the experience questions with an additional option for those currently working in the area.

Other demographics were considered for inclusion as well. Experts could have given

Table 10: Possible Responses for Time Since Experience with Civil Airliners

Time Since Experience with Civil Airliners
Prefer not to answer
No experience with airliners
Currently work with airliners
0–1 year
1–2 years
2–4 years
4–8 years
More than 8 years

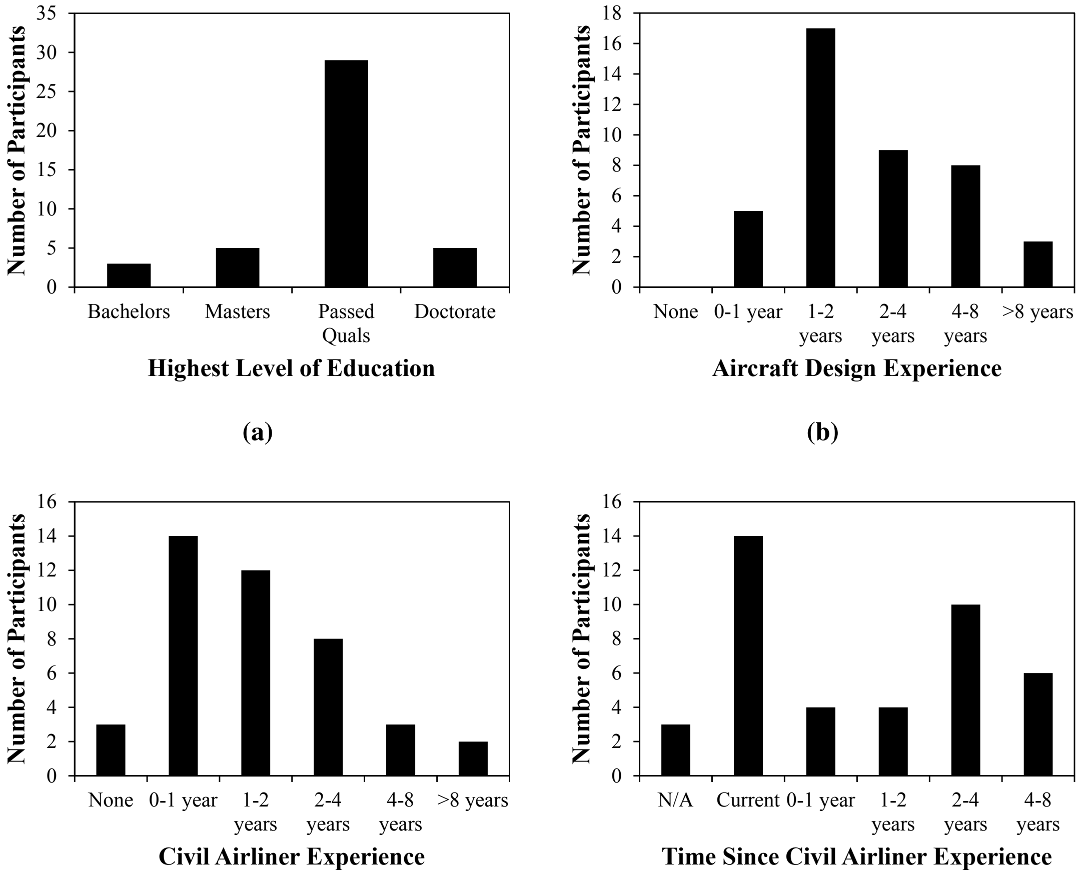
their specialty in an open-ended format where typical answers might be ‘propulsion systems’, ‘noise and emissions’, or ‘structural design’. It was unlikely that there would have been enough of any one subject area to be useful for analysis and this information, if included, may result in the reported data being too specific and risk losing the anonymity of participants. This information would be important to include when this process is performed within an organization to ensure that a variety of viewpoints are included and for documentation purposes.

Gender was also considered. It is an important factor in both physiology and psychology. In some cases, failure to capture any differences between genders may affect the publishability of the results. The American Physiological Society recently announced that all future publications involving human studies must report the gender as part of their findings [95]. The United States branch of the Institute of Electrical and Electronics Engineers identified this requirement as a direction other journals, including engineering journals, should follow[85].

While no physiological effects were measured as part of this research, state of mind, and intuition with different scales or interfaces may have an effect on the quality of information that an expert gives. At the end, gender was not included for several reasons. The pool of applicants could not guarantee sufficient volunteers of each gender to be useful for analysis. In addition, using experts in this fashion is still relatively immature and such tests would be better suited for a follow-on after the method has been further refined.

5.2.2.1 Distribution of Participants by Demographics

Forty two individuals volunteered in total. With the exception of only a single individual, all participants' highest field of education was aerospace engineering. Without a second option with enough individuals to compare against, there's no need to perform any further analysis with it. Figure 38 shows the distribution of the forty two volunteers across each of the other four demographic questions.



(a) **(b)**

(c) **(d)**

Figure 38: Summary of Education and Experience of Participants

Figure 38a shows that the participants were mostly PhD students who had completed their course work. The author tried to encourage students working towards their masters degree to volunteer, but very few responded. The author also tried to encourage research faculty who had completed their doctorates to participate, but there were fewer of those in the population. Because there are so few individuals in the groups other than PhD students who have completed the qualifiers, it is unlikely that any statistically significant analysis could be performed, even if it was partitioned into two groups.

Figure 38b shows that nearly all participants had at least one full year of aircraft design experience. There are fewer individuals with higher levels of experience as would be expected. Individuals change research projects over time and there is less funding in areas of aircraft design. Additionally, fewer individuals with higher amounts of experience remain as students and leave the Institute, thus moving out of the pool of potential volunteers. There are enough individuals in each group to perform comparisons between their results in Chapter 6. For some comparisons, the group may be split into those with less than two years of experience and those with more than two years of experience.

Figure 38c shows that individuals with experiences specialized in civil airliner performance mirror the distribution of individuals with aircraft design experience, but shifted down one level of experience. It makes sense that as a field gets more specific fewer individuals will have as much experience as the more general grouping. If another question asked about the experience with civil airliner aerodynamics performance, the distribution would continue to shift towards fewer in each bin above none.

Figure 38d shows that just under half of the participants are either currently working with civil aircraft or are less than a year away from working with civil aircraft. There is also a large group of individuals who have not worked with civil airliners in more than two years. When comparing the responses for both years of civil airliner experience and time since having that experience, most of the individuals who have not worked with airliners in a while have less than a year of experience. This suggests that many of those individuals have not worked with airliners since their course work in their first year of graduate school. These individuals would generally be considered the least experienced next to those with no civil aircraft experience at all.

These demographics and how they correspond to the experts' degrees of agreement and accuracy will be analysed in more detail in Chapter 6.

5.2.3 Data Collection Interfaces

In the strictest sense, experts could have provided their information using only paper-based surveys. Much of the information that is requested could be displayed on a printed scale and experts could circle the appropriate value. Other data collection methods have used an Excel-based form or a series of displayed questions and provide each expert with a remote control to provide his or her answer. In this case, it was preferable to have a data collection tool that was easy to distribute to each participant and made it easy for participants to return their data. In addition, since there would be three modes of interfaces for the three scales of interest, it was preferable to have a common back-end so that different participants had as few differences as possible. Using a single common tool also minimized the amount of repetition and coding required, reducing the chance of bugs in the tool.

The data collection tool is coded in the JMP scripting language and is based around four main tabs or screens. The first tab is a brief written summary of the instructions. With only a few minor differences for different interfaces, the instructions were consistent for all modes. These were included as a point of reference for participants or in the off chance that they were unable to receive the instructions verbally and in person. The second tab provided an interface for the participant to provide relationships between the design variables and intermediate metrics. Regardless of which interface is used, the design variables are listed across the bottom of the screen with their corresponding design ranges. On the left-hand side of the screen, the intermediate metrics are listed. Each metric has two editable text boxes, one labeled max and one labeled min. These boxes allow participants to estimate the minimum and maximum values of each intermediate metric over the ranges of all the design variables.

The third tab uses the same layout as the second but lists the intermediate metrics across the bottom and the system requirements along the left. The ranges of the intermediate metrics provided in the previous step are transferred to this tab. The last tab contains the demographics questions as well as a series of questions asking how difficult each step was. At the bottom of these questions is a space to allow participants to add any comments they would like to pass along.

Since the mode that a participant used was directly controlled by the researcher, it was much easier to ensure uniform distribution between the modes than with the levels of education and experience as shown in Figure 39. The slight difference is due to a small number of participants volunteering and being assigned to an interface but never actually providing data. Participants were assigned to a particular interface with the goal of distributing different levels of education and experience equally across all three. Due to the limited number of participants within certain groups of education and experience, this was not always possible.

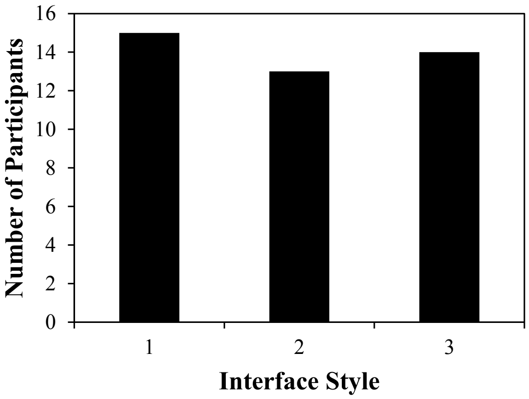
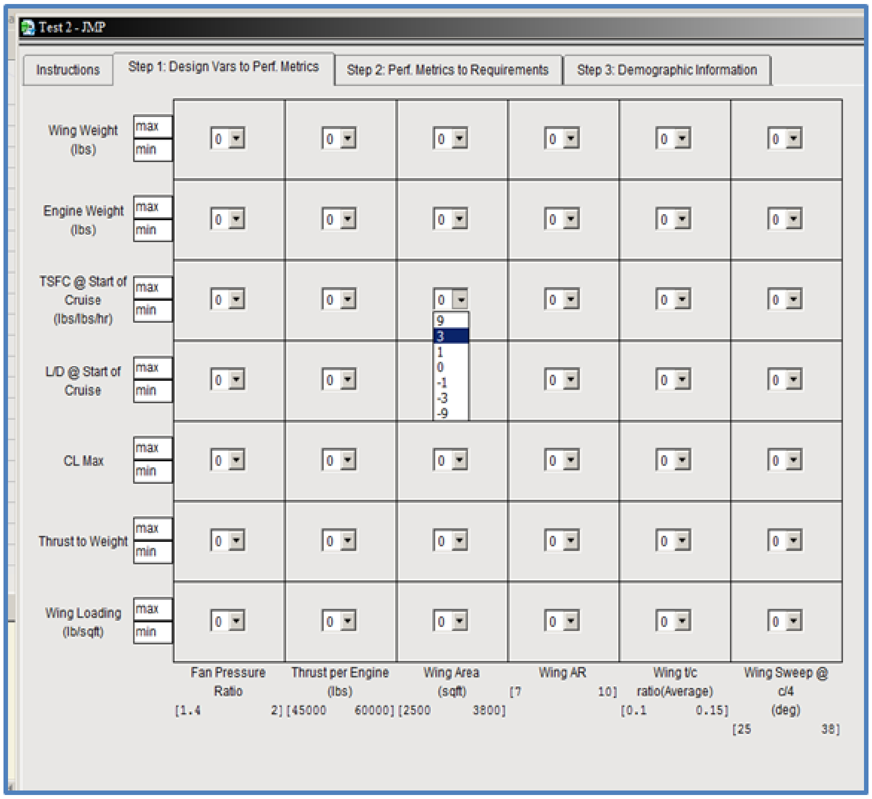


Figure 39: Distribution of Participants on Each Interface

5.2.3.1 Mode 1: Standard QFD Scale

The first interface mode is based on the standard QFD scale. Where the original scale would be Weak, Moderate and Strong, here it has been translated to the scores [0, 1, 3, 9]. It also includes the negative values of the scores for a full list of options: [-9, -3, -1, 0, 1, 3, 9]. This interface uses a drop-down box for each relationship to present the scores since the choices are discrete and nonlinear. A partial screenshot of this interface is shown in Figure 40.


Test 2 - JMP
Instructions
Step 1: Design Vars to Perf. Metrics
Step 2: Perf. Metrics to Requirements
Step 3: Demographic Information
9
3
1
0
-1
-3
-9

		Fan Pressure Ratio [1.4 2]	Thrust per Engine (lbs) [45000 60000]	Wing Area (sqft) [2500 3800]	Wing AR [7 10]	Wing t/c ratio(Average) [0.1 0.15]	Wing Sweep @ c/4 (deg) [25 38]
Wing Weight (lbs)	max min	0	0	0	0	0	0
Engine Weight (lbs)	max min	0	0	0	0	0	0
TSFC @ Start of Cruise (lbs/lbs/hr)	max min	0	0	0	0	0	0
L/D @ Start of Cruise	max min	0	0		0	0	0
CL Max	max min	0	0	0	0	0	0
Thrust to Weight	max min	0	0	0	0	0	0
Wing Loading (lb/sqft)	max min	0	0	0	0	0	0

Figure 40: Interface Mode 1: Traditional QFD Scale with Drop-Down Boxes

The relationships are aligned in a grid to make it easier to follow along. The design variables along with their units and ranges are across the bottom. The text boxes at the left still read "min" and "max" since data has not yet been provided for the ranges on each intermediate metric.

5.2.3.2 Mode 2: Integer Scale

Mode 2 generalizes the scale of Mode 1 to include all integers in the range between -9 and 9, for a total of 19 possible scores. A drop-down box listing this many choices would be difficult to use. At this point, and because the scale is more regular, a slide bar is more intuitive. Next to the slide bar is a read-out of the value of the current setting. Otherwise this interface has the same layout as the previous one, as shown in Figure 41.

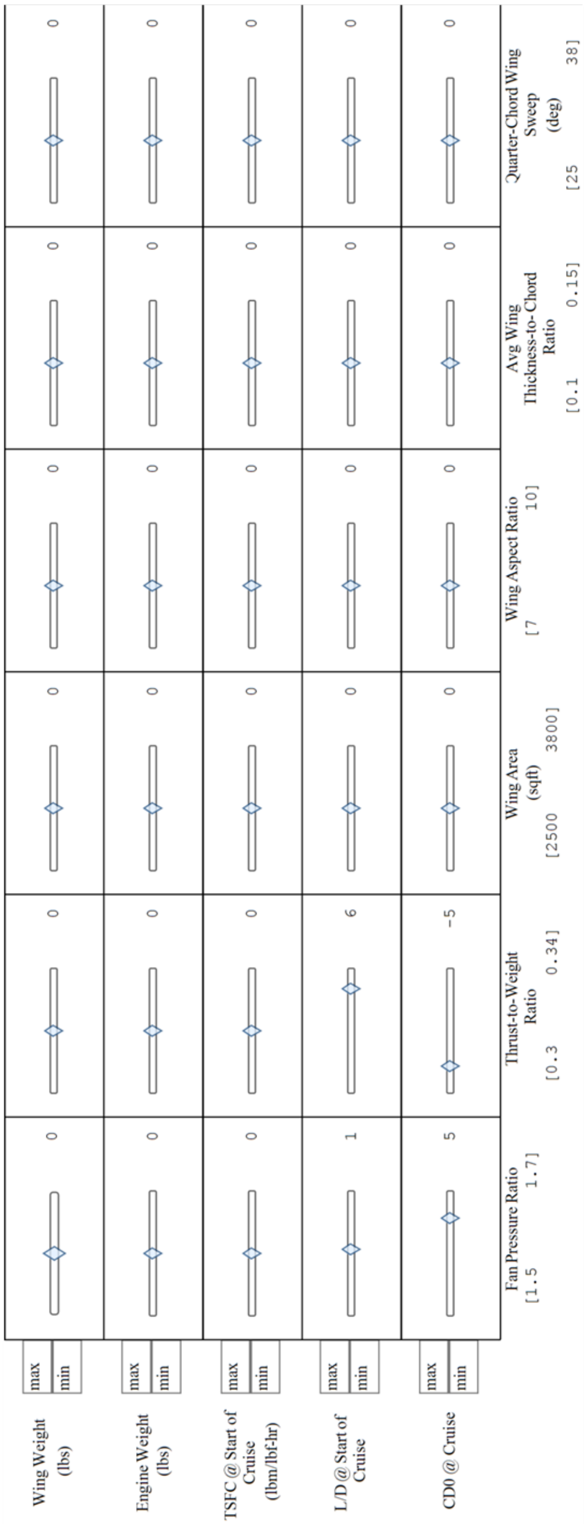


Figure 41: Interface Mode 2: Integer Scale with Slidebars

One of the noticeable effects of including a slidebar is the change in aspect ratio. For more than six variables across the bottom, this begins to become an issue of fitting onto the screens of participants. Those with widescreen monitors could handle 2-3 more variables, but those with conventional 4:3 ratio monitors would be limited in what was visible at once.

5.2.3.3 Mode 3: Graphical Slopes

Mode 3 is a further generalization of Mode 2 since it includes the same range of values but also permits quadratic relationships to be given. This mode also includes a graphical depiction of each score. Figure 42 shows a notional example where only the lower left corner has been filled out. For this example, the majority of the relationships are linear and include the same slide bar as in the previous mode. Under each plot was also a check box that allowed a participant to switch from linear relationships to quadratic relationships. When this box is checked, the slide bar is hidden and three points appear along the line. Each of these points is fixed in the horizontal direction but may moved vertically by clicking and dragging with the mouse. These points are limited to the bounds of the plot so that unreasonable relationships are not given.

Several participants suggested scaling each graph by the same normalization that would occur after scores were compiled. This would give a more accurate feel for the effect of the impact normalization has and ensure that experts are truly giving the relative contribution of each term rather than just giving independent raw scores. This would also provide a more accurate match up against the interface that individuals comfortable with prediction profilers are accustomed to. Unfortunately, the way this was coded, it was difficult to retrofit that functionality without breaking the quadratic interface. Future versions may experiment with this to see how it affects the quality of scores and the impression of the interface for experts.

Originally the slide bar was to include an entirely continuous range of all values between -9 and 9. Like any other digital implementation, it could not be truly continuous and

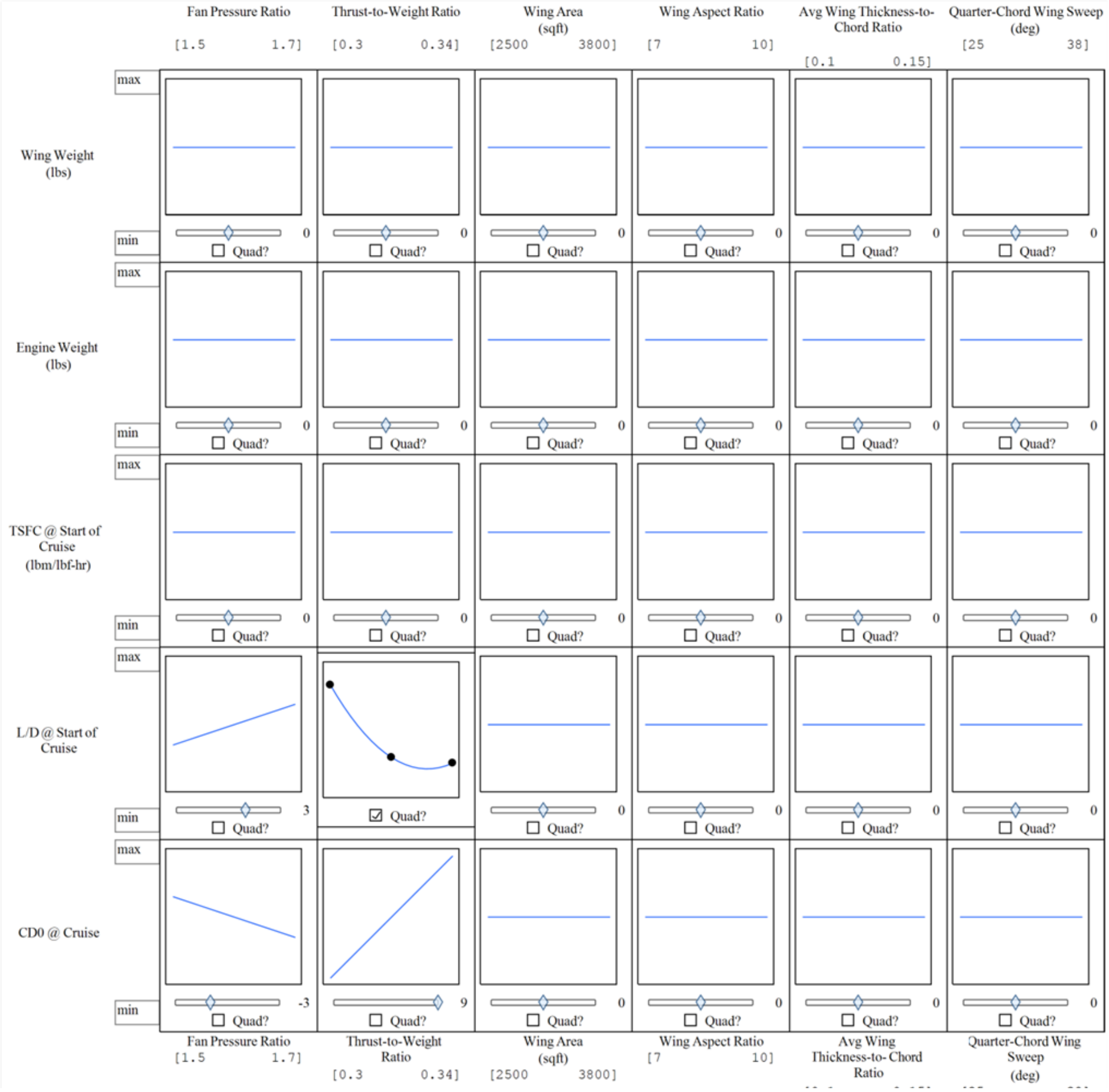


Figure 42: Interface Mode 3: Graphical Scale with Possible Quadratic Terms

so, due to formatting and limits of the number of incremental steps in a slide bar, the scale included all values in that range to the nearest tenth. During one of the preliminary interface bug testings used to make sure that everything was working properly, several testers commented that a continuous scale felt more difficult to use than the previous integer scale. When not given specific units, humans tend to think better in natural numbers, especially in subjective scoring. The slide bar was limited to the same integer steps as the previous mode so that, for linear scores, the only difference between this mode and the previous mode was that this one included a graphical depiction of each slope.

5.2.3.4 Estimation of Ranges of Intermediate Metrics and System Requirements

All three interface modes included boxes where experts were asked to give their estimates of the largest and smallest possible values for each intermediate metric and system requirement over the ranges of the given design variables. Ranges of design variables were given to provide experts with a unified starting point and because different ranges would produce different relationships. These same ranges should have been necessary for the relationships between intermediate metrics and system requirements. Units were listed throughout so that experts would give values on a consistent scale.

Some of the ranges experts gave for the intermediate metrics and requirements had to be corrected afterwards to interpret their intentions. This was only done when it was clear that the error was due to the interface or an understanding of it rather than a failure to understand the problem or what a correct value should be. For example, the interface asked for acquisition cost in millions of dollars ($M). A realistic answer would be on the order of hundreds or large tens. Several experts gave values on the order of tens of millions or hundreds of millions. In this case, it is unlikely that they actually meant to suggest that an aircraft would cost trillions of dollars, so these values were normalized down. Not all values were corrected in this way. For example, zero lift drag coefficient values should generally be on the order of 1×10^{-2} and are almost always listed as such. Some participants gave values on the order of 1×10^{-1} or 1×10^{0}. While these are incorrect, they are not so obviously incorrect that it is safe to assume that the error was with the interface or understanding of it.

Other values had the minimum and maximum values switched. Since this was another case where the error was clearly due to the interface or understanding of it, these were put in the correct order. A very small number of experts elected not to provide ranges for some or all of the values. These individuals did not give reason for not including them, so there is no data to support what made an individual more or less likely to provide the ranges. These were not included in analysis.

The results for the ranges and how they compare to the truth model are discussed in Section 6.6. The distributions of the estimated ranges for all metrics and requirements are shown in Appendix C.

5.2.4 Instructions to Participants

After volunteering and scheduling a time to participate, each individual was given a set of verbal instructions on how to use the data collection tool and what information was being asked of them. On a small number of occasions, multiple participants asked to be given instructions at the same time. In these cases, there was no significant deviation from the standard set of instructions that were given.

Upon arriving at a participant's desk, the researcher distributed a paper copy of the informed consent document (which had been included in the original recruiting email) and a paper copy of the IRMA. Recall that the preliminary experiment described in Section 3.2.1 showed the effect of different classes of vehicles on the relationships between variables. A small number of experts, after having been given the IRMA asked what it was for and why they needed it. The baseline information was quickly explained, but that some experts did not realize that the size of the vehicle would affect the relationships was disconcerting. The interface was loaded with the interface mode and random unique identifier set beforehand.

The first tab was explained as the instructions but quickly moved to the second tab. The layout of the screen was explained briefly.

Participants were instructed to work in a single row for an intermediate metric at a time and start by identify the design variable which has the greatest impact on the variability of the metric. Once a design variable was selected, the direction of the relationship was determined. Depending on the direction, participants were instructed to give this relationship a large positive or negative score. Other design variables were considered in order of decreasing effect. While participants using the Mode 1 interface had a clear delineation between strong, moderate and weak scores in each direction, it was less clear to individuals with the other interfaces. They were encouraged to try to work around the same 1-3-9 scale and make adjustments higher or lower to meet the specific problem. The importance of the assigning values relative to each other was highlighted, explaining that giving all the relationships scores of 9 had the same effect as giving them all scores of 1. This concept seemed very intuitive to most participants. Participants were also encouraged to use a good balance of scores and told that not every relationship had to receive a score and that low values and zeros were acceptable scores. The importance of the ranges of the design variables were highlighted by pointing out that, as an example, wing aspect ratio would have a much more significant impact if it varied from 1 to 200 than from 7 to 10 as it does here. Individuals responded that they understood without being biased towards giving a particular score.

Individuals who had been assigned to the graphical interface were also shown how to provide relationships using the quadratic format. Those individuals were also encouraged to start by defining the linear relationship and then switch to the quadratic form to make adjustments from it.

After providing the relationships for each of the rows, participants were asked to provide the ranges for each of the intermediate metrics. Many of the participants were familiar with the concept of a design of experiments. Therefore, the ranges were explained as the minimum and maximum values if a design of experiments were performed over the ranges for the design variables. Many participants indicated that they were uncertain about the exact values, but were encouraged to give it their best effort in spite of their doubts.

Once the relationships and ranges for the intermediate metrics were complete, experts were instructed to move to the next tab which followed the same process at one level higher. They were then shown the final tab and the demographic and difficulty questions listed there.

Finally, participants were asked to focus on giving data and try to avoid distractions while working on it. Participants were encouraged to ask any questions, which the moderator answered to the best of his ability. The entire instructional period typically required between 9 and 11 minutes. The researcher returned to the participant's desk about 15 minutes later to check that participant was making progress and to answer any additional questions that had come up. When the participants finished giving data, they returned the output file containing their data to the researcher.

5.2.5 Distributions of Raw Data

The scores for all participants were combined into a single data table for analysis. The distributions of those scores are shown in Figures 43 and 44. For these plots, each separate bar chart is the distribution of all scores for that particular relationship across all interfaces and participants. These plots represent the raw scores before normalization. Only the linear component is shown of quadratic relationships. For each relationship, the mean and standard deviation of the raw scores in shown in the upper left. The mean is indicated by a red vertical line. Detailed analysis of the agreement between experts is discussed in Section 6.2.

These plots show trends between experts for each relationship. But there are also trends between relationships for each expert. That is, there are some participants that tend to give most relationships a higher score than average and some that tend to give most relationships

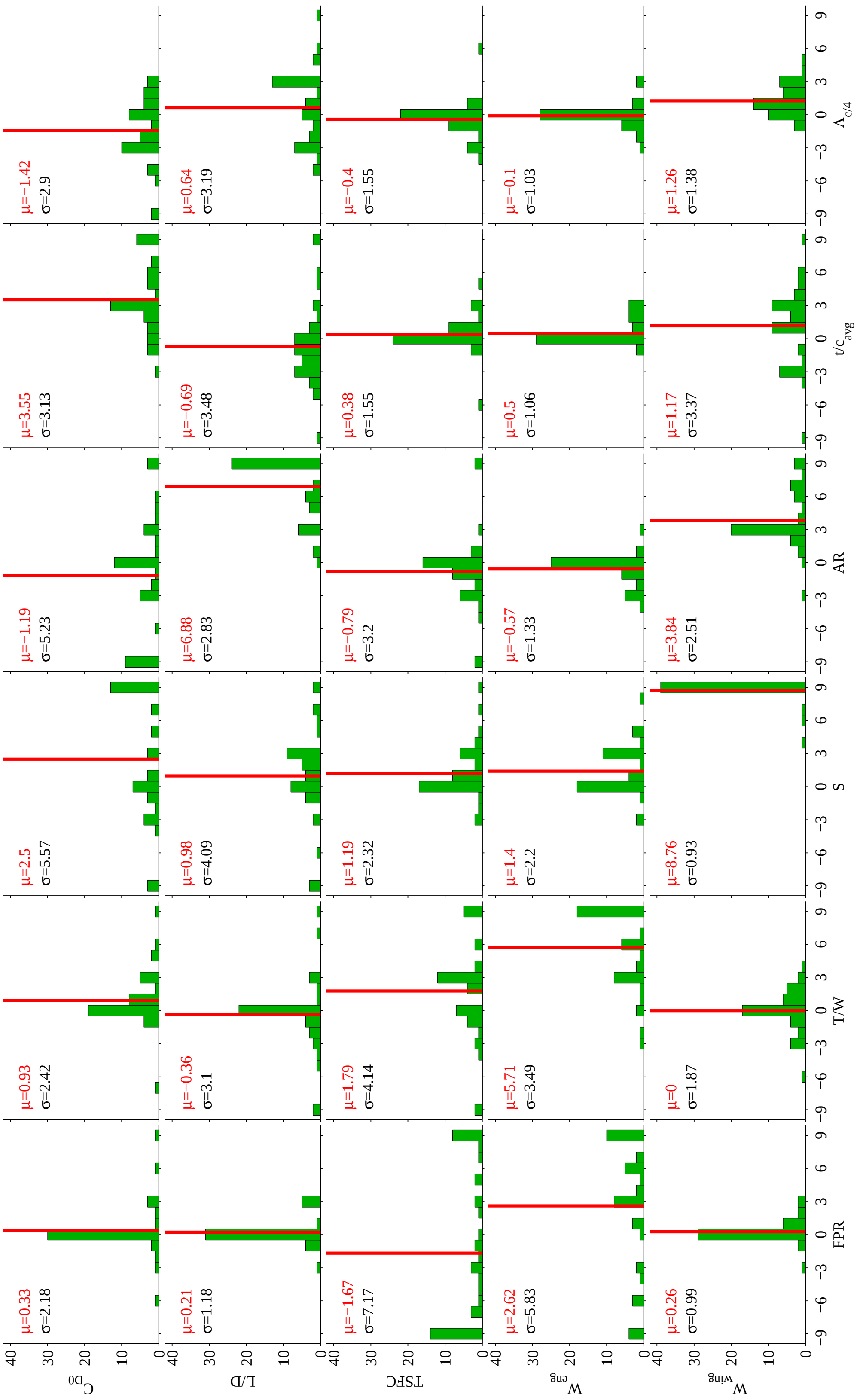


Figure 43: Distributions of Uncorrected Scores for Intermediate Metrics vs. Design Variables

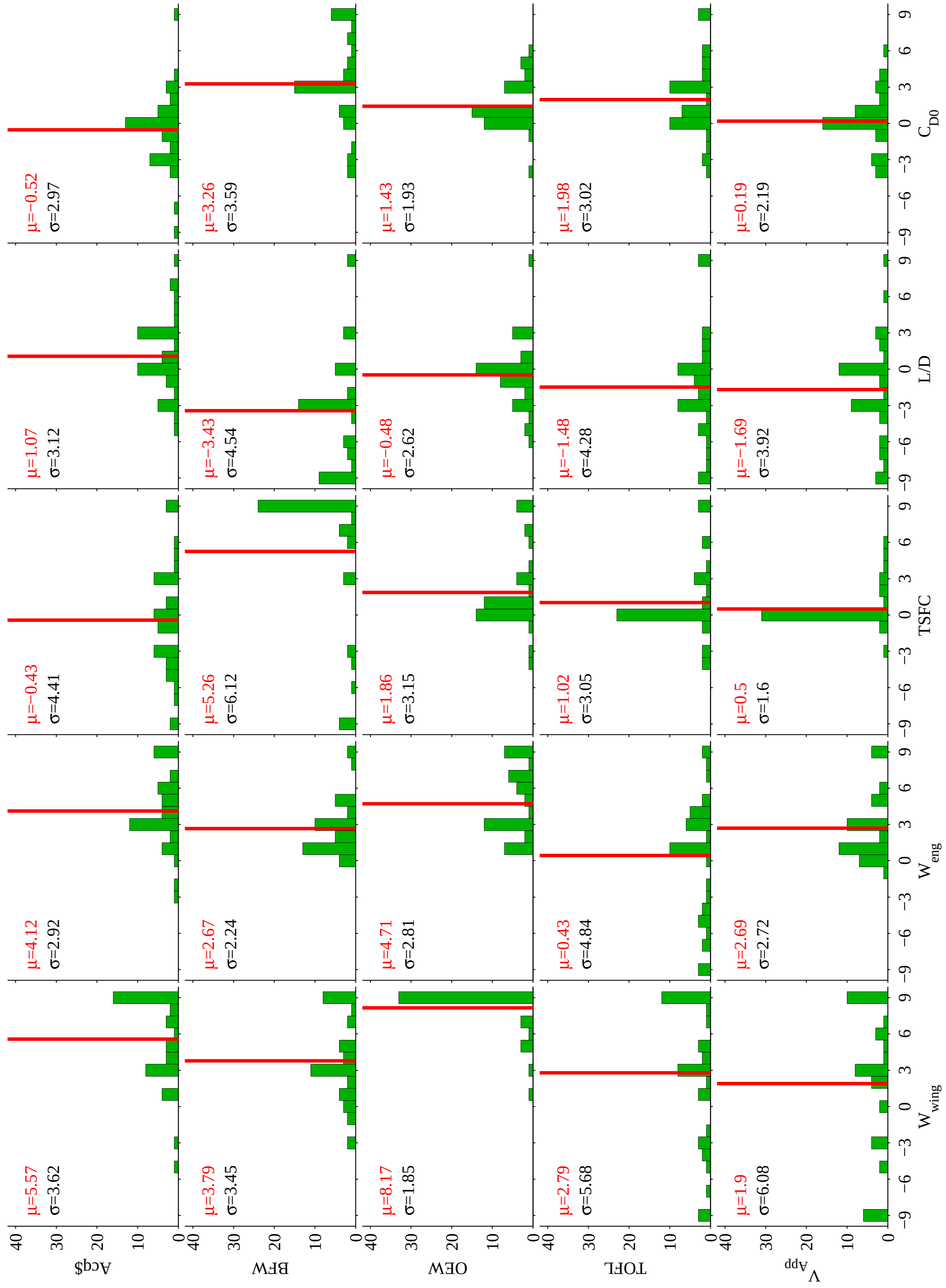


Figure 44: Distributions of Uncorrected Scores for Requirements vs. Intermediate Metrics

a lower score than the average person. The effect of this is that the ratio in the scores of the relationships for a given row for a single individual is much higher for individuals who give lower scores on average and the ratio is much lower for individuals who score everything a bit higher.

For some individuals it's difficult to follow the 1-3-9 scoring as it matches up to the Pareto principle and give fewer low or zero scores. This pushes their overall means toward the center. It is possible to adjust these scores to increase the difference between them, but it may be better to simply filter these individuals out of the pool of experts. An expert who claims that all of the relationships are strong or that all the relationships are weak does not have a good grasp of what is being asked of them and no amount of adjustment can correct them. The best solution is to provide better instruction about what a good distribution of scores looks like with examples of the two extremes: when all but one relationship is zero and when all relationships are given equal scores.

5.3 Create the Model

The form of the model was pre-determined during the method development. That form determined the type of data needed. Once the data is collected, model creation comes down to translating that data into a usable set of equations. The two primary steps here are normalizing the responses for consistency and and combining the data from multiple experts into a single model.

5.3.1 Normalizing Responses

The value displayed to the experts for Modes 1, 2 and the linear portion of Mode 3 was scaled from -9 to 9 to match up with typical scoring. Humans tend to provide better quality scores in natural numbers. The plots for Mode 3 translates this to value to a slope of -1 to +1 for the purposes of plotting it and consistency with the scale used for quadratic relationships. For consistency across interfaces, the other modes were normalized to the same -1 to +1 range when saved. When plotted as scores, as in Figures 43 and 44, these are returned the -9 to +9 scale to be easier to read and compare to original ratings.

For analysis and processing, however, the scores for each expert for each row were normalized such that the range of each intermediate metric or system requirement was between -1 and +1. The ranges supplied by experts were only included as an after-effect. An example of the scaling and normalization for two sets of example scores are shown in Table 11.

Table 11: Examples of Scaling and Normalization Calculations

Raw Scores		**Scaled**		**Normalized**	
Ex. 1	Ex. 2	Ex. 1	Ex. 2	Ex. 1	Ex. 2
9	9	1.000	1.000	0.450	0.300
0	1	0.000	0.111	0.000	0.033
0	2	0.000	0.222	0.000	0.067
1	-2	0.111	-0.222	0.050	-0.067
1	-2	0.111	-0.222	0.050	-0.067
3	6	0.333	0.667	0.150	0.200
6	8	0.667	0.889	0.300	0.267
Abs Sum:		2.222	3.333	1.000	1.000

The first two columns are the raw scores that two individuals, Expert 1 and Expert 2 would have given for the row of a particular intermediate metric. In this example, the row contains seven design variables and thus, seven scores. Both experts agree that the first item is the most important and give it a score of 9. They also both agree that the seventh and sixth variables are the next most important and in that order. Expert 2's scoring suggests that the seventh variable is only slightly less influential than the first.

The second set of columns show those scores scaled to between -1 and +1. Note that this does not mean that the largest value is set to +1 and the smallest value is set to -1. Rather, it means that the maximum possible value is +1 and the minimum possible value is -1 corresponding to a +9 and -9 raw score. The sum of those columns is listed as the last row. This is the sum of the magnitude of the slopes, not the value.

The third set of columns is the result of the normalization after taking the scaled scores from the second set of columns and dividing it by the sum of the magnitudes. The absolute sum for these columns is both 1. This set of scores is what will be used for analysis from this point forward.

The effect of this process is that, because Expert 2 used more high scores, the individual impact of each of the high scores is lower than Expert 1 who used them more sparingly. In some cases Expert 2 might be more correct, but in most cases, Expert 1 will more closely match the actual physical trends.

5.3.2 Combining Responses

Combining the relationships from two or more individuals is performed by averaging their normalized scores. The process is still valid when one of the participants uses a quadratic relationship. In that case, the pure linear relationships are treated as quadratic with a coefficient of 0 on the second-order term. Continuing the data from the previous example, an example of the average is shown in Table 12.

Table 12: Examples of Combining Relationships from Multiple Experts

Raw Scores		**Normalized**		**Combined**	
Ex. 1	Ex. 2	Ex. 1	Ex. 2	Norm.	Unscaled
9	9	0.450	0.300	0.375	3.375
0	1	0.000	0.033	0.016	0.150
0	2	0.000	0.067	0.033	0.300
1	-2	0.050	-0.067	-0.008	-0.075
1	-2	0.050	-0.067	-0.008	-0.075
3	6	0.150	0.200	0.175	1.575
6	8	0.300	0.267	0.283	2.550

The first set of columns are the original raw scores. The second set are the scaled, normalized scores that were the third column in Table 11. The third set of columns are the average scaled, normalized score and the same score but returned to the original -9 to +9 scale. The normalization has brought the scores closer to zero to reflect the effects of relative importance values.

The more difficult aspect is determining which participants to include. Of the forty-two who gave data, not all are experts and not all are correct. Different combinations of experts who agreed the most, had who had certain levels of education or experience or who used a particular interface will be combined separately to identify the effects of each of these differences on the results. It is also possible to use an optimization method to identify which group of experts produces the model that most accurately matches the truth model. Chapter 6 investigates each of these possibilities and discusses the implications of the results of these different comparisons.

5.4 Test the Model

For this experiment, testing the model involves analyzing the model parameters themselves. The accuracy of the output of the models will be tested as part of the *Use the Model* step. There are two primary comparisons here. The first is that the experts can be compared against each other. Hypothesis 1a claims that there will be different levels of agreement among the experts who use different interfaces. The other comparison is comparing experts, both individually and combined, against the model parameters of a truth model.

Two different metrics are used to describe the truth model, just as in the preliminary experiments. The first is the coefficients of a linear model created by regressing against points generated by a design of experiments. The second is the correlations of the values. The first is the truer representation of only the main effects according to the data and seeks to eliminate errors due to correlated inputs. It seeks to only model the causal relationships it sees. The second compares the trends in behavior, which may better match the likely trends that experts included, whether or not they were causal.

Both are measured by finding the degree of agreement between individual experts or a combined model and the truth model.

5.4.1 Measuring Agreement

Two metrics are used to measure the agreement of experts with each other. Pearson's product-moment correlation coefficient ρ_{xy} is used most frequently. Despite being limited to only comparing data sets at a time, it is fast and readily available on the multiple platforms used to perform analysis. Krippendorff's alpha α_k is used less frequently because of it's problems handling large data sets. When Krippendorff's alpha was calculated, Jana Eggink's MATLAB code was used. This code was based on Krippendorff's original implementation and compared against sample data sets to confirm that it produced the correct results. When compared against each other for multiple sets of data, the average of the ρ_{xy} for all unique pairs and α_k for the same data frequently follow similar trends. Both measures are discussed in more detail in Section 3.5.3. [46, 79]

To perform the pairwise comparison for ρ_{xy}, the relationships scores for each participant were ordered into a single vector. Both the linear and quadratic coefficients were included, but experts who provided linear relationships had the quadratic coefficient set to zero. Each unique pair of experts were compared against each other. An example of the comparison in graphical form is shown in Figure 45. Each point is the same relationship with the normalized score from one expert as the X value and the normalized score from the other expert as the Y value. This particular example has a correlation between these individuals of 0.8265.

These tests can be performed on the various subgroups defined by the demographics and the interface mode as well as any groups formed by the filtering described next. This example shows to individuals compared against each other, but the same process is used for comparing an individual against the truth model.

5.4.2 Filtering Responses

The large number of participants who provided data means that many of them can be filtered out for different reasons to whittle down to an elite group of experts. This group serves as

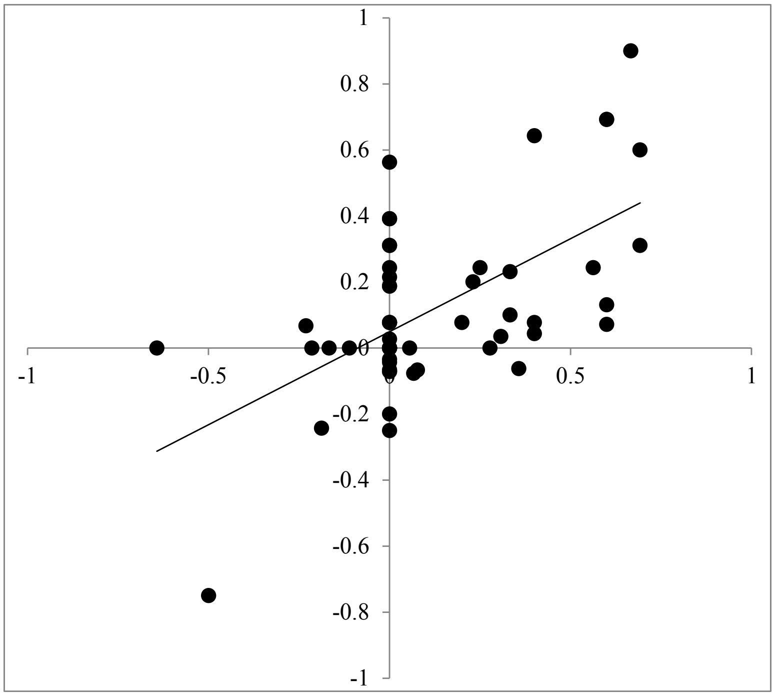


Figure 45: Graphical Depiction of Pairwise Correlation of Relationship Scores

the most ideal set of individuals to demonstrate the capabilities of this method. Individuals can be filtered out from the group for a number of reasons. One obvious one is if they disagree significantly with the group. While it is likely that outliers are less accurate, it is not guaranteed. Considering the small proportion of participants who have more than four years of experience with civil airliner performance, it is possible that the outliers are the only ones who are correct. Outliers would be identified by having a much lower average correlation compared to all other experts than the overall average of correlations.

It is also possible to filter out individuals based on how well they match individual known relationships. Some amount of information may be available about particular relationships prior to creating the expert-based model. For example, some relationships may be known to be positive or negative. Participants who provide a score with the wrong direction could be removed. The danger with this is that an expert may be removed for an area they know very little about while the rest of their data is in areas where they have much more expertise. Eliminating such an individual may hurt the overall quality of the model in order to satisfy a single relationship.

The data from the truth model can be used to cheat slightly by eliminating individuals who are very different than the true relationships with the understanding that such an elimination would not be possible normally. This is useful especially to prevent a small number of individuals from disrupting the ability to perform other tests and analysis.

5.5 Use the Model

Using the model typically refers to packaging it and applying it to the problem that motivated its development. For this research, using the model refers to using it to generate results for the purpose of comparing against truth data. The simplicity of the model means that a large number of cases can be run very quickly. In order to guarantee that the comparison of any of the models to be tested are as thorough as possible, the same design points used with the truth model in the next section will be calculated here.

The input data must be slightly modified in order to work with the expert-based models since they are based on scale from -1 to +1 for each variable at each level rather than the true united numbers used. It is relatively common and well-accepted to code designs of experiments using this scale [133]. The values for the system requirement and intermediate metrics were calculated from the combined expert-based models and the coded design of experiments using a series of MATLAB scripts wrapped around the `bsxfun` function to automate the process.

The outputs of the model are tested using Spearman's rank correlation coefficient r_s to test how similar the rank ordering of points are for each system requirement. Pearson's product-moment correlation coefficient ρ_{xy} are applied the same way to test how similar the outputs of the expert-based models are to the truth model in terms of proportionality. And lastly, the accuracy of the estimated ranges are considered and, if a consensus can be drawn from them applied to the values for the system requirements and intermediate metrics to compare how well they approximate the true values.

5.6 Truth Models

Of the three types of expert-based methods, only predictive methods typically have any sort of validation associated with them. Judging methods are tested for consistency between judges and relationship-based methods are tested primarily by the change in quality of the resulting products. In order to properly test the method, the relationship models produced by it must be compared against this trusted model. It is important to test both the accuracy of the outputs of the model as well as the accuracy of the individual relationships. This need drives a need to have suitable data to compare against. This means having a source of information that is trusted as a point of comparison that is similar in nature and behavior to what is desired and that is readily available.

Identifying and using such a source of information is split into two steps here: finding a trusted source of information and manipulating that information source into the form needed for comparison. The term “truth model” refers to a model that is considered to accurately capture the actual relationships present in the system. It is used as the authoritative answer to compare a new model against. While not part of the method, a truth model is used to test the method.

The types of problems where the proposed method is best applied are those where there are no existing models readily available or that can be used quickly enough. This is the reason that validation is performed without the use of this information. However, within the scope of this research, it is important that the method be tested against a trusted external baseline. This baseline serves as a point of comparison to ensure that the expert-based modeling process can produce models that match with true physics. When the baseline is another model, it is referred to as a “truth model” as a way to recognize that it is being treated as absolute truth.

5.6.1 Aircraft Performance Truth Model

There are several attributes of interest in selecting the truth model. It must be reliable and trusted and it must be available to the author. It should also be easy to use to minimize the chance of producing untrustworthy results. Ideally it would also be flexible enough to be reused for each of the experiments for whichever example problems had been selected. The Flight Optimization System (FLOPS) maintained by Arnie McCullers of NASA Langley meets all of these requirements [92]. If the choice of truth model was a central tenet of the proposed methodology, further investigation into other options would be necessary. In this case, it's akin to selecting the glassware for a chemical experiment or the type of pressure ports used in a wind tunnel experiment. So long as the equipment used is sufficient, it is unnecessary to compare all possible choices.

The ease of use of FLOPS has enabled is use in academic design problems as a teaching tool [89]. At the same time, it is also trusted to be used to support research that eventually helps drive federal and international regulations [74]. It can model a wide variety of aircraft from the civil airliners used here to supersonic transports and military fighter jets [25, 90]. It can be modified and run quickly, minimizing the time spent creating and validating models instead of testing the method. It is an excellent example of a model that, if it were available, precludes the need for expert-based modeling.

Since cost is typically poorly modeled, but an essential part of any design activity, the Aircraft Life Cycle Cost Analysis tool (ALCCA) was used for a simple historical weight-based cost estimation model [88]. ALCCA has the capability to include complex process-based cost estimation. However, this information requires significantly more detailed information than would be available at an early stage in the design process and would require a field of expertise not readily available outside of commercial aircraft manufacturing. Since this information would also likely be well-protected and proprietary, it would be difficult to gather sufficient information to test this aspect of a model. As discussed in Section 2.2.3, such process-based models are outside the scope of this research but would serve as an excellent area of future development. Only the simplified, but historically accurate, weight-based methods will be brought forward.

The design problems shown here were selected because they were examples of well-understood problems that had a large number of experts available to provide information for this research as well as because of the availability of truth models to support them. This makes them ideal as an example problem for testing a new method. A civil aircraft design problem is an unlikely motivating problem to require expert-based design outside of original research.

5.6.2 Aircraft Engine Truth Model

In order to capture the effect of fan pressure ratio on the thrust-specific fuel consumption, an engine cycle model was needed. Full engine analysis codes were considered, such as the Numerical Propulsion System Simulation (NPSS) maintained by the NASA Glenn Research Center. Complex tools such as this are highly accurate, but also very sensitive and dependent on a well-calibrated model of a specific engine. Since the aircraft being modeled is a notional one, no specific engine was selected. Instead a rubberized parametric engine that was suitable for this class was ideal and still valid as shown previously by Bullock and Niedzwiecki [24, 98]. A reduced-order MATLAB-based tool created by Lee, Nam and Perullo was identified as a likely fit [82]. This model accepted the inputs and ranges shown in Table 13 at a design point of Mach 0.8 at an altitude of 35,000 ft. The engine covers a similar thrust class as the Pratt and Whitney PW4000-94 and General Electric CF6-80C2 that were used on the Boeing 767-200ER, an aircraft similar to the notional aircraft.

FLOPS scales the engine thrust to meet its required thrust-to-weight on its own and the only other variable included in the model was fan pressure ratio. On the advice of one of the authors of the parametric model, the high pressure pressure ratio was fixed at 18 and the other metrics were held constant at their mid-level values with the exception of the low

Table 13: List of Parametric Engine Deck Model Inputs and Ranges

Input	Minimum	Maximum
Extraction Ratio	1.0	1.2
Fan Pressure Ratio	1.5	1.7
High Pressure Compressor Pressure Ratio	18	22
Low Pressure Compressor Pressure Ratio	1.2	1.6
Max Turbine Inlet Temperature (deg R)	3200	3600
Max Sea Level Static Thrust (lbf)	50000	80000

pressure compressor ratio. The low pressure compressor ratio was adjusted such that the overall pressure ratio was constant no matter what value was selected for the fan pressure ratio.

The weight of the parametric engine was calculated using a separate parametric regression around a General Electric GE90 with the same design parameters and ranges as the parametric engine deck model.

5.6.3 Linearization and Regression of Truth Models

The truth models as they are exist primarily as “black boxes”. That is, they produce a set of outputs given a set of inputs. This makes it easy to calculate the performance of a single set of inputs or a population of sets of inputs, but makes it very difficult to see the relationships between the inputs and the outputs from the raw data. Regression methods are a common approach to reduce such sets of raw data into recognizable and simplified trends. It is these trends that this method is attempting to capture, so finding these trends from the trusted models is necessary to test the individual accuracy of the model components.

The most difficult part of this step was identifying the proper measures and variables to use at each level. This process is discussed in detail in Section 5.1.3. Using the variables and ranges determined there, the relationships were identified through regression.

The engine weight and engine performance models were linked with FLOPS within a Phoenix Integration ModelCenter environment. The full assembly was used to run a 5000-case Latin hypercube design of experiments. This data served as the truth data when specific data points were needed. When model coefficients and relationship of the truth model were needed, a series of regressions of this data was used. Depending on the specific need the regressions were a simple linear regression or a second order polynomial regression with no cross terms. The regressions were performed at each level to mirror the type of relationships the experts were providing. One set of models were created for the intermediate metrics as a function of the design variables and another set of the system requirements as a function of the intermediate metrics using the 5000 cases as training data.

These regressions demonstrate the best-possible fit one could hope for from the experts, but also demonstrate the level of accuracy that is possible using only a two-level series of linear models. The coefficient of determination (R^2) was calculated for each of the system requirement models when stacked with with the models for the intermediate metrics and is shown in Table 14.

Table 14: Accuracy of Layered Linear Truth Models

Model	R^2
Approach Velocity	0.970
Takeoff Field Length	0.935
Operating Empty Weight	0.961
Block Fuel Weight	0.874
Acquisition Price	0.934

While the results are not perfect, they are very high for most of the fits considering that the relationship tested (system requirements as a function of design variables) was not regressed directly. This truth model, combined with the information gathered from experts serves as the basis for the analysis and discussion presented in the next chapter.

CHAPTER VI

DISCUSSION OF RESULTS

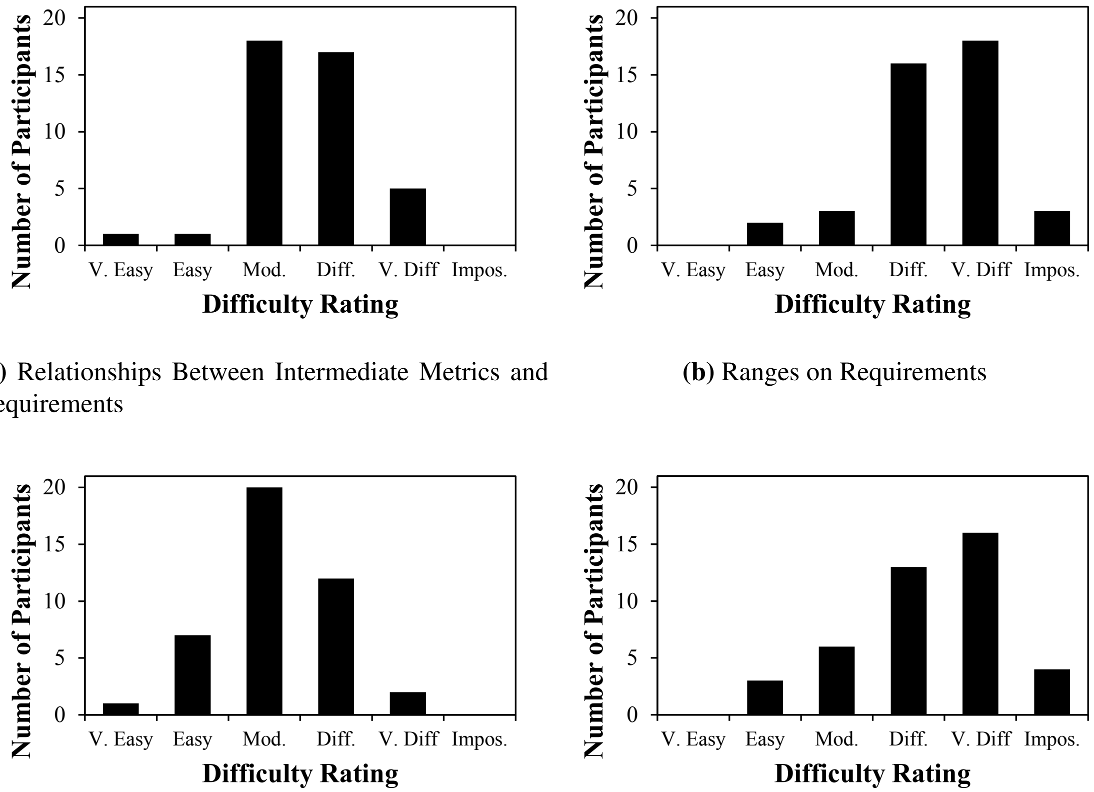
Chapter 3 describes the origins and reasons for the ALTER methodology that were tested according to the procedure in Chapter 5. This chapter discusses the results of that process to evaluate the usefulness of this approach.

6.1 Feedback from Participants

The most immediate information available after gathering data from participants was the feedback and opinions of those individuals. This came in the form of written comments included as part of the data collection tool and verbal comments to the researcher afterwards.

As part of giving data, experts were asked to report on how difficult it was to give the relationships and the ranges at each of the levels, for a total of four different areas to rate in difficulty. For each of these options, there were seven possible options: Very Easy, Easy, Moderate, Difficult, Very Difficult, Impossible and the option to leave no response at all. Figure 46 shows the distribution of scores for each of those.

In general, participants found it slightly easier to provide the relationships between the design variables and the intermediate metrics than between the intermediate metrics and the system requirements. There are several possible causes for this. The first is that the lower level relationships are closer to well-understood physics where there are clear physical connections between design variable inputs and the intermediate metrics. The higher level required a greater level of abstraction. For example, the relationship between wing area and wing weight is easy for even a lay person to estimate. At the higher level, translating wing weight to approach velocity requires the expert to consider what design variables changed in order to cause the wing weight to change and how they would also



(a) Relationships Between Intermediate Metrics and Requirements

(b) Ranges on Requirements

(c) Relationships Between Design Variables and Intermediate Metrics

(d) Ranges on Intermediate Metrics

Figure 46: Experts' Ratings of Difficulty Providing Information

affect approach velocity. [1]

Even in a perfect computational model, each additional level of abstraction adds some noise. In this case, the additional noise must be both identified and accounted for by the expert. This is part of why the lower level relationships in the first preliminary experiment were so much crisper than the higher level relationships.

Another source of difficulty is that experts must rely on their own estimates for ranges of the intermediate metrics. Experts also expressed difficulty in estimating the ranges for the intermediate metrics which likely impacted their confidence about the higher level relationships.

[1] Unlike technology *k*-factors, the values of intermediate metrics can only change as a result of lower-level parameters and are not varied independently.

Many expert-based methods also ask experts to provide some measure of their confidence in their answers. This is commonly used as an additional factor in adjusting how strongly their information is weighted when combined with others. Other methods use confidence to estimate a standard deviation around the individual's estimate to capture the most likely region and a level of certainty associated with it. While participants were not asked for their confidence explicitly, their difficulty ratings can offer some insight into how confident they felt. Those who felt that a section was easy are likely more confident about their responses than those who found it very difficult. [35]

Participants found it more difficult to provide estimates for the ranges at both levels than to estimate the relationships. Though 3-4 individuals rated the ranges as impossible to estimate, only two individuals left those portions unanswered. There is a slight difference suggesting that participants found it easier to estimate the lower level, but not significantly so.

Many participants provided informal verbal feedback about their opinion of the process and their own confidence. Some themes were common across most participants' comments. Almost universally, participants stated that the process was quite challenging and that it was an excellent test of their knowledge and intuition as an aerospace engineer. Several went further to suggest that it should be incorporated into a classroom environment as either a test or a demonstration of how little a student truly knows. Others commented that it was difficult to provide some of the relationships despite knowing relevant equations and trends. The difference was that most of those equations and rules of thumb don't include trades between design variables and frequently relate to changing one variable at a time.

Participants were generally confident of the relationships they provided, but had very little confidence in their estimates of the ranges for the intermediate metrics and system requirements. Participants were encouraged to provide estimates even if they were not confident to facilitate later analysis. Several individuals admitted that they had no idea at actual values and had just guessed. The verbal feedback combined with the difficulty ratings for the ranges explains the poor accuracy of the ranges, discussed in detail in Section 6.6.

6.2 Agreement between Experts for Relationship Estimates

The *Test the Model* step in the modeling process provides for a way to validate a model by using co-validation. This co-validation can be performed on the relationship estimates that form the model parameters or on the outputs of the combined models. The comparison of the model parameters is more stringent and the comparison of the outputs is generally more forgiving. For the sake of detail, the tests here will be on model parameters and model outputs will be analyzed in Section 6.5.

There are several measures of how similar expert ratings are to each other, generally known as inter-rater reliability. Section 3.5.3 discusses several of these and narrows them down to two most applicable metrics: Pearson's product-moment correlation coefficient for pairwise comparisons and Krippendorff's alpha for group comparisons.

6.2.1 Pearson's Product-Moment Correlation Coefficient

To perform this comparison, the relationships scores for each participant were ordered into a single vector. Both the linear and quadratic coefficients were included, but experts who provided linear relationships had the quadratic coefficient set to zero. Each unique pair of experts were compared against each other. An example of this for two experts is shown in Section 5.4.1.

For 42 participants, there is a total of 861 unique pairings. This large number makes it difficult to analyze these individually. A heat map was created to quickly identify trends and groupings in a visual fashion and is shown in Figure 47. Pairing with a very low correlation are shown in bright red while perfect correlation is shown in bright green. Black is in between the two extremes. There is a line of perfect correlation along the diagonal where each expert is perfectly correlated with themselves. The plot is symmetric about the diagonal since correlation is commutative. Arbitrarily assigned row numbers are listed on each edge.

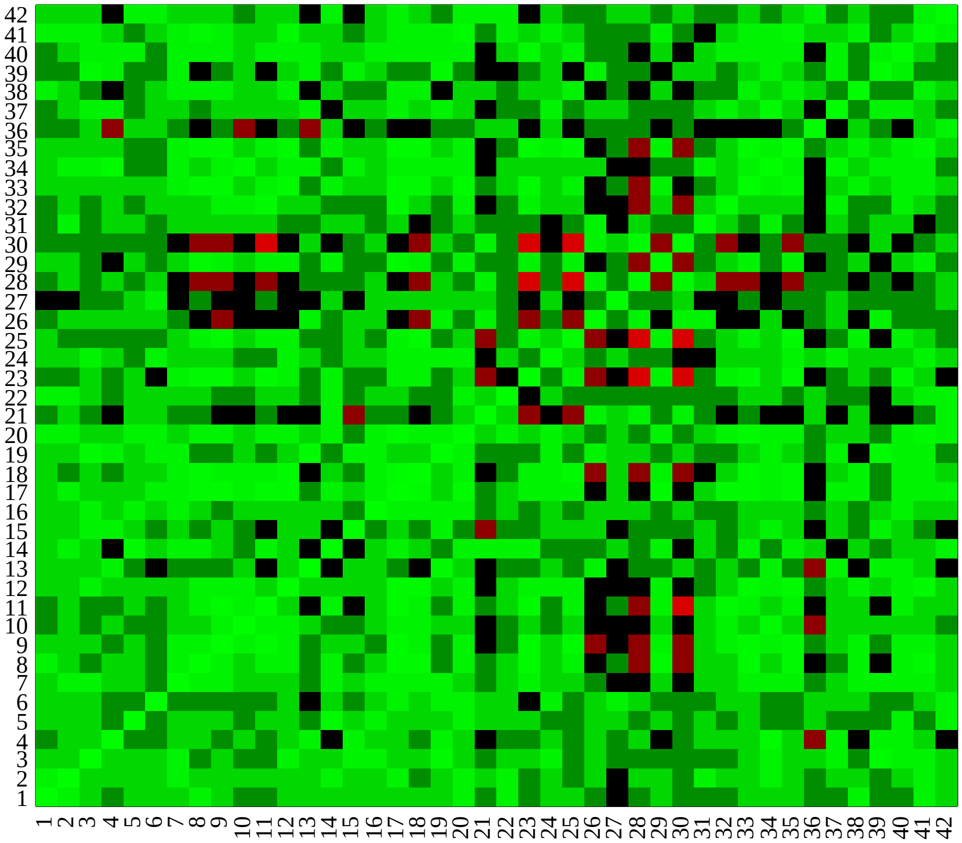


Figure 47: Pairwise Correlation Comparisons Between Experts

There are a small number of individuals who appear to be poorly correlated with most other experts. Rows 21, 26, 28, 30 and 36 have more dark green, black and red values than the others. Rows 26, 28 and 30 agree very well with each other which suggests they're either all wrong together or potentially all correct together. It is tempting to eliminate these individuals as outliers, but since they have clustered together, there is a possibility they are uniquely experienced in a certain aspect of the problem. Rows 21 and and 36 are not highly correlated with any other experts and are more likely candidates to be removed as outliers.

The heat map is useful for identifying trends, but it is difficult to get a good feel for the distribution and range of correlations present. Figure 48 shows the distribution of all 861 pairwise comparisons. A normal distribution has been fit to the histogram with a mean of 0.3179 and standard deviation of 0.2084. Note that the pairwise correlations here are not truly normally distributed since they are skewed slightly with the tail to the left and since they are limited to values between -1 and 1 rather than from -∞ and ∞ like a true normal distribution. However, the approximation is close enough for the sake of a simple intuitive analysis.

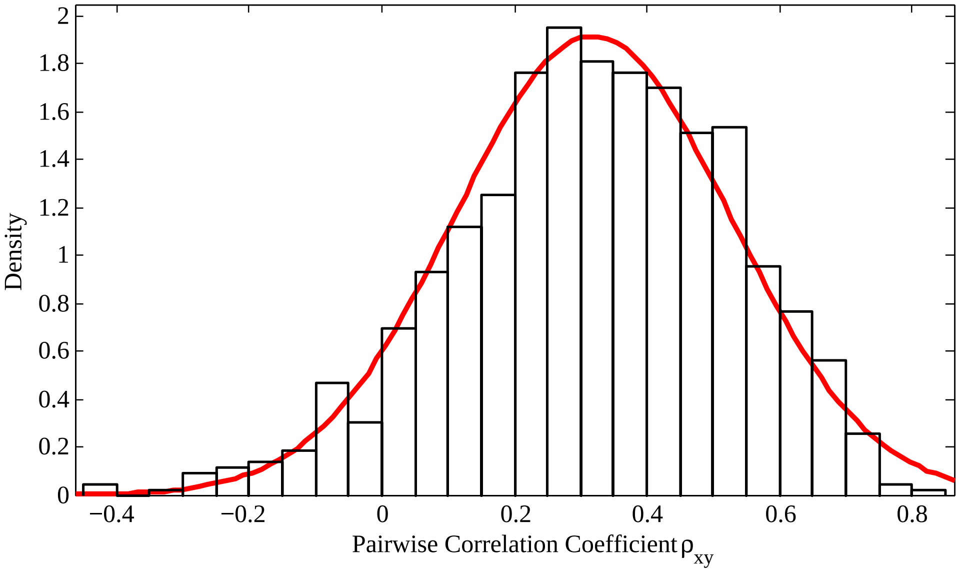


Figure 48: Distribution of Pairwise Correlations Between Experts

An ideal distribution of scores would have a much higher mean in the 0.90 – 0.95 range with a smaller standard deviation. This would correspond to all experts agreeing to a very high degree and any variation between experts being more consistent. Not all experts are equally capable. It is expected that, even with a high mean overall, some of them would be different and produce a small second bump at a lower agreement. This would be readily identified as a group of outliers and considered for removal from the data set used to create the models. In this data, a small group is visible, but this group is at the far left tail around the -0.4 point. This bump corresponds to the rows mentioned previously in the heat map.

This distribution of ranges is not encouraging. However ρ_{xy} has a high penalty for outliers and disagreements between a small proportion of the relationships has a strong negative effect.

6.2.1.1 *Effect of Interface Mode on Agreement*

Hypothesis 1a claimed that participants who used the QFD scale with only seven possible scores would have the greatest agreement between them. The theory is that experts should be more likely to pick the correct value if there are fewer subtle choices to select from and only large steps. In order to test this, each group of experts is separated by interface type. The average of all unique pairwise correlations for that interface is calculated and compared against the same metric for the full population, as shown in Table 15. For further analysis, the calculations were performed for just the relationships between design variables and intermediate metrics, just the relationships between intermediate metrics and system requirements, and all of the relationships together.

Table 15: Agreement of Experts by Interface Type Measured by Average ρ_{xy}

Mode	All Relationships	High-Level Relationships	Low-Level Relationships
All	0.3501	0.3568	0.3575
1: QFD	0.3365	0.3608	0.3302
2: Integers	0.3679	0.3537	0.3973
3: Graphical	0.3454	0.3546	0.3437

All of the average ρ_{xy} of all combinations fall between 0.3454 and 0.3973. While the QFD did have a slightly higher agreement for the high-level relationships, it was lower for the low-level relationships and overall. Even if it were higher, the size of the differences and the distribution of all scores means that it is not statistically significant at any reasonable α confidence level. *These results do not support Hypothesis 1a.* It is likely that the difference in possible scores of the scale had a significant effect on poor correlations. Even though there were fewer choices and experts were more likely to agree just by chance, there was a greater penalty associated with disagreeing since, for some cases, a single disagreement would shift the raw score from a 9 to a 3.

The average ρ_{xy} for Mode 2 has a greater difference from the overall average than any other for both the low-level relationships and all relationships. Mode appears to have the greatest effect on the agreement among low-level relationships and very little impact at the higher-level relationships. While these observations are interesting, it is important to note that they are still not statistically significant observations due to the high variance of ρ_{xy} within the total population and within each group.

6.2.2 Krippendorff's Alpha

The limitation using Pearson's ρ_{xy} is that it can only be performed between two items at a time. Krippendorff's alpha is a better measure of how a group agrees overall. It was calculated for all experts for all relationships as $\alpha_k = 0.3372$. One of the major downsides of this metric is that it does not scale well for large populations. That calculation took approximately two hours to calculate on a computer with an Intel Core i7 3.4 GHz processor. The calculation of all pairwise ρ_{xy} for the same set of data took less than a second. This time requirement made exploring potential relationships among different combinations of groups was impractical and so this method was used sparingly for specific tests.

One such test was the same test of Hypotheses 1a. Table 16 shows the values of α_k for each mode as well as the full population as calculated for all relationships, just the higher-level relationships, and just the lower-level relationships. The layout mirrors that of Table 15.

Table 16: Agreement of Experts by Interface Type Measured by Krippendorff's α

	All	High-Level	Low-Level
Mode	Relationships	Relationships	Relationships
All	0.3372	0.2863	0.3149
1: QFD	0.3229	0.2700	0.2907
2: Integers	0.3638	0.3226	0.3605
3: Graphical	0.3255	0.3361	0.3165

The impact of scale and interface on agreement is clearer for α_k. In this case, the QFD scale, has the lowest agreement in all three sets of relationships. For both the high-level and low-level relationships, the difference is larger than for all agreements. Considering the difference in values, *this suggests rejecting Hypothesis 1a.* Statistical significance of these findings is not available as α_k does not follow a standard distribution or statistical test. It is possible to find significance using bootstrapping, but the combination of the number of values of α_k required and the time to calculate each one would require weeks of constant compute time.

The integer mode continues to have a higher degree of agreement with the lower-level relationships when measured by α_k. And in this case, the higher-level relationships generally have a poorer agreement than the lower-level relationships, which is in line with what was expected from the distributions of difficulty ratings.

6.2.3 Use of Quadratic Relationships

The graphical interface allowed participants to give a nonlinear relationship wherever they deemed it appropriate. Another test of how well the participants agree is whether or not a consensus is present about which relationships required a quadratic form rather than a linear one.

Of the 14 people who used the graphical interface, only seven elected to provide any nonlinear relationships. Of those seven, four provided only a single nonlinear relationship out of 55 possible relationships. Two identified three relationships as nonlinear and one identified six. There was no consensus about which relationships required a quadratic relationships since they were all used in different locations. No relationship had more than two experts agree that the relationships was sufficiently nonlinear to require describing it as such.

The low agreement with the use of quadratic relationships may be partly due to the problem or the general level of expertise. Most of the relationships could be adequately described with linear relationships. If the problem had one or two relationships which could not be described with a linear relationship, there might be more agreement at those locations. There was a weak trend with those with more years of expertise being more likely to use a quadratic relationship.

6.3 Accuracy of Individual Relationships

The previous tests looked only at how similar the scores experts gave were to those of other experts. This is a measure of agreement, but not a measure of accuracy. If all experts share similar misconceptions, they will agree on an incorrect value for the relationship. Depending on the particular problem, that may be the extent of what information is available. As part of this research, a truth model is available to test the accuracy of their relationship values. This section compares the estimates for responses provided by participants against values from the truth model for several selected relationships. Section 6.4 analyzes the accuracy of the relationships as a group.

There are two points of comparison from the truth model: the correlations of the relationships and the coefficients of the relationships. For lower-level relationships these are nearly identical. For the higher level relationships there is a greater degree of variation.

A few examples are shown here to give a feeling for different types of agreement between individual experts and particular relationships. Figure 49 shows the distribution of scaled scores of all 42 participants for the relationship between intermediate metric C_{D0} and design variable t/c_{avg}.

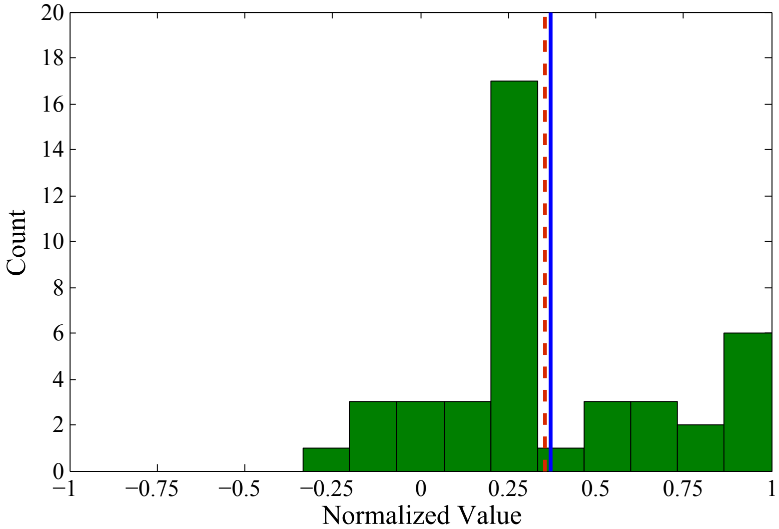


Figure 49: Distribution of Scaled Scores of C_{D0} vs. t/c_{avg} with Truth Data

For this relationship, the average normalized score is 0.394 with a standard deviation of 0.347. While not universally agreed upon, there is a clear mode slightly lower than the average. The plot also shows the correlation of the truth data for this relationship as a solid blue line and the coefficient of the truth model that corresponds to this relationship as a dashed red line. These are 0.372 and 0.355 respectively. For this relationship, the average of the experts matches with the truth data very well despite the spread in the scores amongst the participants.

Figure 50 shows the distribution of scaled scores of all participants for the relationship between the intermediate metric W_{wing} and the design variable S. For this relationship there was a very high degree of agreement between experts with a mean normalized score of 0.973 and a standard deviation of 0.103, entirely due to a few experts with differing opinions. The truth data correlation of 0.426 is shown with a solid blue line. The corresponding truth model coefficient of 0.417 is represented with a red dashed line.

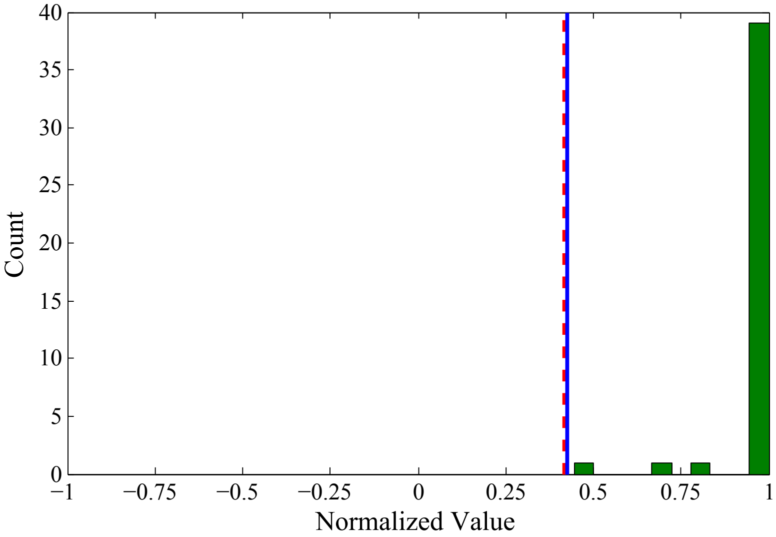


Figure 50: Distribution of Scaled Scores of W_{wing} vs. S with Truth Data

For this relationship there is an exceptionally tight agreement between participants, but this agreement does not match up with the truth data. While the wing area is a major component of wing weight, the weight of the internal structures of the wing are also heavily influenced by the thickness and aspect ratio of the wing. A long, thin wing has a higher span-wise moment arm and a much smaller moment of inertia for the structural spar's cross-section. In order to compensate, the spar must be reinforced, adding to the weight of the wing. This relationship is an example where a structures background would have a benefit towards a better interpretation of these.

Figure 51 shows the distribution of scaled scores for the relationship between the system requirement V_{app} and the intermediate metric W_{wing}. The scores here have a mean of 0.211 and a standard deviation of 0.676. While there are some scores in the center, the distribution approaches a multimodal distribution at the extreme positive and negative scores. As before, the coefficient of the truth model is shown with a solid blue line at a value of -0.238. The correlation of the truth data is shown as a red dashed line at -0.356.

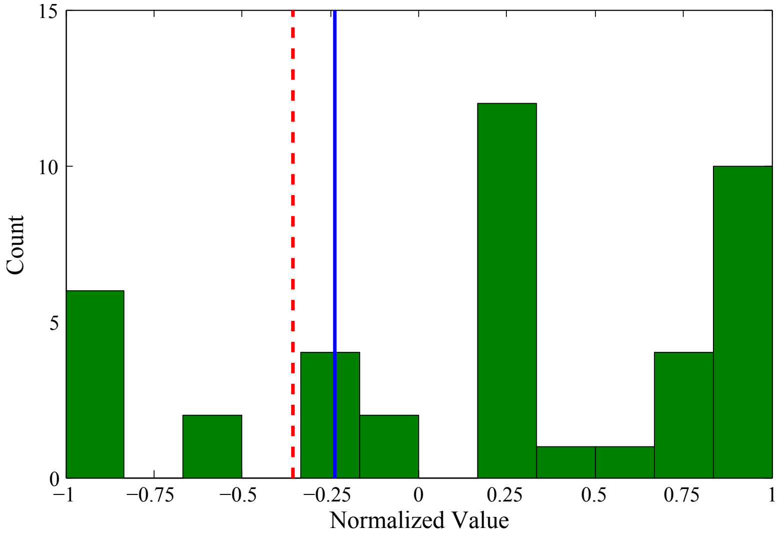


Figure 51: Distribution of Scaled Scores of V_{app} vs. W_{wing} with Truth Data

This is an example of when there was no agreement. There are two interpretations of this relationship that tend to push it toward either the positive or negative extremes. The first is that heavier wing, independent of other variables is usually most indicative of a heavier aircraft. This is trend is common enough in production aircraft that some initial sizing methods estimate operating empty weight of an aircraft by applying a scaling factor to wing weight based on historical data. A heavier aircraft overall will require more lift to stay in flight during approach. More lift is created by flying faster, which results in a higher approach velocity. This interpretation is the result of treating the weight of the wing as totally independent of other aspects of the wing's design as would be typical of how it would be treated as a technology *k*-factor. Since technology is held constant for this design, the change in wing weight would be a result of adding lead or other ballast to the wings or using less structural materials than is desired for safety.

The second interpretation is that a heavier wing weight is the result of having a larger wing due to the previously identified relationship between wing area and wing weight. All else held constant, a larger wing generates more lift at a given speed than a smaller wing. Since the aircraft is capable of generating the needed lift at a slower speed, it can approach at a lower velocity.

Individuals between the extremes are trading between the two interpretations or do not believe that wing weight was the dominant influence on approach velocity. Both trends are true in part, but the truth model demonstrates that the second one is the dominant characteristic. Participants were instructed that intermediate metrics change for a reason, but considering the amount of information that was delivered and the awkwardness of a new interface and a difficult problem, some of that information may not have sunk in or individuals just didn't take it into consideration.

6.3.1 Using Known Information to Support Experts

For some design problems, partial knowledge about the problem is available. This information may be known to some individuals, but not to others or may not have been included in the thought process of all individuals. Including this information explicitly may improve the accuracy of the resulting relationship information. A small, simple test was devised to test the impact of providing additional information using the approach velocity versus wing weight relationship above. A small number of individuals were asked to circle an integer from -9 to +9 and were given the following information.

Approach velocity is correlated to the weight of the vehicle and the lift provided by the wing and can be simplified to

$$V_{app} \propto \sqrt{\frac{W_{landing}}{S}} \qquad (19)$$

The weight of the wing is about 10-15% of the total weight of the vehicle at landing.

There is a strong positive correlation between wing area and wing weight.

When scaled and normalized based on existing information, the scores given by participants are as shown in Figure 52. Because this was not a full data collection exercise, normalization was done using average values of other relationships from existing data. This new data has a mean of -0.294, fitting nicely between the coefficient and correlation of the truth model at -0.238 and -0.356 respectively.

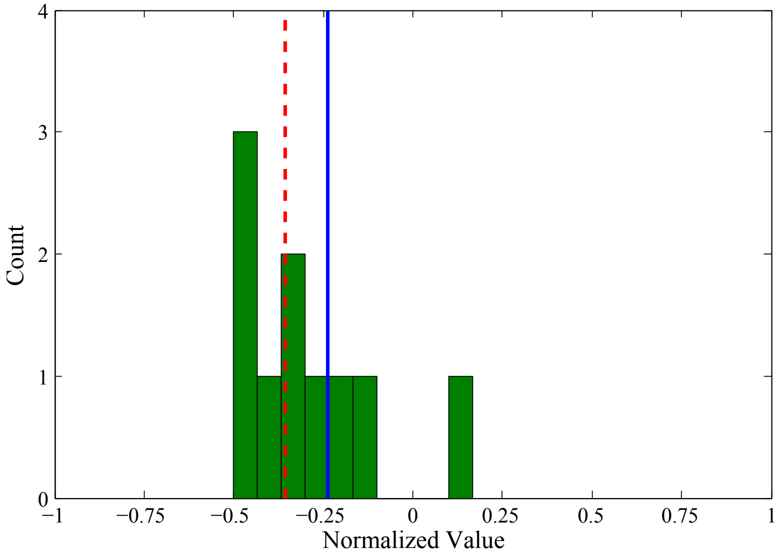


Figure 52: Distribution of Scaled Scores of V_{app} vs. W_{wing} with Truth Data from Experts with Additional Information

This is a significant shift from above and much closer to the true behavior. Individuals responding to a single relationship with information spent more time than individuals spent per-relationship in the full data collection and put more effort into making the necessary trades between the contributing factors in their head. It is likely that both having additional information and spending a greater deal of mental effort on the relationship contributed to more correct estimates. If such information were available for the problem at hand, it would be helpful to include it where appropriate. Care must be taken not to bias participants toward the moderator's view or to overload participants with too much information.

6.3.2 Using Known Information as Filter of Experts

For some problems, some relationships may be known with certainty due to physical relationships and past proven knowledge. As an example, the simplified definition of C_{D0} is shown in Equation (20). This definition is true no matter for all fixed wing aircraft. We can use this information to gain insight into what the relationships for C_{D0} should or must be.

$$C_{D0} = \frac{D_{parasite}}{\frac{1}{2}\rho V^2 S} \tag{20}$$

The velocity V at cruise is defined based on the cruise Mach and the cruise altitude given in the IRMA. The value of density ρ is also based solely on altitude. The only remaining variables are the parasite drag $D_{parasite}$, which includes all drag not due to lift, and the wing area S. The parasite drag is a function of the wing area. A larger wing generally has higher drag. For the relationship between C_{D0} and S, both the numerator and denominator will increase at the same time. Therefore, the rate of change, and the value of the relationship, will be positive if drag increases faster than wing area and negative if it increases slower than wing area. Physics and past experience tells us that the latter is true and that a larger wing tends to be more efficient.

Figure 53 shows the distribution of all normalized relationship scores provided by participants. It also shows the normalized truth values. The truth model coefficient is shown in a red dashed line while the truth model correlation is shown a blue solid line.

Both of the estimates of the truth data are negative, as expected from the equation above. Most of the relationships scores are positive and the greatest concentration is near

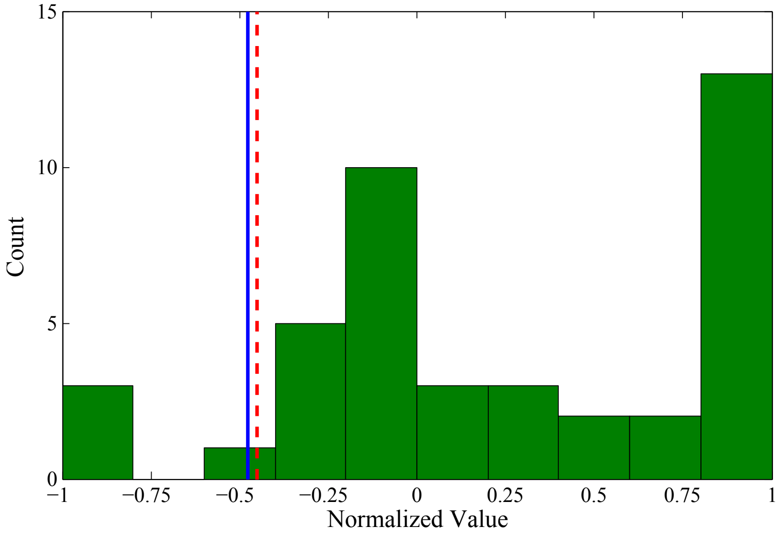


Figure 53: Distribution of Normalized Scores of C_{D0} vs. S with Truth Data

the maximum possible value. Based on this information, all of these experts are wrong beyond a difference of opinion. It is more difficult to identify whether this is a sign of overall bad performance or a single isolated error.

The thought process that leads to a strong positive score is due to the difference between what C_{D0} means and what most aerospace engineers intuitively think it means. Much of aerospace engineering operates almost entirely with nondimensional coefficients rather than full dimensional numbers. Outside of force diagrams and measurements, C_L and C_D are preferred over lift and drag. Pressure over a surface is measured by C_p rather than actual pressure. For most calculations this simplifies the process and eliminates effects primarily due to scale rather than shape. This means that for many aerospace engineers, the term C_D *means* drag intuitively, even though they know its correct definition. And engineers intuitively know that drag increases with wing area.

This relationship highlights the importance of selecting the best possible variables. An alternative measure that would not have had this problem would have been the drag-area f, defined as $C_{D0} \times S$. This measure would have eliminated this problem, but is not commonly used and would have to be explained to participants. Some simple explanation can make sure that all participants are on the same page to minimize misconceptions. Too much explanation biases the experts responses to those of the moderator, which negatively impacts how representative the resulting model is of the experts understanding.

6.4 Similarity of Relationships in Expert-based Models to Truth Model

The previous section considered a single relationship at a time. This is helpful for examining why a particular relationship has a consensus or not, but individual relationships do not define the model themselves. The combination of relationships translate from a set of design variables to a set of system requirements. One expert may overestimate one relationship and underestimate another. A second expert may underestimate the first and overestimate the second. When combined together to produce a model they balance each other out. Looking at individual relationships only shows that neither of them is correct.

The similarity between individual experts or their combined models and the model parameters of the truth model is measured in the same fashion as individuals were compared against each other in Section 6.2. The next few sections compare the accuracy of the set of relationships scores for different classifications of the participants.

6.4.1 Effects of Education and Experience

The level of expertise of each participant was measured in part by the information they volunteered about their level of education and years of experience. While these were identified as less than ideal measures through literature, they are still useful pieces of information to have about the experts contributing to a model.

Hypothesis 2 predicted that those individuals with more than two years of experience with civil aircraft and those who had passed qualifiers or had a PhD would have greater accuracy than those who did not.

Figure 54 shows the values for the correlation of each individual's normalized relationships to the coefficients of the truth model. Each individual is represented as a black dot and is split up by education level. The education levels here are coded for clarity of the plot. The individuals coded as 3 have a bachelor's degree. Those coded as 4 have a master's degree. Those coded as 5 have completed PhD qualifying exams and those coded as 6 have completed a PhD. The uneven distribution of participants between education levels is discussed in greater detail in Section 5.2.2.

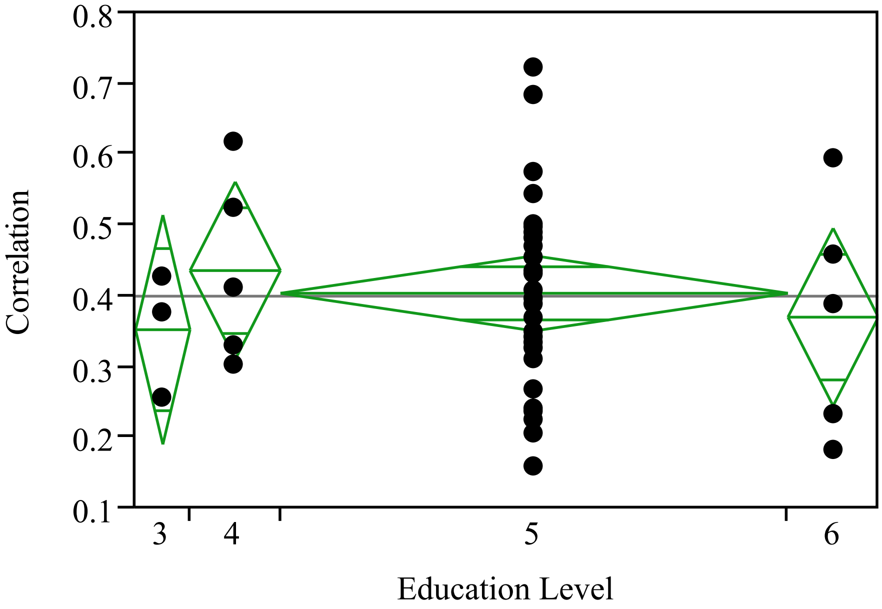


Figure 54: Impact of Level of Education on Similarity to Coefficients of Truth Model

Figure 54 also contains green diamonds around each group. The width of these diamonds is proportional to the number of samples in that group. This is useful to identify when points are overlapped and obscure the number of points visible. The height of each triangle is the confidence interval of the mean. In this figure, the large number of data points who have passed qualifiers allow for a much smaller confidence interval than the other groups.

Graphically, all four of the confidence intervals overlap the overall mean of the group (0.391) and each other. An F-test provides an F-ratio of 0.3143 corresponding to a null-hypothesis probability of 0.815. This value is nowhere near an acceptable α-level and so this data *does not support Hypothesis 2*. It should be noted that a more uniform sampling would provide tighter confidence intervals of the other levels of education.

Figure 55 uses a similar depiction as Figure 54 does to compare the amount of experience different participants had with aircraft design. The population mean remains the same as before and will be constant for this entire series of plots at 0.391. The participants are split into different groups ranging from 0–1 years of experience to more than 8 years of experience. There were no individuals who claimed to have no aircraft design experience.

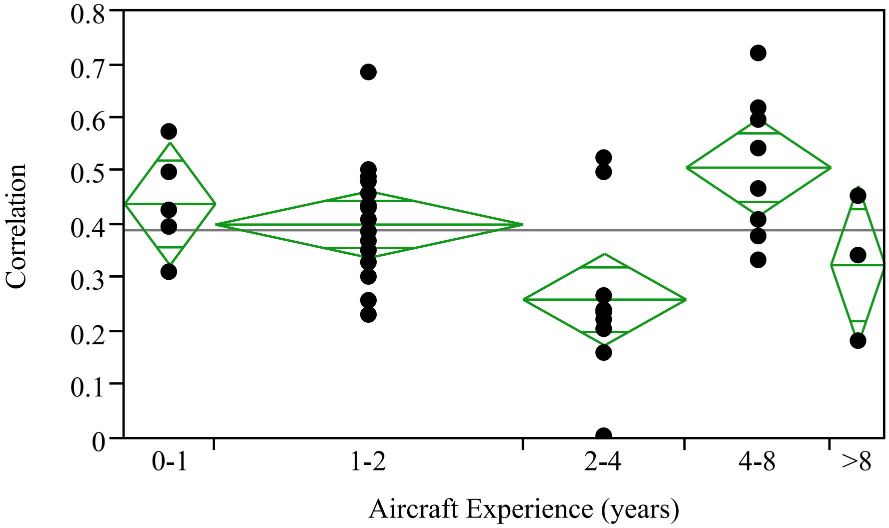


Figure 55: Impact of Years of Experience on Similarity to Coefficients of Truth Model

The distribution of scores and corresponding confidence intervals here is interesting. The two highest groups are those who have very little experience, and those who have 4–8 years of experience. The group with 2–4 years of experience had the lowest scores overall and low enough that it is statistically significant at the $\alpha = 0.05$ level. This research would have accepted significance at the $\alpha = 0.10$ level, but this particular comparison meets the more stringent degree of confidence.

The 1–2 year group is very close to the overall average partially because they comprise the largest group but also because that population in particular is much more sensitive to individual experience since it covers the widest range of ability levels. The 0-1 years of experience group may have performed better than average either because they were closest to the material and are still in aircraft design classes or because they realized their lack of intuition and so focused more on thinking through each of the relationships.

The few individuals with more than 8 years of experience may just be coincidental outliers from the population of all individuals with that much experience. However, it is also possible that some bias due to recruiting from an academic institution is at play here. Very few individuals were available with that much knowledge and those who would truly be considered experts are likely to have graduated and found employment elsewhere. It is also likely that individuals with this much experience have moved from technical roles to management roles within their respective research groups. That shift may have dulled their intuition about these trades.

Those with 4–8 years of experience are the group that most would consider skilled practitioners at a company. That is long enough to learn the ins and outs of a particular problem and still be actively involved with the technical aspects.

At first glance, it is unclear why the individuals with 2–4 years of experience performed so poorly compared to those with more and less experience. Even if the outlier near 0 was removed, the 2–4 are still statistically significantly lower than others. In many areas, beginners or those with less experience consciously put more effort into an activity and thus tend to do well. At some point, individuals get more comfortable with the activity and no longer feel the need to put in the same level of effort. While they do have more knowledge of the subject, they do not yet have the mastery and so their overall performance suffers. With time, the individuals reach that level of mastery and again excel. Such behavior is seen in US Army helicopter pilots. Individuals with 800-1300 hours of flight time (roughly 2–4 years of experience including a deployment) have the highest accident rates [131]. This behavior is mirrored in automobile accident rates where drivers aged 18–20 typically have the highest accident rates, roughly 2–4 years since earning their license.

It is believed that this behavior is also true for experts. That participants during this period of their career know just enough to feel confident enough to not consider the intricacies of secondary effects but do not yet know enough to have an inherent intuition about those effects. Further experimentation would be necessary to prove this conclusively. However, based on the existing analysis, *there is sufficient evidence to reject Hypothesis 2*.

6.4.2 Effects of Interface

Earlier it was shown that the type of interface had little effect on the agreement between experts, providing no support for Hypothesis 1a. Hypothesis 1b predicts that the greater detail, precision and presence of visual feedback associated with the graphical interface (Mode 3) will cause those individuals to provide information with greater accuracy.

Figure 56 compares the accuracy of individuals using each of the three modes. The confidence intervals around the means for each group are similar since this attribute was directly controlled by assigning individuals to a particular interface.

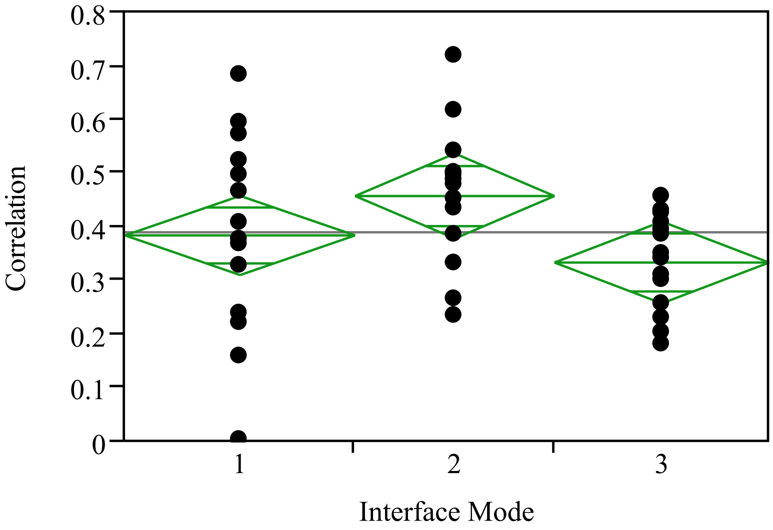


Figure 56: Impact of Interface Mode on Similarity to Coefficients of Truth Model

Interface Mode 1, the standard 7-value QFD scale, has a mean similarity almost identical to the overall mean. This suggests that the traditional approach serves as an excellent baseline for comparison. Because it is also the simplest interface, it is likely that individuals who are less comfortable with numeric methods would perform better with this approach then other more numerical approaches. And those that are comfortable with numeric data will do no worse for using it.

Interface Mode 2 used slide bars to allow participants to give scores at any integer between -9 and +9. Though it is close, the participants in this group performed sufficiently better than the overall mean to be statistically significant at the $\alpha = 0.10$ level. This mode had the benefit of offering users a greater degree of freedom and precision than Mode 1 did. When experts were unsure of a score, they could split the difference between the rigid choices the QFD scale provides. While the slide bars added a small degree of visual feedback, users had to focus on the numeric values of the relationships they were giving, which required additional effort and thought. Less thought is required to decide between a 3 and a 9 than between a 5 and a 6. This additional effort likely caused experts to think more critically when giving a score.

Interface Mode 3 was focused around a graphical interface and also allowed users to provide a quadratic relationship instead of the standard linear relationship. As shown above, this group had the poorest accuracy of the three interfaces. The mean is statistically significantly different than the population at the $\alpha = 0.10$ level. The difference between this mode and Mode 2 is significant at an even stricter level. Only a few individuals used the quadratic interface and those that did, did so sparingly. While that may have had some effect, the proportion of nonlinear relationships is too small to have had much effect. The linear interface used the same integer scale as Mode 2, so the only difference was the presence of graphical feedback in the fashion of a prediction profiler.

As mentioned in Section 5.2.4, the researcher returned to check on participants approximately 15 minutes after finishing giving instructions. This provided an opportunity to observe participants while they were using the different interfaces. Those using Mode 3 tended to focus much more on the graphical depiction than on the numerical value. While that was useful for looking at the comparisons between adjacent relationships at a glance, humans are extremely poor at estimating slopes of lines without the use of grids. The result of this was that participants with mode three were more likely to provide information based on what looked right rather than making a well thought-out decision about a particular score.

Due to the poor accuracy of participants using Mode 3 and the improved accuracy of participants using Mode 2, *this evidence rejects Hypothesis 1b*.

6.4.3 Effect of Filtering Experts Based on Individual Relationships

Section 6.3.2 suggested the notion of filtering out participants who failed to give correct values for a known relationship. Is was also established that the relationships between C_{D0} and S should be negative. Figure 57 shows the distribution of those individuals who correctly identified that relationship as negative and those who identified it as a positive relationship. The trend shows very clearly that the population of individuals who provided the incorrect sign were less accurate than those who provided the correct sign. The difference in the means is statistically significant at the $\alpha = 0.01$ level and is the strongest relationship observed thus far. This difference is also larger than the effect of this single relationship. If this relationship were the only one that was wrong, the means would be much closer together.

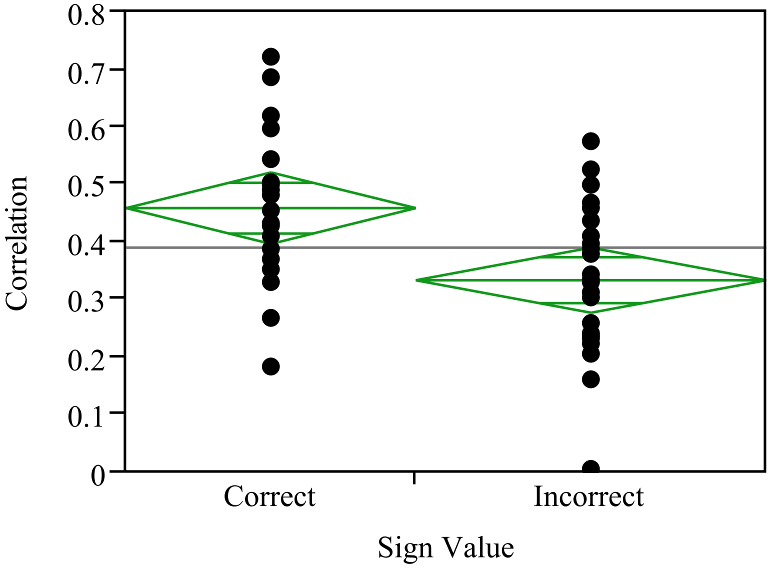


Figure 57: Relationship between Similarity to Truth Model Coefficients and Correct Sign of C_{D0} vs. S

Though this trend is present for the means, it is not a complete step function between the two groups. There are many individuals who gave an incorrectly signed score for this relationship who outperformed individuals who gave a correctly signed score. This is useful for identifying those who are less likely to be correct, but it is not sufficient on its own. Combining it with other known values may be a better indicator.

Three additional filters were added to test whether or not each expert identified the correct signs for W_{wing} vs. t/c_{avg} and V_{app} vs. W_{wing} and provided estimates for all ranges that were correct within an order of magnitude. The average accuracy of model output is shown in Figure 58 compared against the number of filters each individual passed. Nearly all the individuals who pass three or four of the filters perform better than those who did not pass at least two. Those who passed two filters span the entire range. Additional testing may create a more continuous trend rather than the step-function visible here. Comparing the number of filters passed against each of the demographics did not show any clear trends suggesting that this relationship is not based on years of experience or education level.

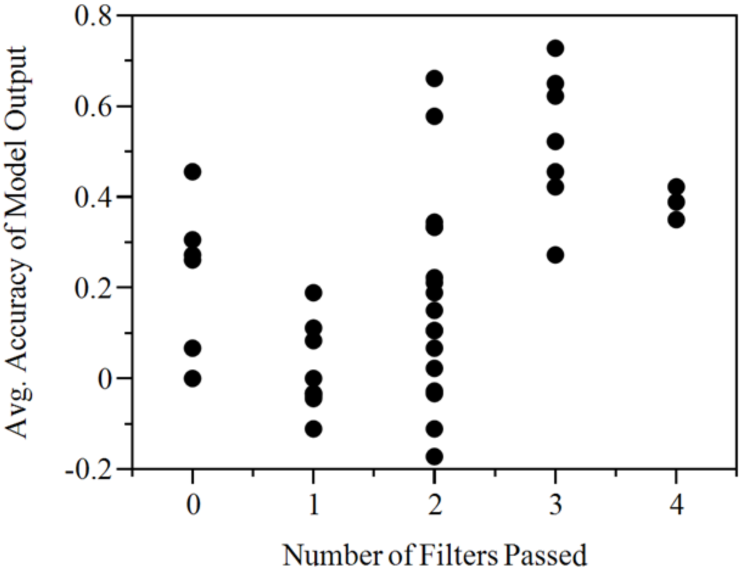


Figure 58: Relationship between Similarity to Truth Model Coefficients and Number of Filters Passed

The filters used here were based on data that happened to be available and have a clear physical meaning with disagreement. For cases where there widespread agreement or where the relationship was close to zero, filtering by sign would be more likely to create a false negative and indicate that individuals should be removed from the pool just for normal variation. In the future, it would be better to identify relationships to test experts as part of the initial design. This could be a particular relationship where good physical data already exists and an expert-based model is not needed. Including such relationships as part of the data collection interface with the relationships that are needed means that it is less likely that participants would identify which is the test. If participants know they are being tested, and should that testing have consequences, they would have an additional motivation to put forth a greater effort.

If a test is included, the entire row becomes part of the test since the scores are all relative to each other. This may mean that more effort is required to provide information than is actually needed for the resulting model. It's also not possible to combine data for one relationship in a row from a trusted source while experts provide all the rest of the relationships in the row. All sources of information must have the same inputs and provide data for all of them (even if that is zero). If a row is split up between different people based on their experience, there is no way to ensure that their scores are truly relative to each other across the entire row.

The downside of these tests is that they almost always require having someone who is more authoritative than the rest of the group identify which relationships can be trusted from an existing model. It is possible for the test itself to be biased even if the facilitator or test-creator is relying on a higher fidelity model, fundamental physics, or historical data. If any such bias is included, this will bias the results of the rest of the model to that particular individual's opinion. One of the reasons that data was collected individually was to avoid the problem of a single most authoritative person or loudest speaker causing others to follow along in his or her voting.

These tests may also be a good way to measure not just who is knowledgeable about an area but also who is skilled at this particular approach. Some of the surprising relationships are explained only by secondary effect that dominate the primary effects. Individuals who are able to look more than a single layer deep are likely to do so for all problems they participate in. In this case, the filtering results should be maintained for each expert and combined with their performance on past modeling efforts (when compared against a higher fidelity model replacing the expert-based model) to produce a method score.

6.5 Accuracy of Output of Expert-Based Models

The most important test of whether or not a model can be trusted is the accuracy of its outputs. The previous set of tests were focused primarily on the correctness of the specific parameters of the models given by experts. Such tests treat all model parameters as equally important, but that may not be the case. If a particular intermediate metric does not have a significant impact on one of the system requirements, the model parameters used to to produce it are less important to the larger system requirement model. In such a case, even if the relationships experts give for the metric are completely wrong, the final model may still be accurate.

To calculate the output, the normalized relationships were used with a coded design of experiments. The same design of experiments was used to calculate outputs from the expert-based models as was used to create the data points from the truth models. It would have been possible to generate a smaller set of design points and calculate them using the linear surrogate models created to compare model parameters in the previous sections, but this would have added a small amount of error due to an additional set of translations. Therefore the data from the full truth modeling environment were used unchanged.

The previous step compared the accuracy of individuals total set of relationships against the relationships present in the truth model, producing a single score. For comparing outputs, it is possible to get a degree of accuracy for each intermediate metric and system requirement. Of the two groups, the ability to match system requirements is more important. In general the accuracy of individual models was significantly lower than the accuracy of group models.

With 42 separate sets of relationships, there are $2^{42} - 1$ or about 4.4 trillion possible combinations of them to form different models. The MATLAB script used to analyze them takes approximately one tenth of a second to calculate the accuracy of each model. With that many combinations and at that rate, it would take just short of 14,000 years and require 64 terabytes of storage to record the results. Instead, to find a number of ideal solutions for analysis, a simple genetic algorithm was used to search the space.

The combined model selected for analysis here is a combination of the relationships from 9 different experts. Of those experts, three used interface Mode 1, four used Mode 2, and two used Mode 3. Five currently work with airliners and one has never worked with airliners. In all other dimensions, they are more or less uniformly distributed. Their combined model was first used to calculate values for the intermediate metrics using a design of experiments. The result of this first calculation is the scatterplot matrix shown in Figure 59.

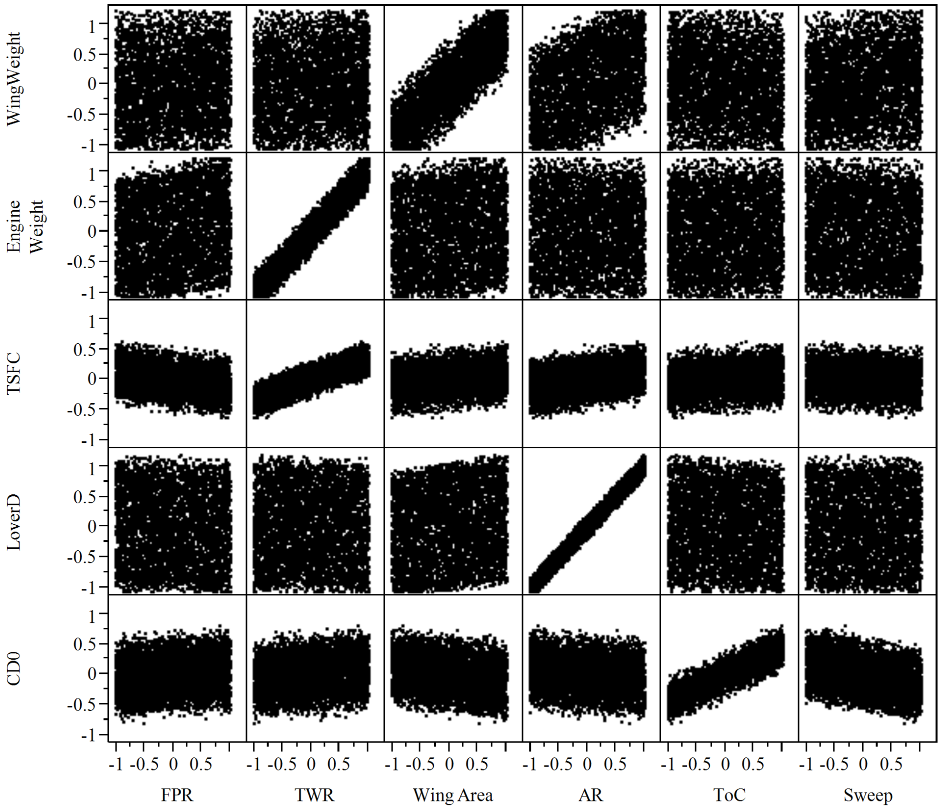


Figure 59: Scatterplot Matrix of Combined Models of Lower-Level Relationships

The subplots where clear trends are visible are those relationships which are most dominant in the combined model. Most of the relationships are filled uniformly which means the value of the combined model parameter is very small or zero. The values of the intermediate metrics are then used as inputs for the higher level relationships and calculated in the same way. This produces the data shown in Figure 60. There are no more relationships that fill up the whole of the box due to further abstraction from the truly independent design variables.

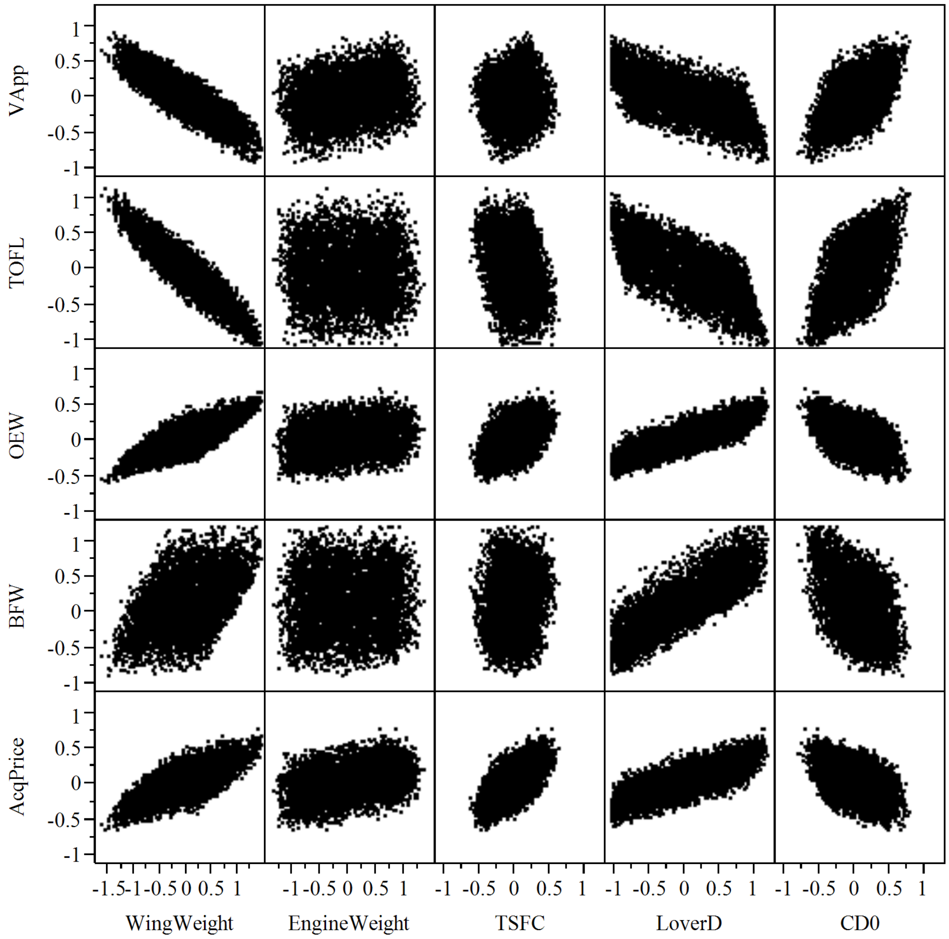
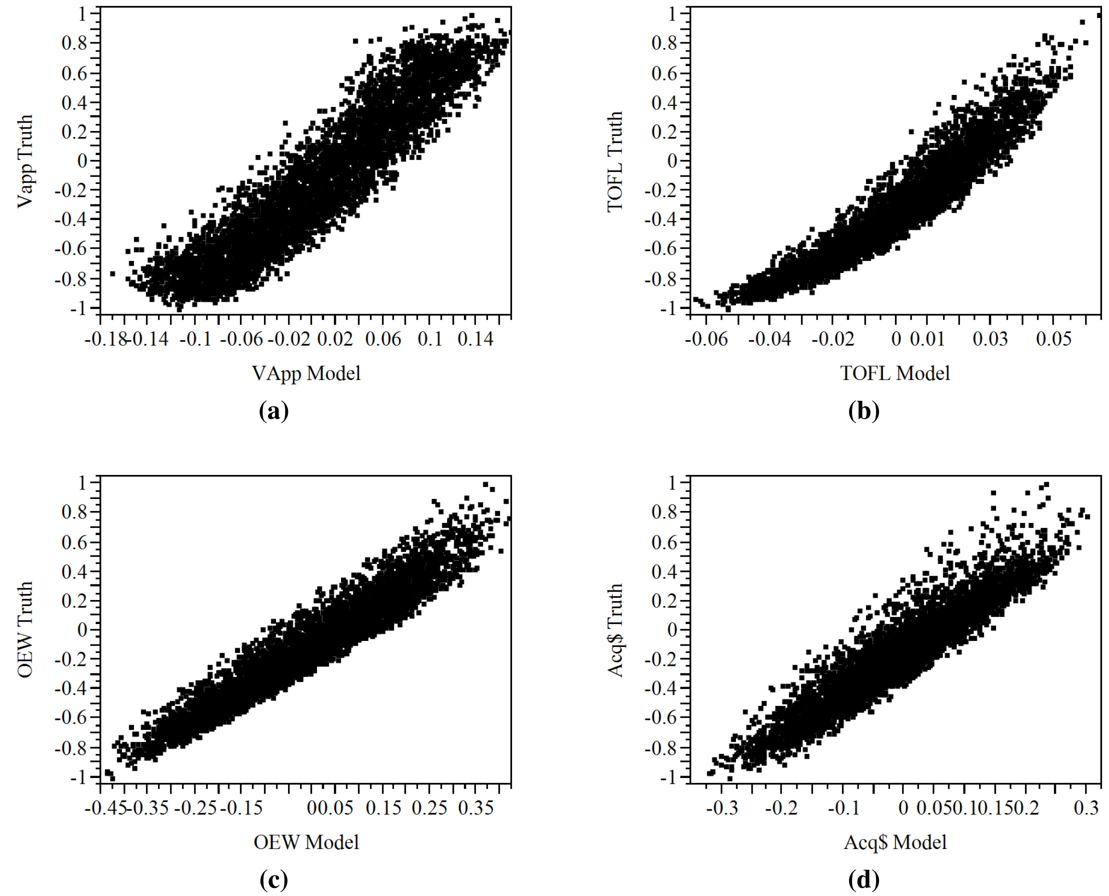


Figure 60: Accuracy of Experts as Pass Filtering Tests

6.5.1 Matching the Correct Ordering and Relative Position of Points

The first two goals for matching the output of the truth model was to be able to get the correct ordering of points and to get the correct proportional spacing between points. If all the relationships are linear, the two become identical goals. If there are a large number of quadratic relationships, there can be a significant difference between the two. Since there are only a few non-linear relationships, the two will be evaluated together.

The same combined model from as above is used for these tests. Actual-by-predicted plots are a common way of testing the quality of a fit of a regression. The values of the truth data are on the vertical axis while the values from the model are along the horizontal axis. The closer each of these is to a perfect line along the diagonal, the better. Figure 61 shows the actual-by-predicted plots for the four system requirements with excellent matchings.



(a) **(b)**

(c) **(d)**

Figure 61: Actual-by-Predicted Plots for System Requirements

One interpretation that can be used with these plots is as a measure of the potential for error in a calculation of an output. Finding a point along one axis, the band of points shows the possible values for that particular point. From these plots, it is clear that the model for operating empty weight has a lower possibility for error associated with it than the approach velocity model. The plot for TOFL has a slight upward curve to it as a result of several of the individual models containing quadratic terms.

The model for block fuel weight did not have the same quality fits as the other system requirements. This may be either due to this being a more difficult set of metrics to capture accurately or it may be that the truth model has more noise in this estimation. The actual-by-predicted for BFW is shown in Figure 62. The relationship still has a slight correlation to it, but it is not nearly as crisp as the other four.

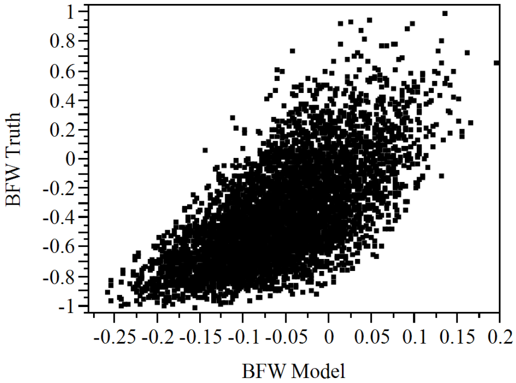


Figure 62: Actual-by-Predicted Plots for Block Fuel Weight

Spearman's r_s is the traditional method for comparing the ranks of ordered data sets. Pearson's ρ_{xy} is the traditional measure of matching proportionality. Both of these are listed for each of the system requirements for the same combined model in Table 17. There is a third column of data for the coefficient of determination, R^2. This is the most common measure for capturing how well a regression fits the original data set. Conveniently, for linear models, R^2 is the square of ρ_{xy}. Considering that these models are based on information created independently of the truth model, these fits are excellent. The model for TOFL showed a slight curve in Figure 61b. This nonlinearity reduces the value of ρ_{xy} for this relationship. Since it is a monotonic curve, r_s does not take the same penalty and has a higher value than ρ_{xy} for TOFL.

This table and the actual-by-predicted plots above are the proof that this method, and

Table 17: Measures of Accuracy of Combined Expert-Based Model for System Requirements

Requirement	r_s	ρ_{xy}	R^2
Approach Velocity	0.93518	0.92733	0.85994
Takeoff Field Length	0.96377	0.94934	0.90125
Operating Empty Weight	0.95877	0.95552	0.91302
Block Fuel Weight	0.65962	0.66628	0.44393
Acquisition Cost	0.93997	0.93736	0.87864

expert-based modeling as a discipline, are capable of producing a useful model. They also show that not all metrics can be modeled accurately, as in the case of block fuel weight. The significantly lower score here is worth investigating.

The intermediate metric with the strongest impact on block fuel weight in the regressed truth model is the lift-to-drag ratio at cruise. This makes sense since it is roughly a measure of the aerodynamic efficiency of the vehicle. When comparing the truth value to the values supplied by experts, the distribution is as shown in Figure 63. The mean is shown with a green line as -0.13 while the true relationship is the blue dashed line at -0.59. While there appears to be a great deal of agrement between experts, none of the experts gave a strong

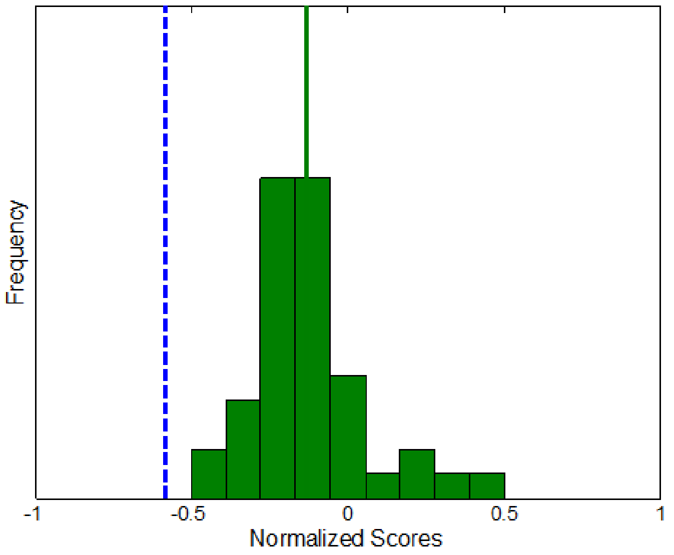


Figure 63: Distribution of Normalized Expert Estimates for BFW vs. L/D

enough negative score to match it. When optimizing to get the best scores, even capturing those individuals who were the most negative was not enough to match the truth model. This is a clear demonstration of a situation where the experts agreed but were still wrong. This particular result suggests that just because experts agree, it is insufficient to guarantee correctness. On the other hand, this is the only relationship out of 55 that has this problem suggesting that if it should happen, it will only be in a minority of cases.

6.5.2 Matching the Correct Dimensional Values for Points

The highest goal was to produce a model that could correctly estimate the physical values of the outputs with correct units and magnitudes rather than just the relative scores that have been presented thus far. Participants were asked to provide the ranges of the physical for each output in order to make that translation. Unfortunately, as discussed in the next section, there was insufficient consensus to be able to determine ranges with any degree of confidence. Without adequate means to identify a physical baseline, it is not possible to take this step. Therefore, for this experiment, the model is incapable of giving the correct values for the points and fails on this account.

6.6 Estimates for Ranges of Outputs

Participants were given design ranges of for each of the design variables as part of the data collection interface. They were asked to estimate the ranges of values for intermediate metrics and system requirements that would be possible for aircraft at the design points within the ranges of the design variables. Units were given, but no baseline information was provided. This proved to be a two-fold test. First, an expert had to have a feeling of what a reasonable starting value for each output would be. Some estimates were very close while others were off by orders of magnitude. Second, the expert had to estimate how much variability would be present in each output. On this count, no participant came close to the ranges identified with the truth model. Considering the comments discussed in Section 6.1, the poor performance here was expected even prior to compiling the responses.

The distributions of all participants' estimates for ranges, along with the ranges of values from the truth model, are shown in Section C.

The estimated ranges with the most consensus and accuracy were those for takeoff field length, shown in Figure 64. The ranges from the truth model have a lower bound at approximately 5000 feet and an upper bound of 8200 feet, indicated by the vertical blue lines. Each horizontal bar indicates the minimum and maximum values for a single participant. One of the reasons that TOFL matches so well is that most experts rounded to the nearest thousand feet, causing a serious of discrete jumps for lower and upper bounds. Also, takeoff field length is a physical measure that is easy to conceptualize, even for individuals who are not aerospace engineers.

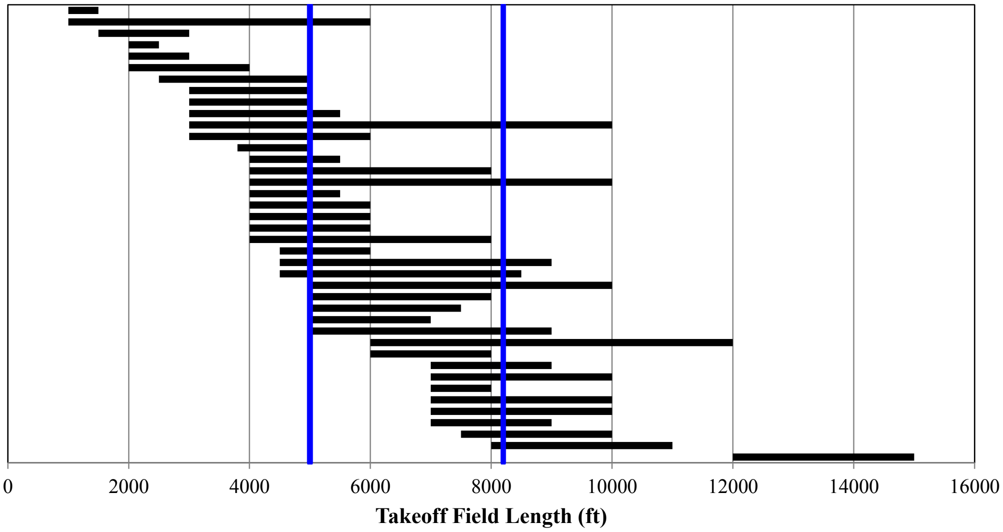


Figure 64: Experts' Estimates of Ranges of Possible Values for Takeoff Field Length

Not all of the ranges lined up so well. In some cases, they showed significant errors in understanding the metric. The ranges for lift-to-drag ratio are shown in Figure 65. The truth model bounds are at 12.9 and 17.8. Many of the ranges are reasonable estimates. It wouldn't be unreasonable for a civil airliner to have a lift-to-drag ratio anywhere between 10 and 20. There are several estimates for values below 1. A quick force diagram shows that an aircraft with a L/D less than 1 requires a thrust-to-weight ratio of more than 1, a trait that belongs mostly to helicopters and military combat aircraft.

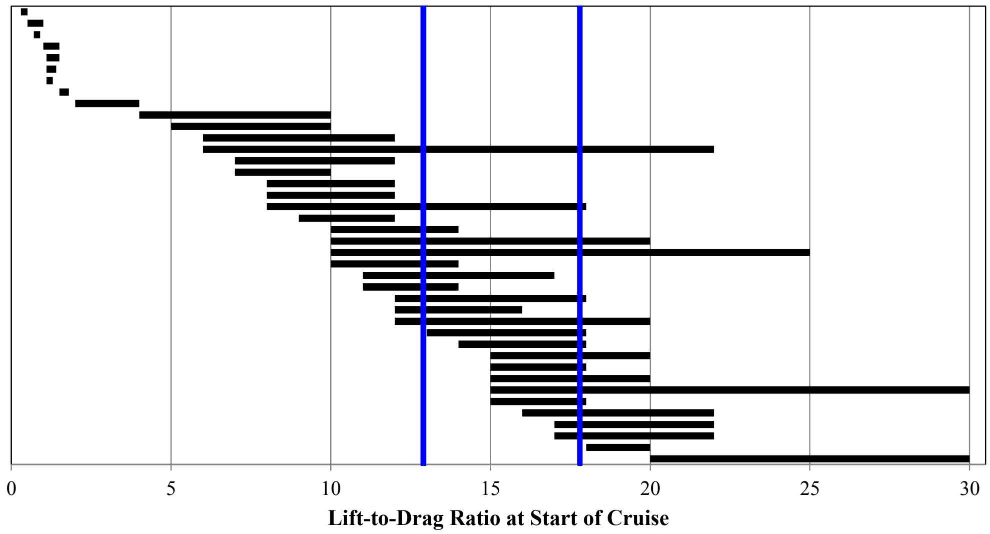


Figure 65: Experts' Estimates of Ranges of Possible Values for Lift-to-Drag Ratio at Start of Cruise

One output in particular was expected to have a high degree of error even before participants began giving data. Cost modeling is a notoriously challenging field and most companies closely guard their own models. It would be unreasonable to expect accurate estimates for cost based solely on memory and intuition.

The truth model includes a cost-estimating tool, ALCCA, but the cost model has not been updated in more than a decade and outputs in 1996 dollars. Fortunately the aircraft defined for this problem uses mostly traditional technologies and materials. Still, material costs fluctuate and inflation changes the value of labor and aircraft. Figure 66 shows the distribution of prices with the truth model estimates as vertical blue lines.

The truth model estimated ranges from approximately $73 to $81.5 million in 1996 US dollars. These have been shifted to 2012 dollars at $106 and $119 million. Boeing identifies the average cost of a new 767-200ER as costing $160.2 million [19]. Though this range is still relatively narrow, reasonable values for acquisition price could be anything from $70 to $170 million depending on what year and what other technologies an individual had in mind. However, with a range that large, cost modeling is little more than a shot in the dark.

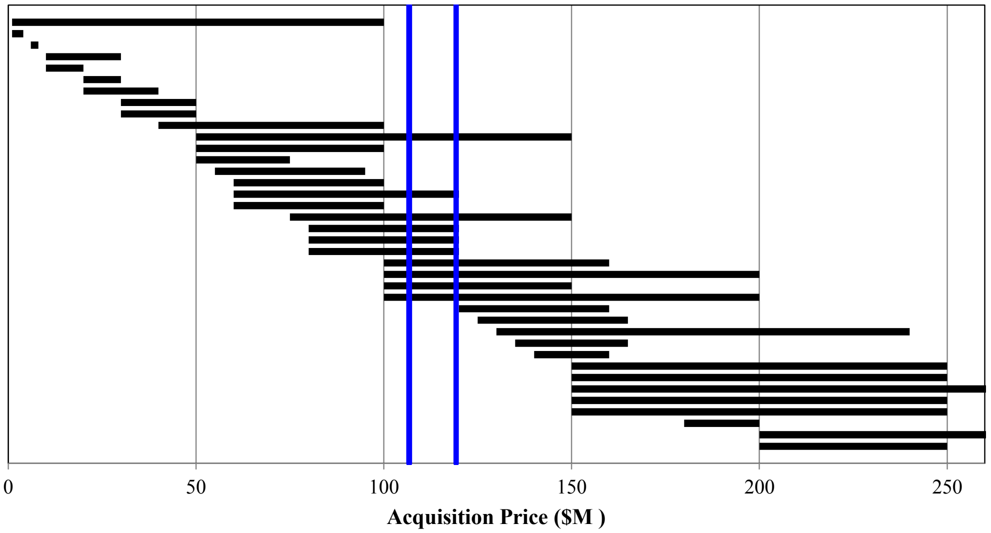


Figure 66: Experts' Estimates of Ranges of Possible Values for Acquisition Price

The ranges for all outputs are consistent in that there is no general consensus among participants for any of them. There is plenty of overlap to identify that a smaller number of experts might have reached agreement if they had communicated. Also, almost all of the ranges are significantly wider than the truth model estimated. It seems reasonable that most experts selected ranges that would be sure to include the design points rather than trying to match the exact ranges.

Without a greater degree of consistency, it is not possible to create any useful combinations. Without ranges to apply, there is no way to test how close the expert-based models could get to estimating actual values and no way to test the third level of "accreditation".

6.7 Number of Participants

One of the concerns with developing a new expert-based method is identifying the correct number of individuals to include. Hypothesis 3 predicted that 5-6 experts would be sufficient to come close to the maximum accuracy of a model and that further experts would not necessarily help. It has been shown already that not all of the participants in this experiment would truly qualify as experts. Including a larger number of individuals who provide poor data will not improve the model over a smaller number of individuals who provide good data.

In order to test Hypothesis 3, only individuals who could qualify as experts should be included. To identify those experts, each participant's individual model was compared against the truth model. The correlations to all five system requirements were averaged and the participants were ranked. The top participants from that list were included as the set of available experts. The test was performed with 10, 15, and 20 experts included in the set to compare the differences. For each of those three sets, models were created using each individual, all members of the group and then random combinations of different numbers of experts. The correlations for all five system requirements were averaged. The accuracy of each combination of experts, from single experts to all members of the group, was plotted against the number of experts included in that model. The results for each of these is shown in Figures 67 through 69. The median score for each number of experts is shown as a green line.

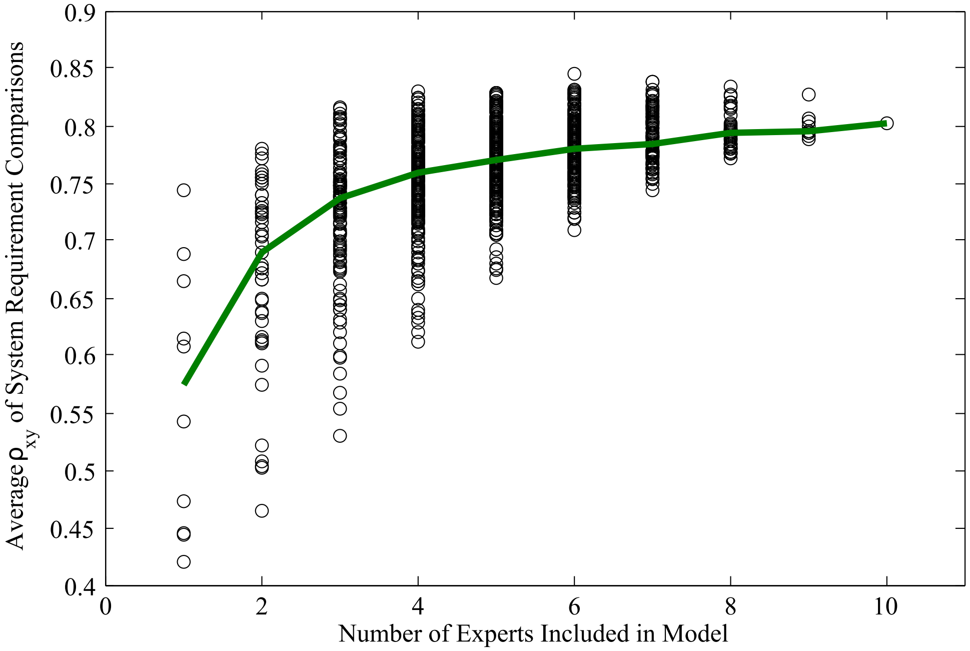


Figure 67: Average ρ_{xy} as More Experts are Included in the Combined Model with 10 Total Experts

With 10 total individuals, the median value of the accuracy score continues to increase up until all experts are included in the model. However, the maximum possible accuracy occurs earlier with the inclusion of six experts. Perhaps most interesting is that, once four or more experts are included, the median score is higher than the maximum individual score. This suggests that, even if one expert is nearly perfect, combining him or her with several slightly less perfect experts will improve their overall accuracy. It was only possible to identify the single best individual with the help of the truth model. So in order to ensure that a model is the best in the absence of perfect knowledge about who will produce the best model, multiple individuals must be included. The median value for six experts is 0.779.

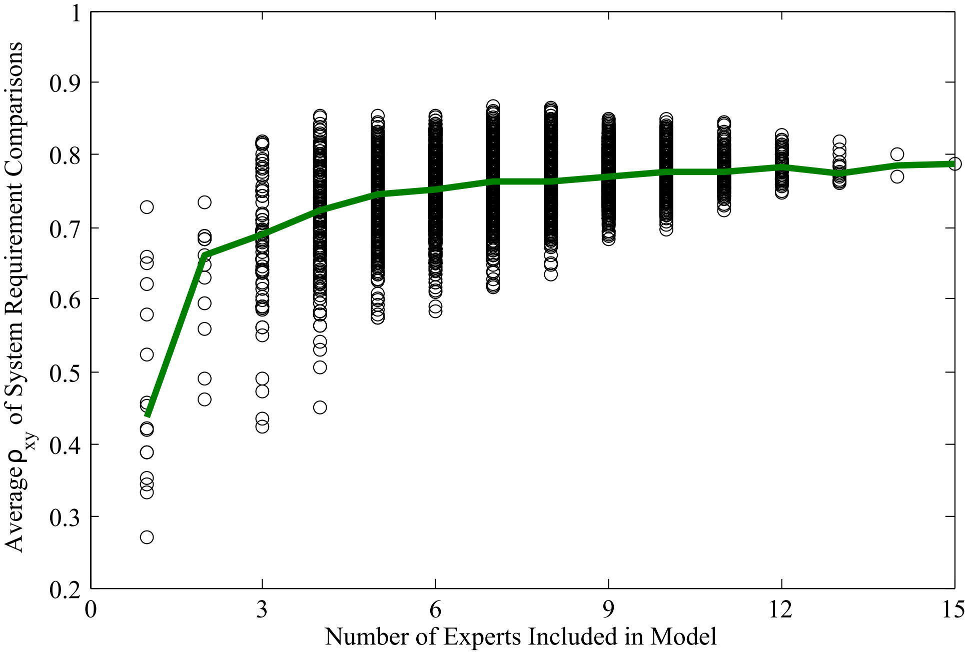


Figure 68: Average ρ_{xy} as More Experts are Included in the Combined Model with 15 Total Experts

Including the next top five experts for a total of 15 increases the size of the cloud of scores. The maximum possible score is higher, but many more model combinations are below the performance. This is expected since more individuals whose own models are included. The initial increase in median score is much steeper. The median score never reaches the same level of accuracy as in the first case, ending at 0.788 instead of 0.803 for the 10-expert population. Still, for a 50% increase in population all of whom are lower performing, the change is relatively small. The median value for six experts is 0.753, only slightly lower than the 10-expert population.

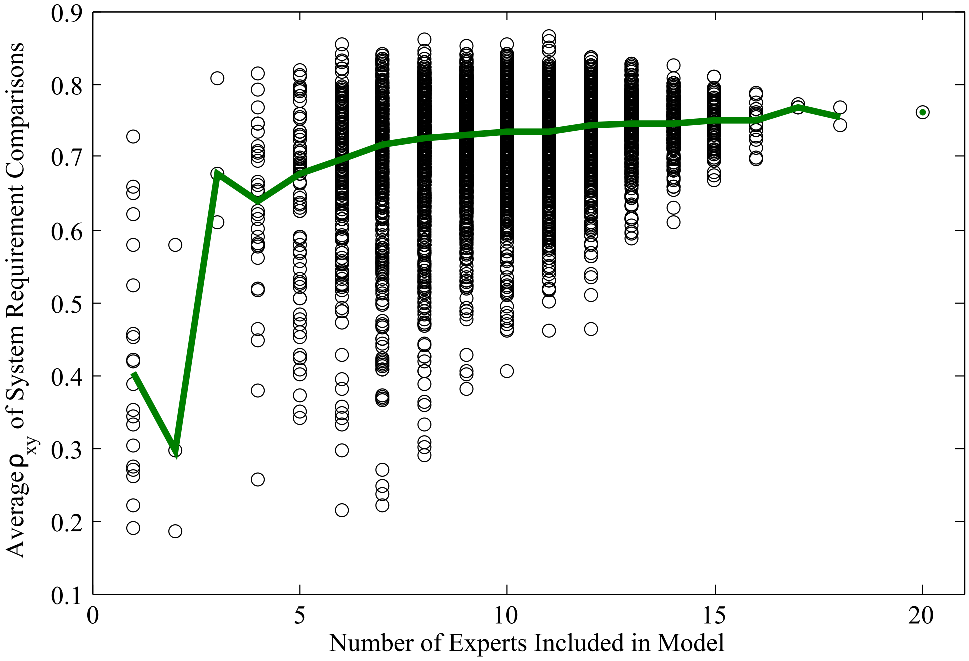


Figure 69: Average ρ_{xy} as More Experts are Included in the Combined Model with 20 Total Experts

When including 20 total individuals, the cloud of scores is much broader. At this point enough of the participants would not be considered experts based on their performance. Any model they are included in is hurt by their joining. The highest possible accuracy score is higher than the other two groups, but there is a higher chance of a poor-performing model. This evidence does not suggest that 20 individuals are too many, but rather that 20 individuals are not necessary due to the likelihood of including individuals who are not experts. It would be better to include a smaller number of more trusted individuals. That said, if the initial pool is larger and individuals are eliminated by other means before inclusion in the model, 20 individuals may be a good original number. The accuracy for all individuals is 0.760 and the median for six experts is 0.696. This represents a much larger decrease changing from a full populations size of 15 to 20 than from 10 to 15.

Based on this analysis, Hypothesis 3 is weakly supported, but this depends heavily on the quality of the individuals who are available. If participants are all even moderately qualified experts, 5-6 individuals seems to be sufficient to approach the diminishing returns of adding an additional expert. If less qualified experts are available, including a larger number will not necessarily improve quality but may allow for some to be excluded based on the filtering demonstrated in Section 6.3.2. If only very marginally qualified individuals are available or they are indistinguishable from highly qualified experts, no number of them will be sufficient to produce a quality model.

6.8 Potential for Impact from Other Factors

Not every part of an experiment can be controlled, especially with human subjects. Any number of things can impact the performance of experts from the time of day the participant gave data to the noise or other distractions present in the room. Many of these are difficult to identify, let alone measure. Two variables in particular were of interest. Since not all experts participated at the same time, there is a chance that the order of the participants had some impact on the results.

Participants were timed automatically while they were giving data using the data collection tool. While the average time spent was 40 minutes including instructions. There was sufficient variation in the times that any relationship it had with accuracy was worth investigating.

6.8.1 Order of Participants

One concern is that the author may have unintentionally impacted the performance of individuals. After giving the instructions several times, they may have changed slightly to address problems other participants may have had. While the author attempted to distribute different demographics throughout the process, it was more important to schedule time with volunteers as soon as convenient to ensure sufficient participants within the time available.

To test this, compare the performance of different experts against the order they were recorded in as shown in Figure 70. There is a very slight positive trend present between the two.

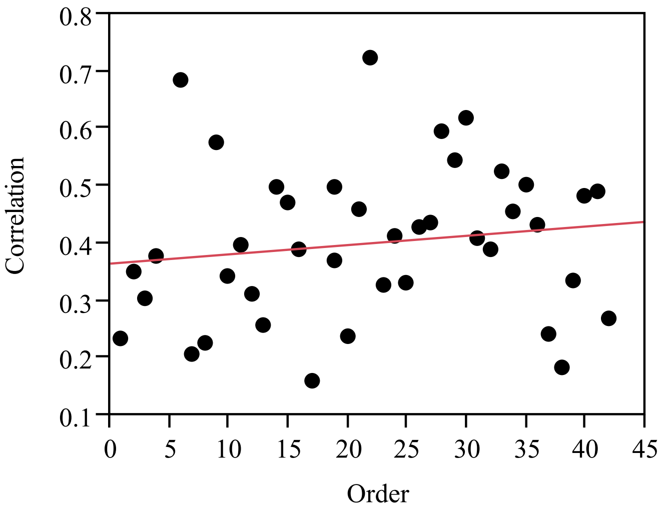


Figure 70: Trend Between Correlation to Truth Model Coefficients and Order of Participation

However, there is another effect that may be at play here. The individuals who had the most experience and best reputation for knowledge and skill in aircraft design were not all available for the early portion of the data collection process and generally gave data later in the process. Unfortunately, reputation could not be measured for reasons of privacy so it cannot be included in the numerical analysis. The individuals later in the data collection period also tended to have more experience with QFD and similar approaches, though not universally so. Figure 71 shows the trend of aircraft experience against the order in which volunteers participated. Since some participants were giving data concurrently, the order here is based upon when participants were given the initial instruction.

Earlier analysis has already shown that experience with aircraft is not sufficient to predict accuracy of an individual. However, the weak relationships here suggest that if there was any effect of giving data early in the process versus late in the process, it is minor. In actual usage of this method, participants would either be experienced enough with the process to not require instruction, the moderator would have already reached a point of

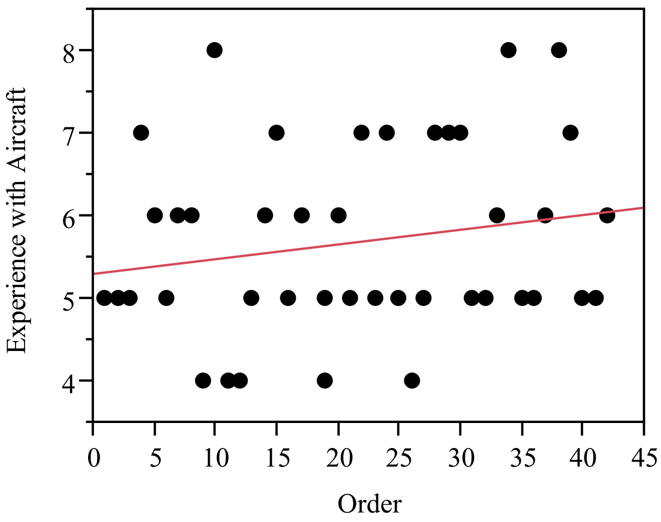


Figure 71: Trend Between Experience and Order of Participation

steady-state effect on participants, or participants would likely receive instruction simultaneously in person or by webcast. Either solution would further eliminate any possibility of bias.

6.8.2 Time Taken by Participants

As part of the data collection, the tool started a timer when it was started and stopped the timer when the participant saved the data. At first this was to ensure that the promised time commitment was respected, but participants were informed of it. This time included a 10 minute period of instruction for most participants. There are several factors which add noise to this measurement. Some participants (especially those with much lower times) either received instruction in pairs or closed the tool after instruction and reopened it at a later time to give data. For those individuals, the 10 minute instruction period is not included in their times. Others may have had momentary distractions while giving data causing an increase in their times. While the time for instructions was fairly consistent, there was a margin of approximately ±2 minutes where some individuals had more experience with the process and required less time and others had additional questions.

The distribution of times participants took to give data is shown in Figure 72 with the

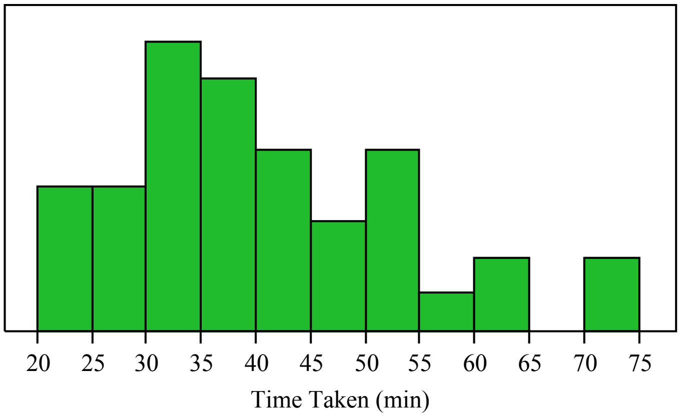


Figure 72: Distribution of Recorded Times Spent Giving Data

exception of one outlier who left the tool open overnight. Excluding the single outlier, the participants averaged 40.8 minutes with a standard deviation of 13.0 minutes.

It was considered that there might be some relationship between the time an individual spent and the correctness of their scores. Those who rushed through the process might not have thought out the trades and relationships thoroughly and given poor data. Alternatively, an expert with greater comfort and knowledge of the material may be able to give more accurate information based on intuition faster. The comparison of the time it took for participants to give data and their correlation to the regressed coefficients of the truth model is shown in Figure 73.

While there is a slight trend, it's very weak and driven primarily by the individuals who took the longest times. Considering the sources of noise for this information and the broad scatter, it is not possible to come to any statistically defensible conclusions. Note also that this trend does not capture whether or not a single individual would be more or less accurate if he or she spent more time and effort giving data. Some testing protocols do not let participants move forward to the next step until they have spent a certain amount of time on an earlier step. Such an approach was not used here since participants were generally giving their best faith efforts.

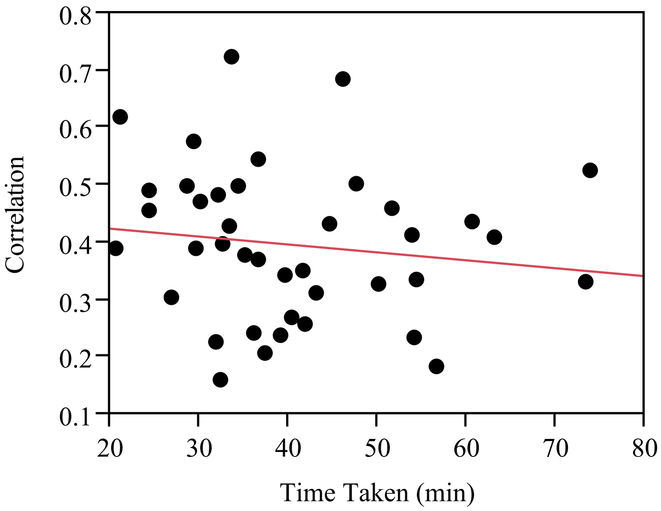


Figure 73: Trend Between Correlation to Truth Model Coefficients and Time Spent Giving Data

6.9 Revised ALTER Process

The results of the analysis presented in this chapter show that this method is valid but also show areas for improvement. In addition, the very act of performing the method revealed other changes that were not visible until tested. This section provides a brief summary of the ALTER method incorporating the changes identified in this chapter as well as Chapter 5.

Define the Problem Start with the motivating problem and system to be modeled as an input. If the concept is not already well defined, define it using an Interactive Reconfigurable Matrix of Alternatives and downselect to a single concept. Identify a series of design variables as inputs, desired system requirements as outputs, and intermediate metrics to connect them. Ensure that the design variables and intermediate metrics are as independent as possible. Define units for all three levels or measures and identify ranges of interest for the design variables.

Gather Information Use the measures, units and ranges that were defined to populate an interactive data collection tool. This tool should represent the possible scores for each relationships using a slide bar that includes all integers from -9 to +9 as well as providing experts a place to estimate ranges of outputs. If a baseline or starting point is available, clearly identify its values to improve the accuracy of the estimates for the ranges. Identify 3-6 relationships which have known signs or values to serve as tests of expertise. These should include both obvious and confusing relationships.

Recruit a group of 5–15 willing and motivated experts with diverse backgrounds but all with experience across design fields for the problem at hand. A larger number will provide more flexibility if some are eliminated later on. These individuals should have at least four years of recent experience in the relevant field and should be well-respected by their peers. A nomination form or an established list of those who perform well on these exercises is an ideal way to identify experts easily. Host a training session to teach the experts how to give useful data using this method. The session should include both a simple example and a more complex example with several "trick" questions to demonstrate the need to think about the relationships carefully. Preferably this training would be performed prior to needing to develop a model so that this process does not need to wait on the training to occur.

Provide each of the participants with the data collection interface and the reference materials and the necessary time to complete it. This step is best performed concurrently to minimize the clock time needed to complete this step.

Create the Model Collect and collate the data from each individual and combine it with any information about the individuals (peer ratings, years of experience, etc). Scale and normalize their scores if it is not already performed as part of the data collection tool. Use the test questions to score each individual as to how well they are expected to do. Discard any individuals who failed all of the tests and as many others who do poorly as can be afforded. Average the normalized scores from all remaining experts.

Test the Model Compare how well the remaining experts agree. For any relationships where there is a disagreement between the experts, identify what additional information is available and relevant and present it to all of the experts again. Allow them to provide a new score for that relationship or adjust the existing score. A group discussion may be helpful, but it is important for the facilitator to ensure that no one person dominates the discussion or unfairly influences the opinions of others. In addition to individual relationships, test the overall agreement of experts' scores. If any one individual disagrees with most others and does not cluster with any other users, consider removing his or her data.

If a larger number of experts (more than 10) remain in the population, use selective sampling from this set and compare the similarity of the model outputs to identify the variation among the group.

Also analyze the agreement between the experts for ranges for each intermediate metric or system requirement. If an individual has an estimate that is far from the others, provide an opportunity to refine his or her estimate or explain it.

Use the Model Package the finalized and tested model into a portable format either still normalized or scaled to fit the consensus of the ranges for each output. Use this model for its desired purpose. As other, higher-fidelity models are developed, compare them to the expert-based model to identify areas for improvement and test the accuracy of the model and any decisions that were made based on it.

CHAPTER VII

CONCLUSIONS

This dissertation began with a high-level goal of being able to rapidly create a quantitative that would be useful for early engineering problems using only expert knowledge and opinion. This research goal motivated a series of literature searches to explore existing work and recent developments in expert-based methods in general and expert-based modeling specifically. The available literature led to the decision to use a QFD-style hierarchical framework for building models while incorporating some of the advances that were part of the ROSETTA and SP2/SOAR methodologies.

In order to develop a structured process for creating expert-based models, a survey was performed of general modeling processes. A generic modeling framework was then used to guide the development of the methodology, starting with *Defining the Problem* and then *Gathering Information* necessary to *Create the Model* with both linear and quadratic terms. During the development, three preliminary experiments were performed to provide analytical information about potential problems and the applicability of solutions found in literature. Approaches were identified to *Test the Model* in the absence of higher fidelity models or historical data for validation so that it would be trusted when one *Uses the Model*. The development of the methodology generated several research questions for investigation and four hypotheses to test.

An experiment was performed to test the method using actual people to ensure that the most essential part of expert-based modeling was included: experts. Forty two participants volunteered to give data using one of three different interfaces and scoring scales. The data was collected and analyzed in detail, both as individual records per person and combined together. These results showed that, while not all participants were at the same level of expertise, it is possible to create a reasonably accurate model based solely on information from experts intuition and opinion.

7.1 Summary of Observations

The process of doing research leads to more than just the answers to the planned tests. Along the way, new and sometimes unexpected observations are made and lessons learned. The most significant observation is the demonstration that it is possible to create accurate expert-based models, but that such models are highly dependent on having the right experts available and willing to participate.

To that end, not all individuals who have the credentials to be ideal experts are the best at providing the needed data. Two measures of an expert that are often left out of the qualification process are their personality and their ability to convey information. Some individuals who are quite knowledgeable are less able to translate that knowledge into a desired format. Conversely, some individuals who did not have the credentials expected of an expert were capable of giving very accurate information. And at times, having knowledge of the process used for gathering information can have a greater impact on success than the technical knowledge being gathered.

During the process of collecting information, the interface that was was expected to be the least interesting turned out to be the most successful. The simple slide bar interface with an integer scale produced the highest average accuracy as well as the lowest average time required to give data. This is likely a result of providing enough flexibility to be accurate without either overloading or oversimplifying the process with a graphical interface or a completely continuous scale.

Even though an interface is simple and straightforward, that does not mean that giving data will be. The amount of difficulty that participants had estimating ranges of outputs highlights the difference between understanding and knowledge. Participants did relatively well at providing relationships based on understanding and intuition, but pitifully bad at providing values based primarily on knowledge. However, if experts had been given a baseline set of values, or in some cases just an order of magnitude, it is likely their ranges would have been much more accurate.

Many of the findings of this research are easily transferred to QFD or SP2/SOAR which extends the usefulness of this research beyond this method.

7.1.1 Hypotheses Revisited

A series of hypotheses presented in Chapter 4 and tested specifically. Each hypothesis is listed here along with the conclusion drawn from analysis of the experimental results and supporting evidence. Each of these hypotheses is associated with selecting or defining a particular aspect of different steps in the ALTER method by comparing it to other possibilities. They were formed based on existing literature and preliminary testing. In the case of a failed hypothesis, this means that the existing intuition was incorrect and that a different approach may be optimal for this method.

> **Hypothesis 1a:** Because of the reduced set of possible values, there will be a greater degree of agreement among experts who use the QFD scale than among experts who use either of the other scales.

The analysis in Section 6.2.1.1 compared the effect of different interface modes on agreement between participants using the average Pearson's product-moment correlation coefficient (ρ_{xy}) for all unique pairs within that mode. The differences between the interfaces modes was below the threshold for statistical significance but the QFD and the Mode 1 interface associated with it had slightly lower agreement between participants than other interfaces.

The analysis in Section 6.2.2 made the same comparisons but used Krippendorff's alpha (α_k) for each subset of the population to measure agreement. No tests of statistical significance were performed, but the QFD scale and the Mode 1 interface had noticeably lower agreement between participants than the other interfaces.

Neither of these analyses support Hypothesis 1a but both provide weak evidence for rejecting Hypothesis 1a. Future efforts should therefore either continue to compare different interfaces for different problems and groups of individuals or may focus on the full integer scale in light of the slightly better agreement among experts using it and Hypothesis 1b.

> **Hypothesis 1b:** The greater degree of discretization and information presented to the participants provided by the graphical display will allow experts to more accurately capture their intended relationships as they believe.

Section 6.4.2 compared the similarity of the model parameters from participants using different interfaces against the model parameters of a truth model. The confidence interval of the means for all three interfaces showed that Mode 1 (QFD scale) was not significantly different than the mean accuracy of all participants, that Mode 2 (integer scale) was statistically significantly better than the mean accuracy of all participants at the $\alpha = 0.10$ level, and Mode 3 (graphical interface and quadratics) was statistically significantly worse than the mean accuracy of all participants at the $\alpha = 0.10$ level.

This analysis provides strong evidence to reject Hypothesis 1b, but also provides strong evidence to support an alternate hypothesis that Mode 2 provides the greatest accuracy. Future efforts should therefore focus on interfaces using the full integer scale to collect data.

> **Hypothesis 2:** Individuals with more than two years of experience with the relevant vehicle type who have demonstrated a general knowledge of aircraft design concepts by passing the doctoral qualifiers will perform noticeably better than those who have not. Individuals beyond this mark will perform similarly.

Section 6.4.1 compared the similarity of the model parameters from participants with different levels of education and different amount of experience against the model parameters of a truth model. The number of participants who were in the group of students who had passed PhD qualifiers but had not completed the PhD was significantly larger than the number of participants in other levels of education. The lack of diversity in this demographic field led to very large confidence intervals for the means of the other other levels of education such that the results were not statistically significant. The comparison of years of experience with aircraft design showed that the group of participants with two to four years of experience was statistically significantly less accurate than the total population at the $\alpha = 0.05$ level. The same analysis showed that the group of participants with four to eight years of experience were statistically significantly more accurate than the total population at the $\alpha = 0.10$ level.

This analysis provides strong evidence to reject Hypothesis 2, but also provides evidence to support an alternate hypothesis that individuals with more than four years of experience will perform noticeably better than those with less. Future efforts should therefore focus on experts with at least four years of experience in the field or type of system they are providing data about.

> **Hypothesis 3:** Increasing the number of experts whose data is included in a model past five to six will not have a significant impact on the accuracy of the combined model.

Section 6.7 performed a series of tests to compare how quickly accuracy of a model converged based on number of experts included relative to the total populations sized available and the quality of the experts in that population. Combined models were created from different numbers of randomly selected experts from each population and scored for accuracy. It was shown that for any size population, the median accuracy score for a given number of experts reached diminishing returns by the point of six experts, though the median continues to increase until the full population is included. The maximum accuracy peaks closer to when half of the experts are included.

This analysis weakly supports Hypothesis 3. Future efforts should therefore continue to use at least six individuals until additional empirical evidence suggests that a smaller number is suitable, but need not include a very large number.

7.2 Suitability of Methodology for Use

The ALTER method presented here is not universally applicable, nor is any method. There are some classes of problems it will likely never be useful for due to some of the early assumptions and decisions. There are also certain types of environments where this method will thrive and others where it will likely fail. Even if both succeed, there are limitations to how the results should be used.

7.2.1 Limitations on Problem Type

There two fundamental decisions that limit problem type are the use of mostly linear relationships, and the failure to include any cross-terms or interactions in the model form. The choice of using linear relationships was one of convenience and because the assumption worked for the problems it was tested on. The choice not to include interactions was a result of being unable to find a way to collect that information from experts in an intuitive and useful fashion.

Problems that have strong interdependencies, especially scenarios where time is an independent value, will not work with this. This approach may be valid for problems which estimate results at the end of a given time, however. Problems describing complex systems, frequently defined by their inability to be modeled by typical methods, will not be valid unless the problem is scoped up to the level that reduces a complex system to a simpler system with context.

This method will also not work on problems that are very sensitive to minor changes. Some structural and aerodynamic problems have sweet spots or trouble spots that drive the behavior in a different direction from the previous trend.

Problems where this approach will work well include those that involve a hierarchical breakdown or a series of processes. The more that a system can be broken down into isolated parts, the easier it will be to model.

7.2.2 Organizational Requirements

Many of the limitations of when ALTER could be used depend on the culture of the organization it is used within and the experts who are available to participate. The most important resource at an organization is experts who have are able to extrapolate beyond problems they have seen before and to simplify down trends to capture the broad strokes. Experts need to see the benefit of participating and in giving it their honest effort. Including follow-up phases to collect their further refinements and to reveal the quality of the resulting model establish a pattern of usefulness and quality.

From a management perspective, an organization must be flexible enough to allow experts to participate in this activity. If this activity is able to easily cut across groups to include experts from slightly different backgrounds, that would help to make sure that no specialty relationships are neglected. This method, like any other, tends to work best with individuals who are familiar with it. Several individuals performed better than peers with similar aircraft backgrounds because they had more experience with QFD and QFD-like tools.

7.3 Summary of Contributions

This research presents a number of contributions to the published literature and to the field. First and foremost, it demonstrates that the ALTER method for expert-based modeling is a valid approach to create useful models. It proves this in spite of imperfect agreement between experts and in spite of imperfect individual accuracy. The ability to create useful models with imperfect information reduces the level of detail and precision required from experts when creating a model. This accuracy can come both from experts who agree with each other about the correct relationships and from experts who disagree just enough to balance each others biases and errors out to average to the correct values.

This research also demonstrated conclusively that a small number of better experts will more consistently provide a more accurate model than a larger number of experts who do not perform as well. However, the larger group has a higher maximum accuracy, despite how unlikely it is that that particular set of experts would be selected from the population. This suggests that, if any form of analysis is available to test the abilities of the experts, the overall model may be better if a larger group is used and then reduced down to a smaller set of optimal participants. However, if it is difficult to perform such analysis for the problem at hand, it is best to select those individuals who have a great deal of experience, but also are most comfortable extrapolating and exploring mental what-if scenarios.

Though IRMA was initially released some time ago and has been incorporated into numerous other methods, this research is the first time it has been tested with users, despite its static representation. The use here demonstrates the difference between its use as a reference tool after being populated and defined rather than as a concept exploration tool.

Most expert-based relationship-describing methods are never tested against the true relationships. The only test that is performed is the change in the quality of the resulting product. Such analysis shows that the method is useful, but not that it is correct. The approach here used to test the accuracy of the relationships and the outputs of the model provides a template for similar comparisons for this and other expert-based methods in the future using the traditional models created later in the development process. This fundamental difference supports the continual improvement of expert-based methods to better approximate the true behavior.

Part of this ability to test the model was based on the ability to take a monolithic black-box model and decompose it into multiple logical levels with roughly independent variables as intermediate metrics. This allowed simplified regression techniques to be employed without the need for accounting for aliasing.

Lastly, the analysis performed produced several results that are applicable for most QFD-style relationship-based expert methods. The use of the full set of integers for a given range were demonstrated to produce significantly more accurate relationships than the typical reduced integer range typical of most QFD exercises. Most methods could use the full set of integers with only a change to the interface and no change to the back-end math used for calculations. This research also demonstrated the benefit of using a series of questions to test and then select a subset of experts to include in the final model from a larger set of participants. This down-selection tends to produce more accurate relationships, but cannot guarantee it. As the number and diversity of the testing questions increases, the level of certainty of who to include or exclude also increases. In cases where experts disagree, this research demonstrated the improvement associated with giving additional information for those areas of disagreement. While existing methods return to provide additional focus on areas of disagreement, most rely on the discussion of the group which may lead to domination by a single individual or small group of individuals. The method here avoids that problem by continuing to keep responses hidden.

7.4 Recommendations for Future Research

One of the great tragedies of completing a dissertation is that, after spending so long with the problem and the material, one comes up with a surplus of "Great Ideas". These ideas may be alternative approaches or curiosities that deserve testing or realizations that a much simpler approach is possible. While a field of research can continue to grow indefinitely, a dissertation cannot. Several ideas which may be helpful for future research are presented here divided between additional experiments that could be performed with the same method and setup and ideas that require furthering the method to develop something beyond it.

7.4.1 Additional Testing

The first improvement to the testing would be to repeat the effort with a group of people who would be considered experts with more certainty. Creating a series of expert-based models for an aircraft with individuals from the early design groups at Boeing or Airbus would be a better test of the limits of accuracy of this method.

It would also be ideal to include some of the other metrics of expertise, such as peer evaluations and rankings. These methods would provide additional demographics about experts that may be helpful in identifying deeper trends about what type of person is most useful to provide data.

A small number of participants in the experiment asked to receive instructions together. Since many expert methods involve experts correcting each other and balancing own opinions with others, it might be interesting to have pairs or small groups of experts each contribute their data together. This might be especially appropriate if clumps of experts are geographically distributed.

7.4.2 Method Improvement and Extension

It is rarely the case that absolutely no information is available about a problem. For this research, it was assumed that such information did not exist because finding the balance of what information did or did not exist and should or should not be presented to the participants was difficult. Since there was really only one chance to perform an experiment, that was not a degree of freedom that was varied. There are two specific types of information that could be integrated. The first is information about certain relationships. Just as some relationships were known to only be positive or negative, this information could be included in the data collection tool to only allow values that meet those limitations. The other is actual values of a baseline design. Such information would use an existing known design point as a starting point to aid experts in estimating ranges of the intermediate metrics and system requirements with greater accuracy than they did in the experiments here. This could be further expanded to develop adaptive modeling techniques that use a sparsely set of data points to reduce the total degrees of freedom in the model and allow experts to assess which known changes in design variables cause which known changes in intermediate metrics and system requirements.

The method as presented is limited to mostly linear models with some small number of quadratic models. Those quadratic models relied on an expert being able to judge the shape of a model rather than its relative parameters. In some cases it is easier for an expert to estimate parameters for a particular model. In addition, there are other known general relationships that apply to certain classes of problem that could be easily incorporated. Experts could use a similar method such as this to estimate the parameters of those known relationships.

Including arbitrary model forms allows expert-based modeling to match closer with systems that do not behave linearly. This also means that they are closer to traditional modeling and can start to be integrated into them. As the difference between the expert-based modeling presented here and a traditional modeling effort shrinks, it is easier to switch from one to the other and for one to improve the other. There are times when expert intuition is more correct than a physics-based model. Developing methods to apply intuition as a correction factor would allow fine-tuning of existing models to new problems.

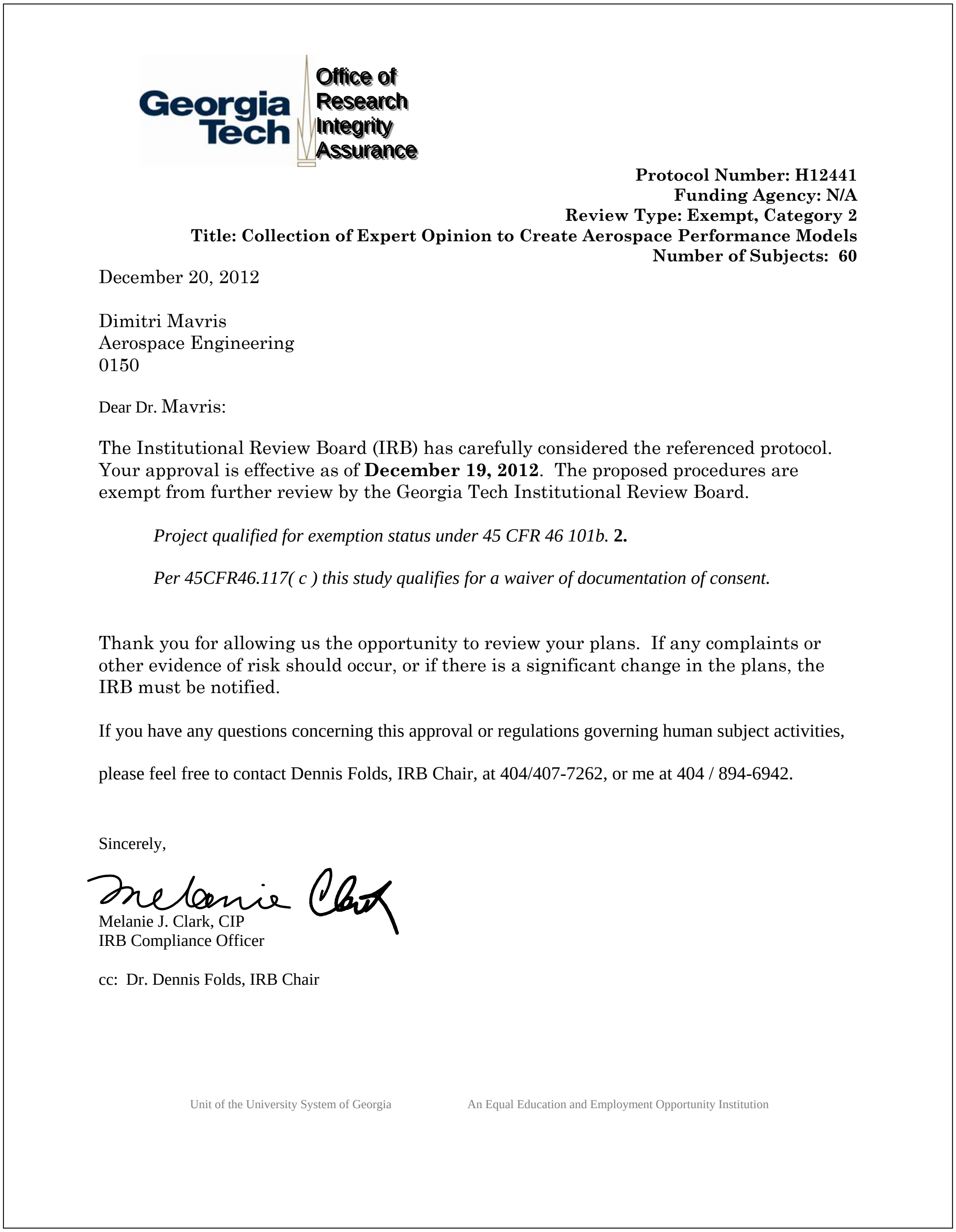
One of the initial problems that inspired this work was the difficulty associated with developing activity models for military mission planning. For many of the individual activities, both the input and output is a set of information with some transformation, addition, or dissemination happening in between. It is not possible to create physics-based models to capture the effect that changing the quality of the input or how well the activity is performed has on the quality of the output. Even models based on human testing are difficult to create and test for new solutions. However, expert-based estimates of these relationships may be sufficient to distinguish between several alternatives with some estimate of the goodness of the information output.

7.5 Closing Remarks

This research initially started with the idea that experts could just draw arbitrary curves and have them translate into relationships and models. After research, prototyping and having to create something that actually works, it is clear that the early goal is not as easy as originally envisioned. It is also clear that there are very many paths and extensions from here and from parallel paths that were not included here. This research has shown what is possible using expert opinion and identified a few more best practices than existed before. It is hoped that this research will aid others in continuing to push the Pareto frontier of speed versus accuracy.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL



Georgia Tech
Office of
Research
Integrity
Assurance

Protocol Number: H12441
Funding Agency: N/A
Review Type: Exempt, Category 2
Title: Collection of Expert Opinion to Create Aerospace Performance Models
Number of Subjects: 60

December 20, 2012

Dimitri Mavris
Aerospace Engineering
0150

Dear Dr. Mavris:

The Institutional Review Board (IRB) has carefully considered the referenced protocol. Your approval is effective as of **December 19, 2012**. The proposed procedures are exempt from further review by the Georgia Tech Institutional Review Board.

Project qualified for exemption status under 45 CFR 46 101b. **2.**

Per 45CFR46.117(c) this study qualifies for a waiver of documentation of consent.

Thank you for allowing us the opportunity to review your plans. If any complaints or other evidence of risk should occur, or if there is a significant change in the plans, the IRB must be notified.

If you have any questions concerning this approval or regulations governing human subject activities,

please feel free to contact Dennis Folds, IRB Chair, at 404/407-7262, or me at 404 / 894-6942.

Sincerely,

Melanie Clark

Melanie J. Clark, CIP
IRB Compliance Officer

cc: Dr. Dennis Folds, IRB Chair

APPENDIX B

HISTOGRAM MATRICES OF RELATIONSHIP SCORES

These histogram matrices show the distribution of the unnormalized and normalized scores compared against the truth model coefficients. For all figures here, the scores have been scaled back to the -9 to +9 ranges for more intuitive interpretation when compared to the data collection activities.

Figure 74 and Figure 75 show the relationship scores before normalization. Figure 76 and Figure 77 show the relationship scores after normalization. One of the most noticeable effects of the normalization is the push of scores to the middle. This is more obvious here because the ranges of each histogram were not rescaled to fit the existing data.

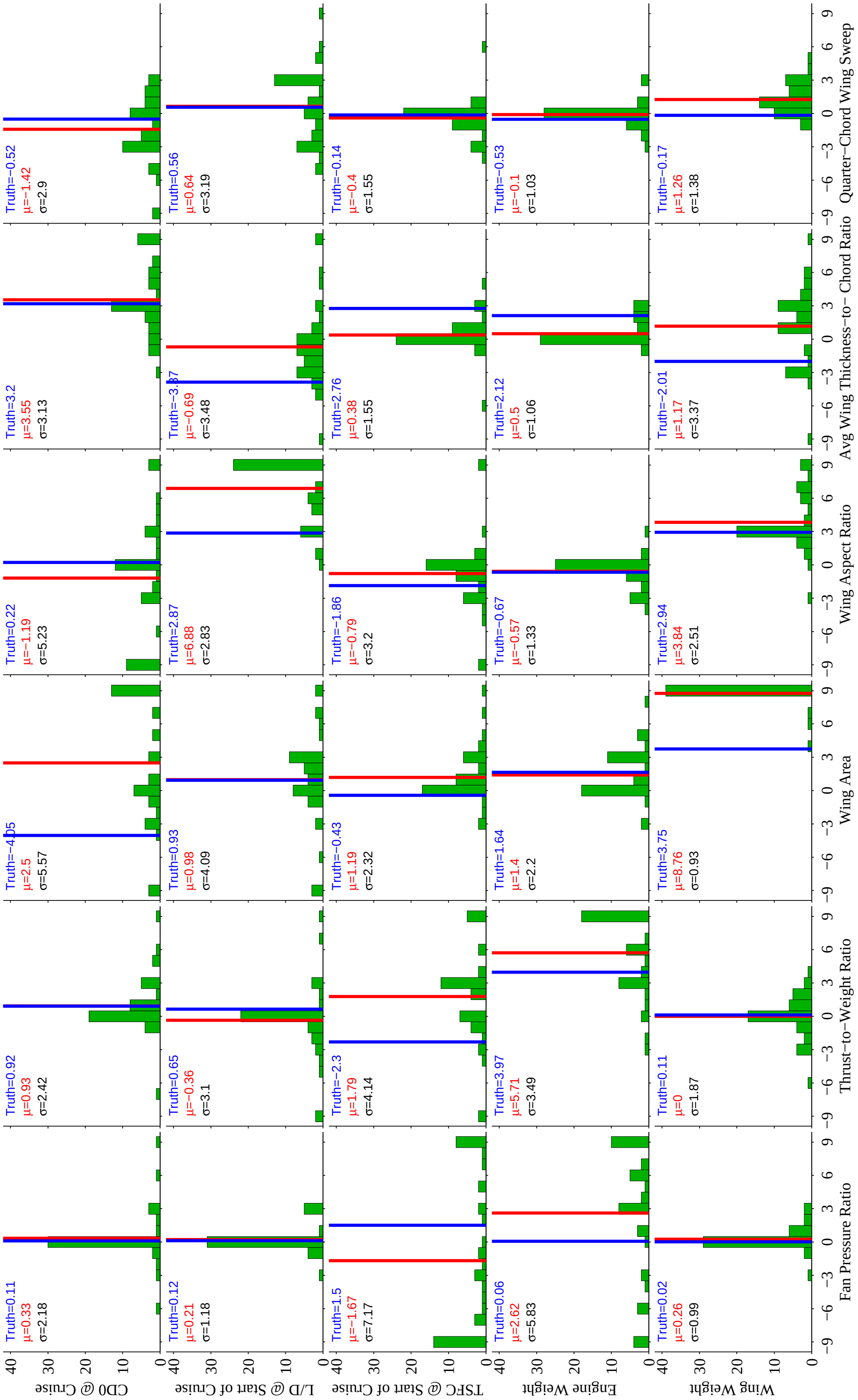


Figure 74: Distributions of Unscaled Scores for Intermediate Metrics vs. Design Variables with Truth Model Coefficients

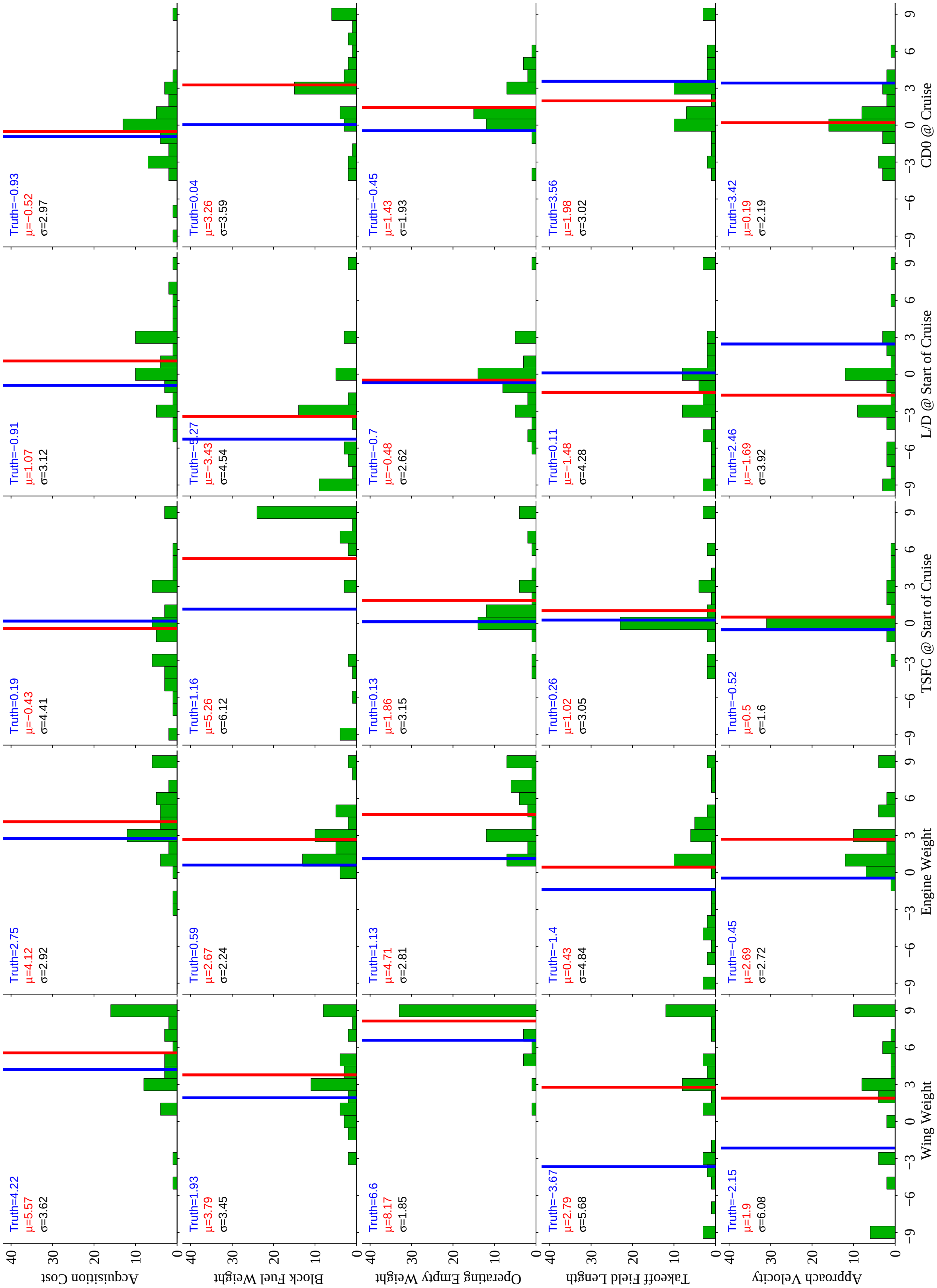


Figure 75: Distributions of Unscaled Scores for Requirements vs. Intermediate Metrics with Truth Model Coefficients

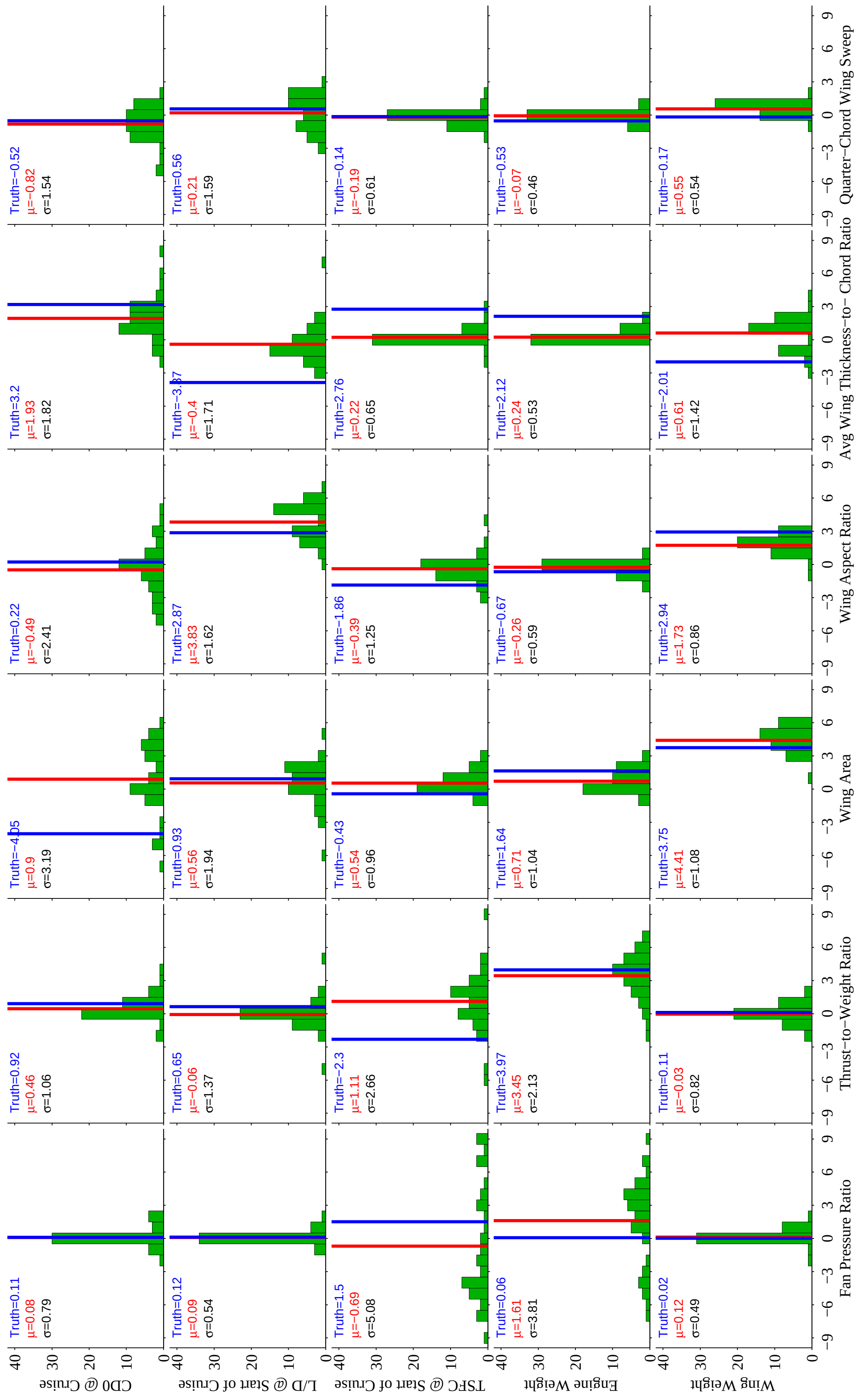


Figure 76: Distributions of Normalized Unscaled Scores for Intermediate Metrics vs. Design Variables with Truth Model Coefficients

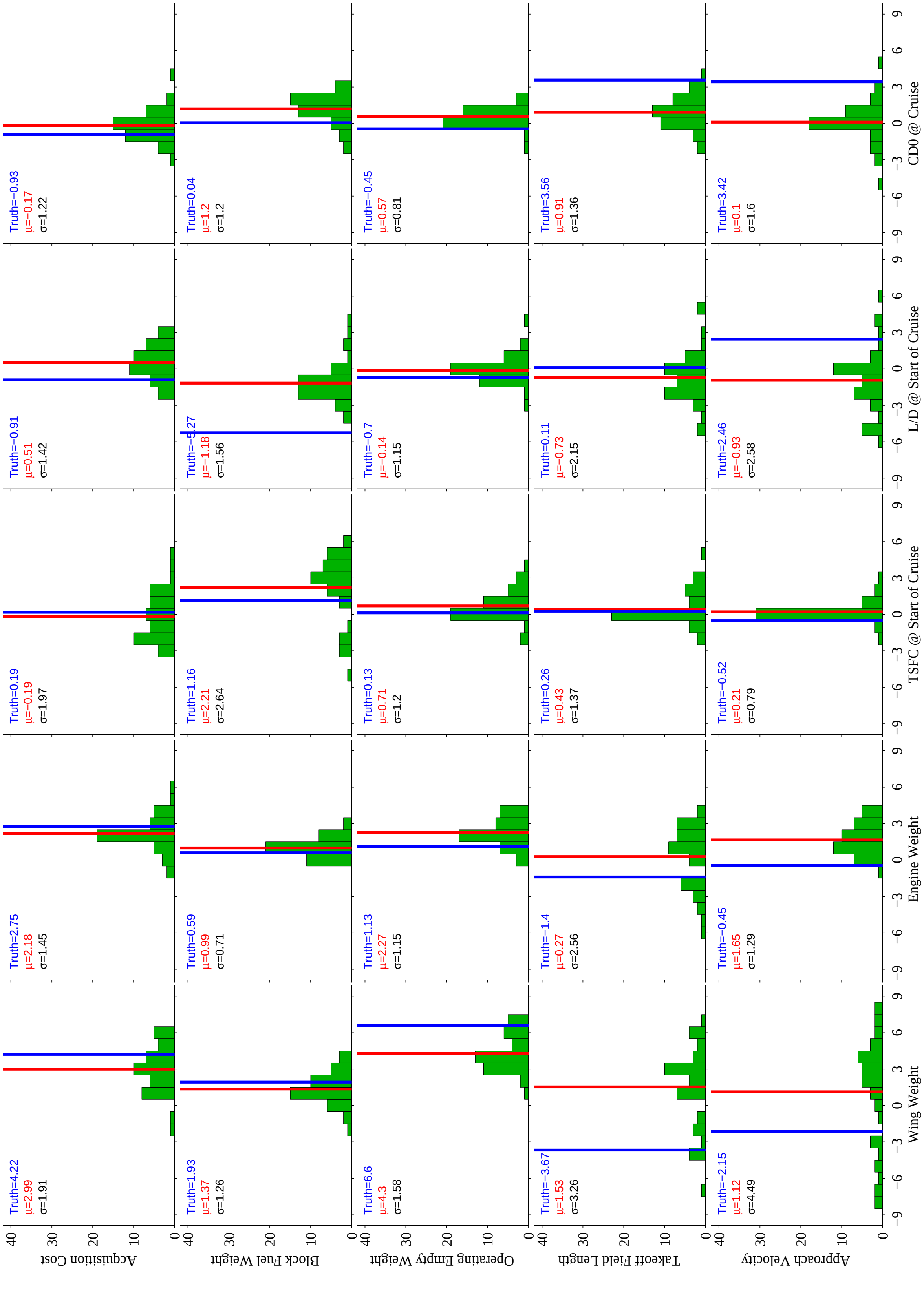


Figure 77: Distributions of Normalized Scores for Requirements vs. Intermediate Metrics with Truth Model Coefficients

APPENDIX C

EXPERTS’ ESTIMATES FOR RANGES

The need for estimates of the ranges of the intermediate metrics and system requirements was discussed in Section 3.6.1. That information was collected along with the relationships. Because they were presented as open-ended rather than selecting from a list or within bounds, there was less agreement than with the relationships, making systematic analysis difficult.

Figures 78 through 87 show the estimates that experts made for the system requirements and intermediate metrics. The ranges for the design variables were given and so are the same for all volunteers. In these figures, each black bar represents the range from estimated minimum value to estimated maximum value. The data here includes the corrections discussed in Section 5.2.3.4.

Several plots do not show all the data from all participants when those estimates were far outside the bounds of most other participants. If this data had been included, the scales of the plots would have been too large to be useful to interpreting the other values. Figure 82 does not include one estimate of $500 to $800 million. Figure 85 does not show the three participants who estimated values of TSFC over 1. Figure 87 excludes seven participants who estimated ranges of C_{D0} greater than 0.1.

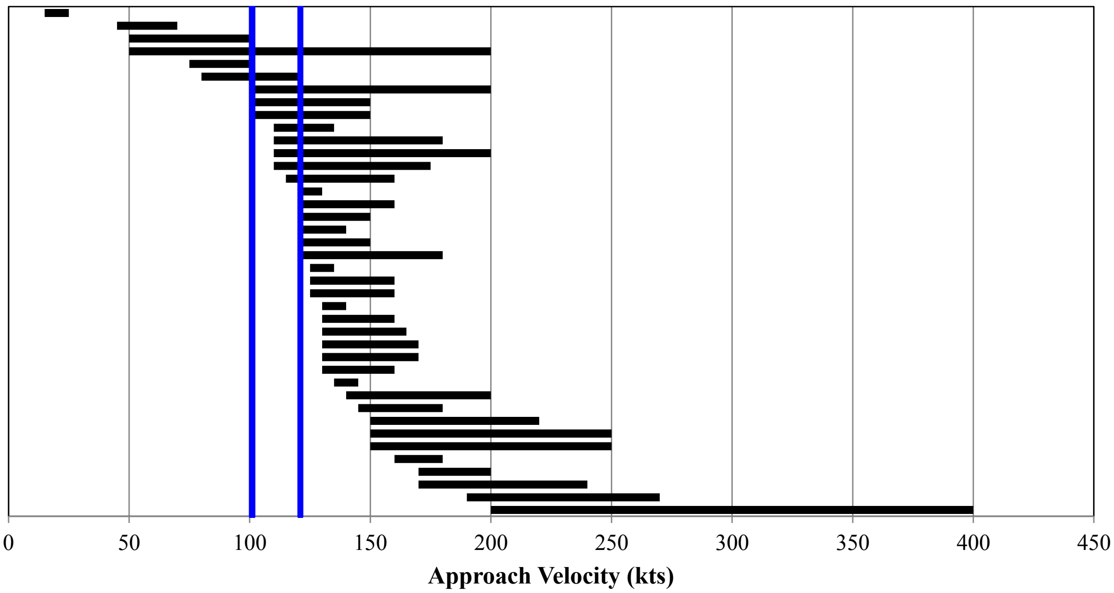


Figure 78: Experts' Estimates of Ranges of Possible Values for Approach Velocity

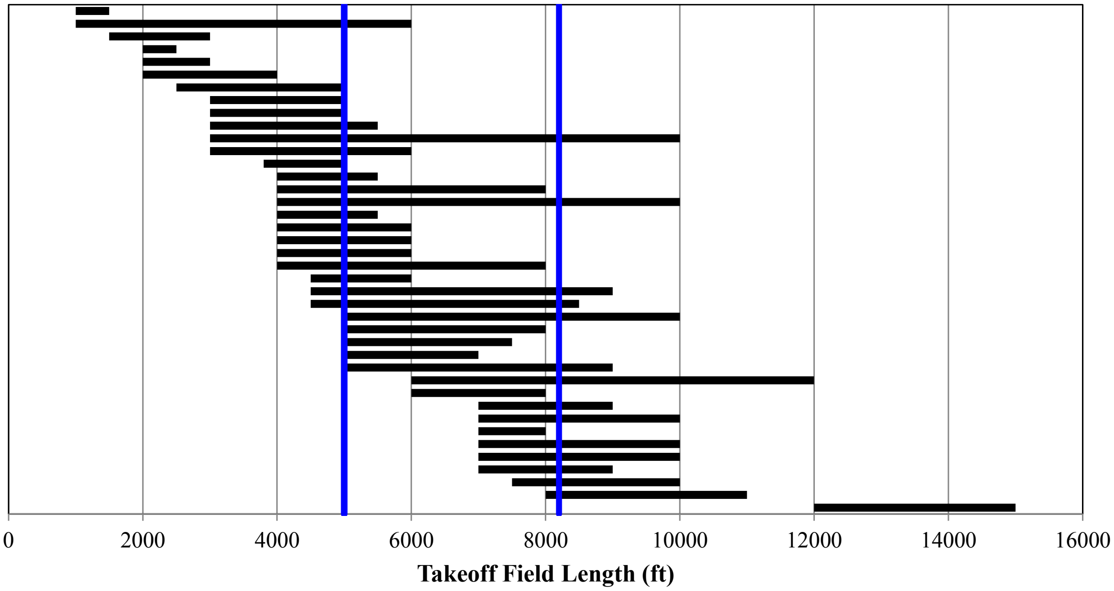


Figure 79: Experts' Estimates of Ranges of Possible Values for Takeoff Field Length

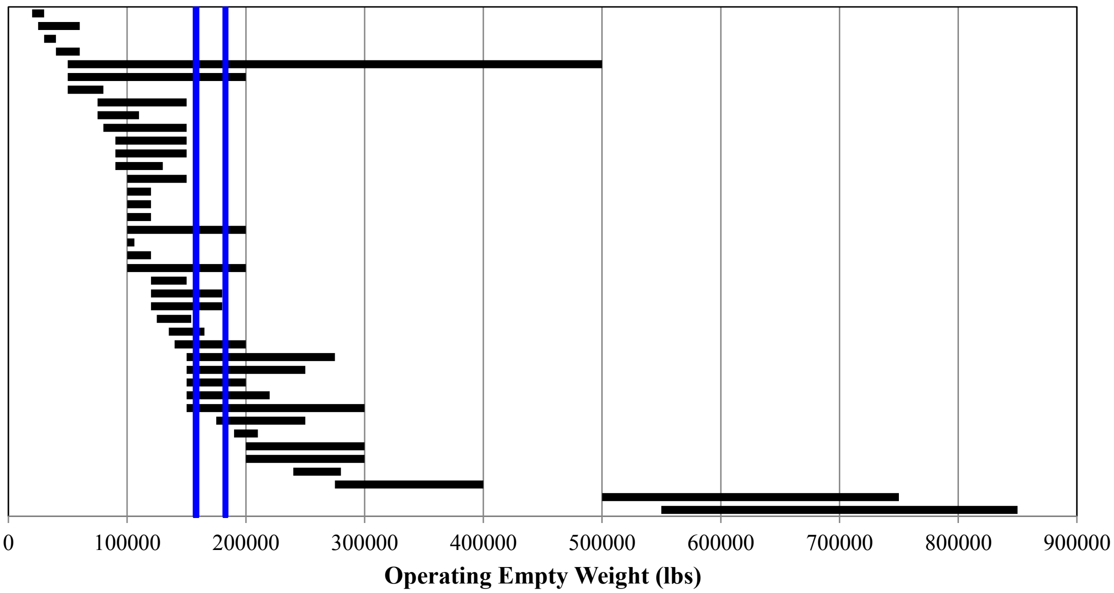


Figure 80: Experts' Estimates of Ranges of Possible Values for Operating Empty Weight

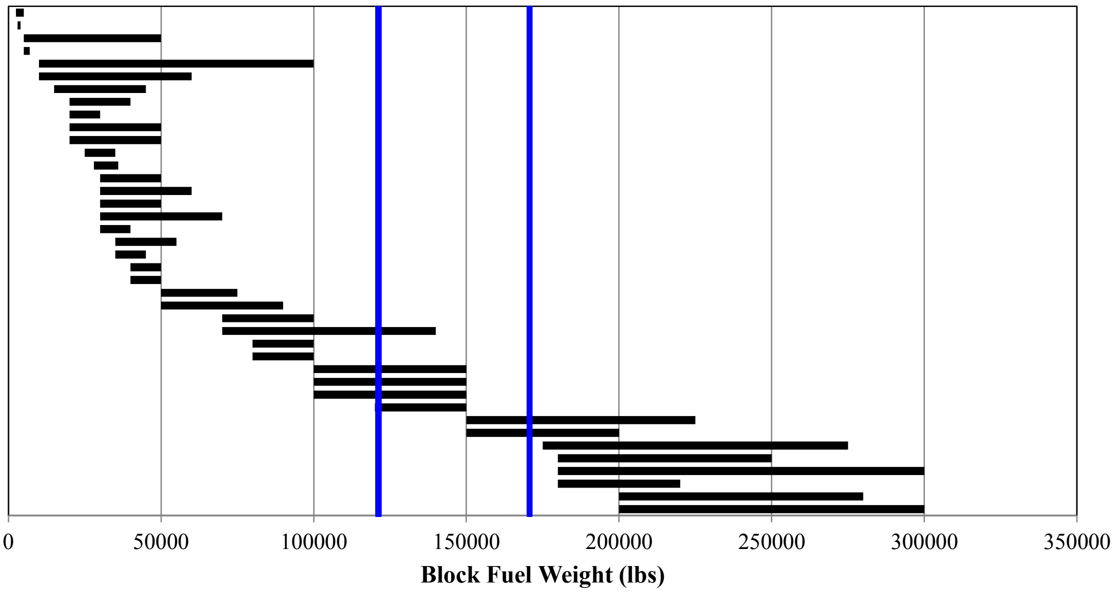


Figure 81: Experts' Estimates of Ranges of Possible Values for Block Fuel Weight

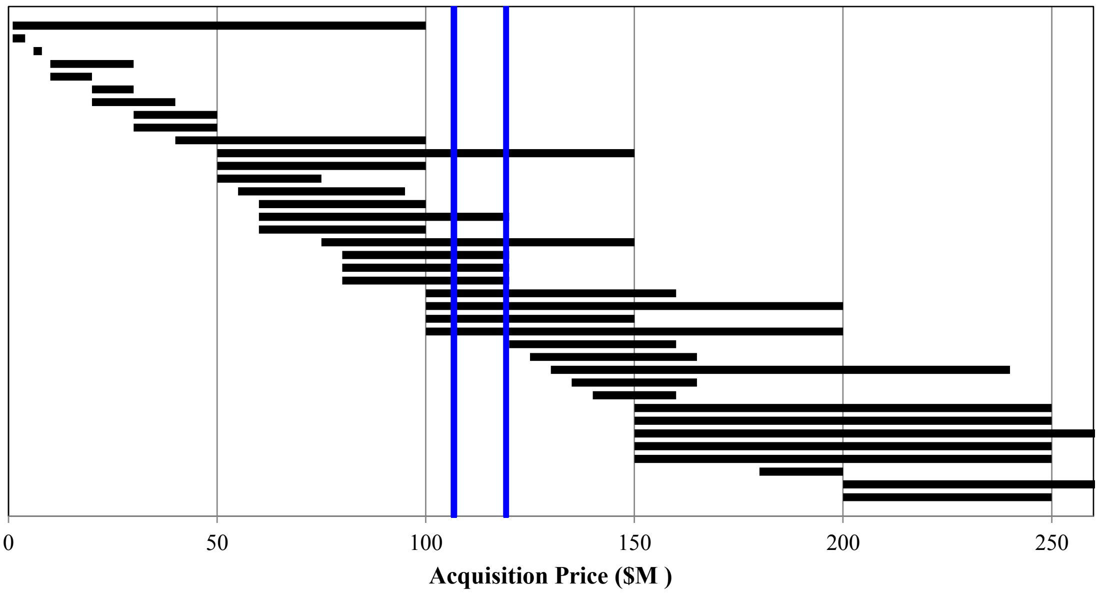


Figure 82: Experts' Estimates of Ranges of Possible Values for Acquisition Price

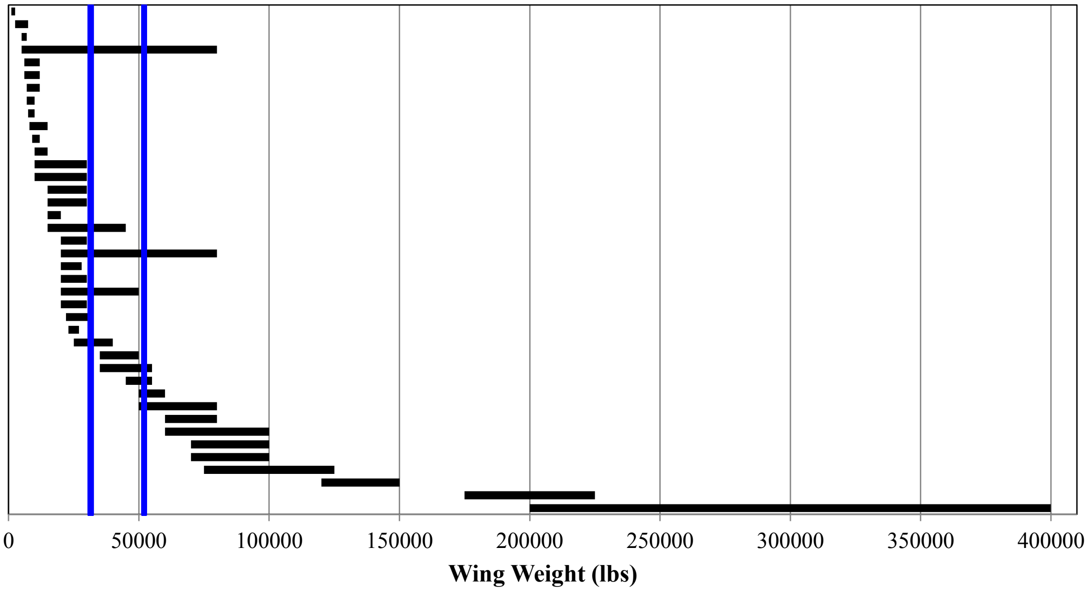


Figure 83: Experts' Estimates of Ranges of Possible Values for Wing Weight

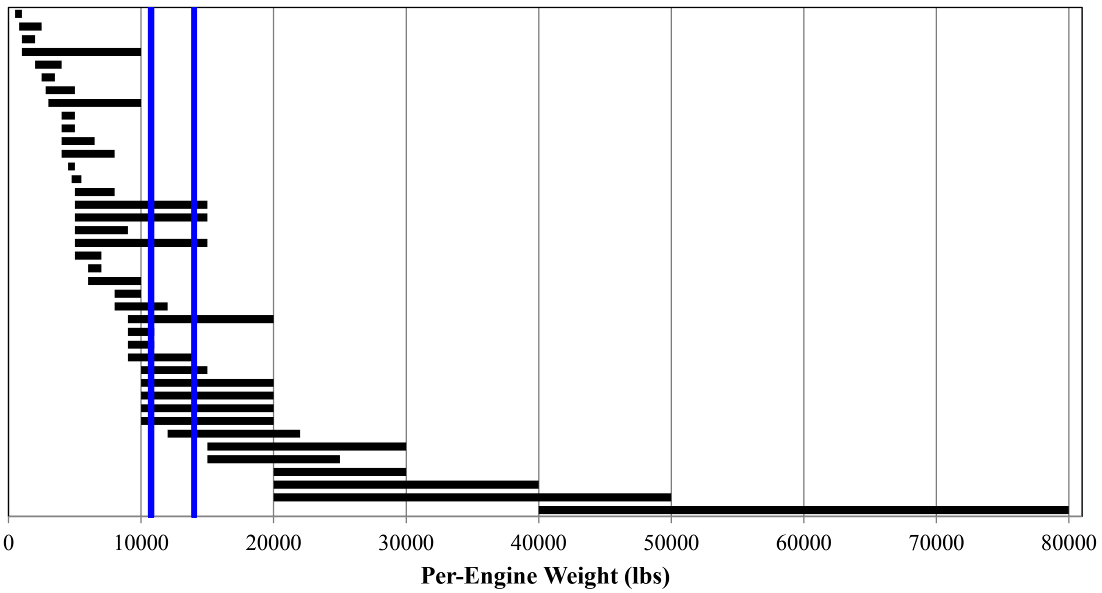


Figure 84: Experts' Estimates of Ranges of Possible Values for Weight of Each Engine

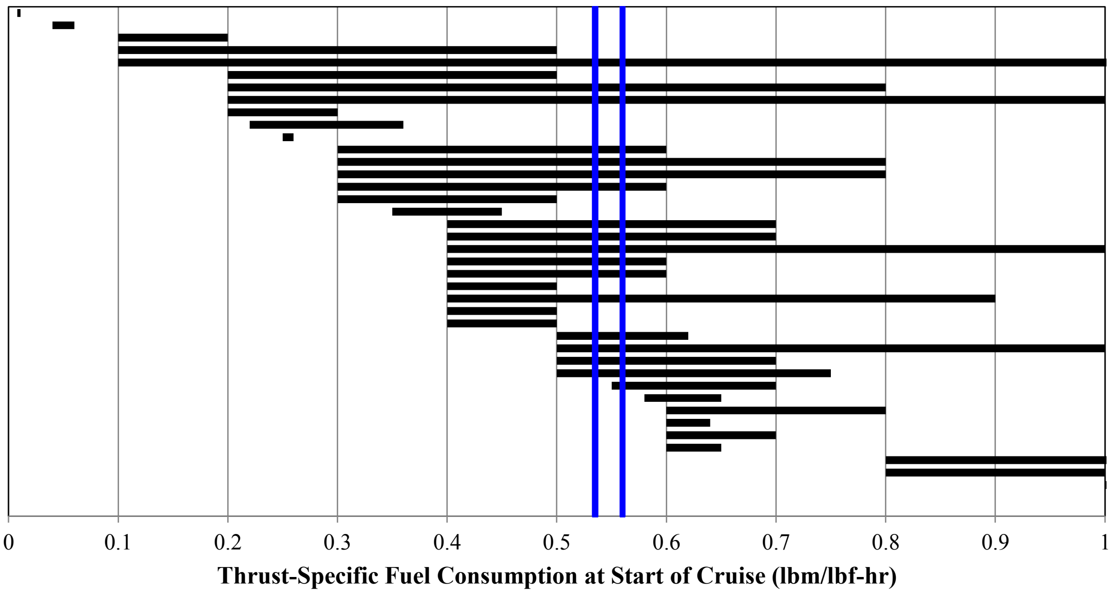


Figure 85: Experts' Estimates of Ranges of Possible Values for Thrust-Specific Fuel Consumption at Start of Cruise

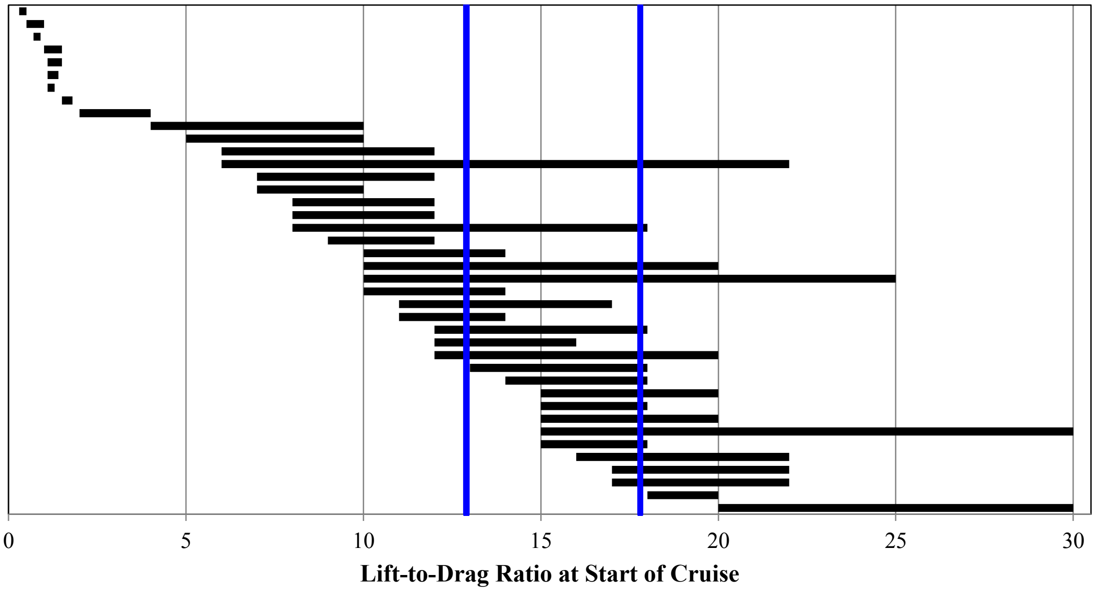


Figure 86: Experts' Estimates of Ranges of Possible Values for Lift-to-Drag Ratio at Start of Cruise

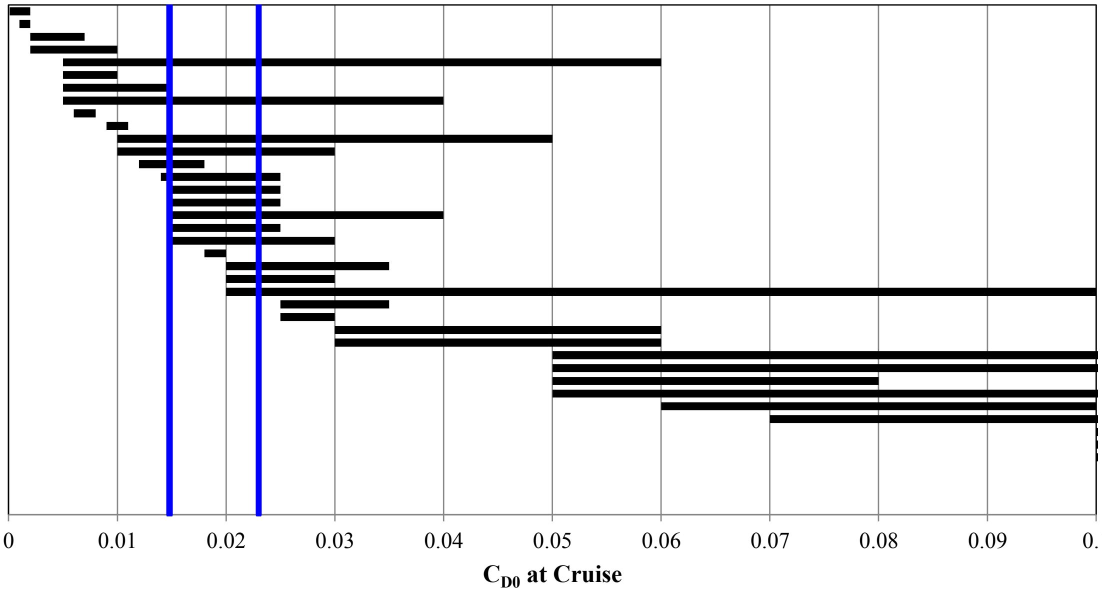


Figure 87: Experts' Estimates of Ranges of Possible Values for Parasite Drag Coefficient at Cruise

REFERENCES

[1] Airbus, *A310 Airplane Characteristics for Airport Planning*, December 2009.

[2] Akao, Y. and Mazur, G. H., "The Leading Edge in QFD: Past, Present and Future," *International Journal of Quality and Reliability Management*, vol. 20, no. 1, pp. 20–35, 2003.

[3] Anderberg, M. R., *Cluster Analysis for Applications*. Academic Press, 1973.

[4] Armstrong, J. S., *Long-Range Forecasting*. Wiley-Interscience, 1985.

[5] Arrow, K. J., *Social Choice and Individual Values*. John Wiley & Sons, second edition ed., 1963.

[6] Aspinall, W., "A route to more tractable expert advice," *Nature*, vol. 463, pp. 294–295, 21 January 2010.

[7] Associated Press, "Pentagon Reopens Bidding for Aerial Tankers and Refines Expectations," *New York Times*, p. C4, August 6 2008.

[8] Balci, O., "Verification, Validation, and Accreditation," in *Proceedings of the 1998 Winter Simulation Conference* (Medeiros, D. J., Watson, E., Carson, J. S., and Manivannan, M., eds.), IEEE, 1998.

[9] Banks, J., Gerstein, D., and Searles, S. P., "Modeling Processes, Validation, and Verification of Complex Simulations: A Survey," in *Methodology and Validation*, Society for Computer Simulation, 1987.

[10] Bartlett, III, J. E., Kotrlik, J. W., and Higgins, C. C., "Organizational Research: Determining Appropriate Sample Size in Survey Research," *Information Technology, Learning, and Performance Journal*, vol. 19, pp. 43–50, Spring 2001.

[11] Bedford, "Expert Elicitation for Reliable System Design," *Statistical Science*, vol. 21, no. 4, pp. 428–450, 2006.

[12] Bellomo, N. and Preziosi, L., *Modelling Mathematical Methods and Scientific Computation*. CRC Mathematical Modelling Series, CRC Press, 1995.

[13] Berg, J. E., Nelson, F. D., and Rietz, T. A., "Prediction Market Accuracy in the Long Run," *International Journal of Forecasting*, vol. 24, pp. 285–300, 2008.

[14] Biltgen, P. T., "Concept Identification and Selection Methods for a Solid-Propellant Target Vehicle to Support the Development of a National Missile Defense (NMD) System," AE8900 Special Problems report, School of Aerospace Engineering, Georgia Institute of Technology, 2004.

[15] Biltgen, P. T., *A Methodology for Capability-Based Technology Evaluation for Systems-of-Systems*. PhD thesis, School of Aerospace Engineering, Georgia Institute of Technology, March 2007.

[16] Blilie, C., *The Promise and Limits of Computer Modeling*. World Scientific Publishing Co., 2007.

[17] Boehm, B. W., "A Spiral Model of Software Development and Enhancement," *Computer*, vol. 21, pp. 61–72, May 1988.

[18] Boeing Commercial Airplanes, *767 Airplane Characteristics for Airport Planning*, September 2005.

[19] Boeing Company, "Commercial Airplanes: Jet Prices." Last accessed Jan 14, 2013.

[20] Bossert, J. L., *Quality Function Deployment: A Practicioner's Approach*. ASQC Quality Press, 1991.

[21] Box, G. E. P. and Draper, N. R., *Empirical Model-Building and Response Surfaces*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, 1st ed., 1987.

[22] Boyne, W. J., "Airpower Classics: SR-71 Blackbird," *Air Force Magazine*, p. 152, May 2012.

[23] Brannen, K., "Competition upended in JLTV program," *Army Times*, March 31 2012. `http://www.armytimes.com/news/2012/03/army-competition-upended-in-jltv-program-033112w/`.

[24] Bullock, R. O., "Analysis of Reynolds Number and Scale Effects on Performance of Turbomachinery," *Journal of Engineering for Power*, vol. 86, pp. 247–256, July 1964.

[25] Buonanno, M. A., *A Method for Aircraft Concept Exploration using Multicriteria Interactive Genetic Algorithms*. PhD thesis, School of Aerospace Engineering, Georgia Institute of Technology, December 2005.

[26] Burke, E., Jack M. Kloeber, J., and Deckro, R. F., "Using and Abusing QFD Scores," *Quality Engineering*, vol. 15, pp. 9–21, March 2002.

[27] Butler, A. and Doyle, J. M., "Defense Dept. says new competition addresses faults in KC-X found by auditors," *Aviation Week and Space Technology*, vol. 169, August 11 2008.

[28] Calandra, A., "Angels on a Pin," *AIChE Journal*, vol. 15, p. 163, March 1969.

[29] Carnevalli, J. A. and Miguel, P. C., "Review, analysis and classification of the literature on QFD – Types of research, difficulties and benefits," *International Journal of Production Economics*, vol. 114, pp. 737–754, 2008.

[30] Chan, L.-K. and Wu, M.-L., "Quality Function Deployment: A Comprehensive Review of Its Concepts and Methods," *Quality Engineering*, vol. 15, no. 03, pp. 23–35, 2002.

[31] Chan, L.-K. and Wu, M.-L., "Quality Function Deployment: A Literature Review," *European Journal of Operational Research*, vol. 143, pp. 463–497, 2002.

[32] Chan, L.-K. and Wu, M.-L., "A systematic approach to quality function deployment with a full illustrative example," *Omega*, vol. 33, pp. 119–139, 2005.

[33] Chuang, P.-T., "Combining the Analytic Hierarchy Process and Quality Function Deployment for a Location Decision from a Requirement Perspective," *The International Journal of Advanced Manufacturing Technology*, vol. 18, no. 11, pp. 842–849, 2001.

[34] Chung, C. A., *Simulation Modeling Handbook: A Practical Approach*. Industrial and Manufacturing Engineering Series, CRC Press, 2004.

[35] Cooke, R. M., *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, 1991.

[36] Cooke, R. M., ElSaadany, S., and Huang, X., "On the Performance of Social Network and Likelihood Based Expert Weighting Schemes," *Reliability Engineering & System Safety*, vol. 93, pp. 745–756, May 2008.

[37] Cooke, R. M. and Goossens, L. L. H. J., "TU Delft Expert Judgment Data Base," *Reliability Engineering & System Safety*, vol. 93, pp. 657–674, May 2008.

[38] Cross, M. and Moscardini, A. O., *Learning the Art of Mathematical Modelling*. Ellis Horwood Series in Mathematics and its Applications, Ellis Horwood Limited, 1985.

[39] Dalkey, N. and Helmer, O., "An Experimental Application of the Delphi Method to the Use of Experts," Tech. Rep. RM-727/1-Abridged, The RAND Corporation, July 1962.

[40] Davis, P. K., "Generalizing Concepts and Methods of Verification, Validation, and Accreditation (VV&A) for Military Simulations," Tech. Rep. R-4249-ACQ, RAND Corporation, 1992.

[41] Davis, P. K. and Dreyer, P., "RAND's Portfolio Analysis Tool (PAT) Theory, Methods, and Reference Manual," Technical Report TR756, RAND Corporation, 2009.

[42] Department of Defense, Defense Modeling and Simulation Office, "Verification, Validation and Accreditation Recommended Practices Guide." Electronic, 2006. `http://vva.msco.mil/Default.htm`.

[43] Dieter, G. E., *Engineering Design: A Materials and Processing Approach*. McGraw-Hill Series in Mechanical Engineering, McGraw-Hill, 3rd ed., 2000.

[44] Dreyer, P. and Davis, P. K., "A Portfolio-Analysis Tool for Missile Defense (PAT-MD) Methodology and User's Manual," Tech. Rep. TR262, RAND Corporation, 2005.

[45] Dym, C. L. and Ivey, E. S., *Principles of Mathematical Modeling*. Academic Press, 1980.

[46] Eggink, J., "Krippendorff's Alpha." MATLAB Central: File Exchange, Apr 2012. `http://www.mathworks.com/matlabcentral/fileexchange/36016-krippendorffs-alpha`.

[47] Ender, T. R., *A Top-Down, Hierarchical, System-of-Systems Approach to the Design of an Air Defense Weapon*. PhD thesis, School of Aerospace Engineering, Georgia Institute of Technology, 2006.

[48] Engler, W. O., "Long Range Strike System Design Case Study," in *Architecture and Principles of Systems Engineering* (Dickerson, C. E. and Mavris, D. N., eds.), Complex and Enterprise Systems Engineering Series, ch. 19, pp. 417–439, CRC Press, 2010.

[49] Engler, W. O., Biltgen, P. T., and Mavris, D. N., "Concept Selection Using an Interactive Reconfigurable Matrix of Alternatives (IRMA)," in *45th AIAA Aerospace Sciences Meeting and Exhibit*, vol. 10, American Institute of Aeronautics and Astronautics, 2007.

[50] Feickert, A., "Joint Light Tactical Vehicle (JLTV): Background and Issues for Congress," tech. rep., Congressional Research Service, April 5 2011.

[51] Fitzharris, A., "Simulation Modelling for Undergraduate Mathematicians," in *Teaching and Learning Mathematical Modelling* (Houston, S. K., Blum, W., Huntley, I., and Niell, N., eds.), ch. 30, pp. 373–384, Albion Publishing, 1997.

[52] Forsberg, K. and Mooz, H., "The Relationship of System Engineering to the Project Life Cycle," in *Joint Conference of National Council on Systems Engineering (NCOSE) and American Society for Engineering Management (ASEM)*, 1991.

[53] Franceschini, F., *Advanced Quality Function Deployment*. St. Lucie Press, 2002.

[54] Franceschini, F. and Rupil, A., "Rating scales and prioritization in QFD," *International Journal of Quality and Reliability Management*, vol. 16, no. 1, pp. 85–97, 1999.

[55] Franta, W. R., *The Process View of Simulation*. Elsevier North-Holland, 1977.

[56] Frigg, R. and Hartmann, S., "Models in Science," *The Stanford Encyclopedia of Philosophy*, Feb 27 2006. `http://plato.stanford.edu/entries/models-science/` Last Accessed Jul 15, 2011.

[57] Galbraith, P. L. and Haines, C. R., "Some Mathematical Characteristics of Students Entering Applied Mathematics Degree Courses," in *Teaching and Learning Mathematical Modelling* (Houston, S. K., Blum, W., Huntley, I., and Niell, N., eds.), pp. 77–91, Albion Publishing, 1997.

[58] Google, Inc., "Google Scholar." `http://scholar.google.com` Last Accessed May 16, 2012.

[59] Gregorski, T., "Akashi Kaikyo Bridge," *Roads and Bridges*, vol. 36, pp. 34–38, August 1998.

[60] Griendling, K. A., *ARCHITECT: The Architecture-Based Technology Evaluation and Capability Tradeoff Method.* PhD thesis, School of Aerospace Engineering, Georgia Institute of Technology, 2011.

[61] Guinta, L. R. and Praizler, N. C., *The QFD Book.* American Management Association, 1993.

[62] Hammond, K. R., *Human Judgement and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice.* Oxford University Press, 1996.

[63] Hannon, B. and Ruth, M., *Dynamic Modeling.* Springer-Verlag, second edition ed., 2001.

[64] Harvard Law Review Board of Editors, "Reliable Evaluation of Expert Testimony," *Harvard Law Review*, vol. 116, pp. 2142–2163, May 2003.

[65] Haskins, C., ed., *INCOSE Systems Engineering Handbook.* International Council on Systems Engineering, 3.2.2 ed., October 2011.

[66] Hauser, J. R. and Clausing, D., "The House of Quality," *The Harvard Business Review*, vol. May-June, pp. 63–73, 1988.

[67] Hazelrigg, G. A., "The implications of Arrow's impossibility theorem on approaches to optimal engineering design," *Journal of Mechanical Design*, vol. 118, pp. 161–164, Jun 1996.

[68] Hegyi, B., "Discussion on Current State of Experts in Meteorology," May 2012.

[69] Ho, W., "Integrated Analytic Hierarchy Process and its Applications — A Literature Review," *European Journal of Operational Research*, vol. 186, pp. 211–228, April 1 2008.

[70] Hoffman, M., "DoD pushes back deadline for JLTV bids," *DoD Buzz*, March 8th 2012. `http://www.dodbuzz.com/2012/03/08/dod-pushes-back-deadline-for-jltv-bids/`.

[71] Hornik, K., Stinchcombe, M., and White, H., "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[72] Ikeda, T., "A Case Study of Instruction and Assessment in Mathematical Modelling — 'the delivering problem'," in *Teaching and Learning Mathematical Modelling* (Houston, S. K., Blum, W., Huntley, I., and Niell, N., eds.), pp. 51–61, Albion Publishing, 1997.

[73] Kano, N., Seraku, N., Takahashi, F., and Tsuji, S., "Attractive Quality and Must-be Quality," *Journal of the Japanese Society for Quality Control*, vol. 14, no. 2, pp. 39–48, 1984.

[74] Kirby, M. R. and Mavris, D. N., "The Environmental Design Space," in *26th International Congress of the Aeronautical Sciences*, 2008.

[75] Kirby, M. R., Raczynski, C., and Mavris, D., "An Approach for Strategic Planning of Future Technology Portfolios," in *6th AIAA Aviation Technology, Integration and Operations Conference*, September 25–27 2006.

[76] Kirby, M. R., *A Methodology for Technology Identification, Evaluation, and Selection in Conceptual and Preliminary Aircraft Design*. PhD thesis, School of Aerospace Engineering, Georgia Institute of Technology, 2001.

[77] Krippendorff, K., "Estimating the Reliability, SystSystem Error, and Random Error of Interval Data.," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[78] Krippendorff, K., "Reliability in Content Analysis: Some Common Misconceptions and Recommendations," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[79] Krippendorff, K., "Computing Krippendorff's Alpha-Reliability," departmental paper, Annenberg School for Communication, University of Pennsylvania, 2011.

[80] Krippendorff, K., *Content Analysis: An Introduction to its Methodology*. SAGE Publications, 3rd ed., 2013.

[81] Kurowski, P. M., *Finite Element Analysis for Design Engineers*. Society of Automotive Engineers, 2004.

[82] Lee, K., Nam, T., Perullo, C., and Mavris, D. N., "Reduced-Order Modeling of a High-Fidelity Propulsion System Simulation," *AIAA Journal*, vol. 49, pp. 1665–1682, August 2011.

[83] Legree, P. J., Psotka, J., Tremble, T., and Bourne, D. R., "Using Consensus Based Measurement to Assess Emotional Intelligence," in *Emotional Intelligence: An International Handbook*, Hogrefe & Huber Publishers, 2004.

[84] Liu, Y.-C., Bligh, T., and Chakrabarti, A., "Towards an 'ideal' approach for concept generation," *Design Studies*, vol. 24, pp. 341–355, 2003.

[85] Malkinson, T., "World Bytes: Sex Bias in Research," *IEEE-USA Today's Engineer*, June 2012. `http://www.todaysengineer.org/2012/Jun/worldbytes.asp` Last accessed July 21, 2012.

[86] Mattson, C. A. and Messac, A., "Concept Selection Using s-Pareto Frontiers," *AIAA Journal*, vol. 41, pp. 1190–1198, June 2003.

[87] Matzler, K. and Hinterhuber, H. H., "How to Make Product Development Projects More Successful by Integrating Kano's Model of Customer Satisfaction into Quality Function Deployment," *Technovation*, vol. 18, no. 1, pp. 25–38, 1998.

[88] Mavris, D., Galloway, T., Marx, W., and Garcia, E., *Aircraft Life Cycle Cost Analysis*. Aerospace Systems Design Laboratory, September 2001. Version 6.0.

[89] Mavris, D. N., Bagdatli, B., and Miller, M., "AE6373 Advanced Design Methods I Design Project Description." School of Aerospace Engineering, Georgia Institute of Technology, Fall 2011.

[90] Mavris, D. N. and DeLaurentis, D., "Methodology for Examining the Simultaneous Impact of Requirements, Vehicle Characteristics, and Technologies on Military Aircraft Design," in *22nd Congress of the International Council on the Aeronautical Sciences*, 2000.

[91] Mavris, D. N. and Griendling, K., "Relational Oriented Systems Engineering and Technology Tradeoff Analysis (ROSETTA) Environment," in *6th IEEE International Conference on Systems of Systems Engineering*, IEEE, 2011.

[92] McCullers, L. A., *Flight Optimization System*. NASA Langley Research Center, 8.11 ed., October 9 2009.

[93] Mihram, G. A., "Some Practical Aspects of the Verification and Validation of Simulation Models," *Operational Research Quarterly*, vol. 23, pp. 17–29, March 1972.

[94] Mihram, G. A., "The Modeling Process," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-2, pp. 621–626, November 1972.

[95] Miller, V. M., "In Pursuit of Scientific Excellence: Sex Matters," *American Journal of Physiology — Heart and Circulatory Physiology*, vol. 302, pp. H1771–H1772, May 2012.

[96] Moore, R. A., "Variable Fidelity Modeling as Applied to Trajectory Optimization for a Hydraulic Backhoe," Master's thesis, School of Mechanical Engineering, Georgia Institute of Technology, May 2009.

[97] Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M., *Response Surface Methodology*. John Wiley & Sons, 2009.

[98] Niedzwiecki, R. W., "Small Engine Technology Programs," Tech. Rep. 92N22532, NASA - Glenn Research Center, Feb 1990.

[99] Office of the Undersecretary of Defense for Aqcuisition and Technology, "DoD Guide to Integrated Product and Process Development," February 5 1996.

[100] Office of the Undersecretary of Defense for Aqcuisition and Technology, "DoD Integrated Product and Process Development Handbook," July 6 1998.

[101] O'Keefe, L. and Sierchio, J., "Developing a Mission Solution: From Mission Gap Analysis to Preferred System Concept," in *13th Annual Systems Engineering Conference*, National Defense Industrial Association, October 2010.

[102] Okudan, G. E. and Tauhid, S., "Concept selection methods – a literature review from 1980 to 2008," *International Journal of Design Engineering*, vol. 1, no. 3, pp. 243–277, 2008.

[103] Ölvander, J., Lundén, B., and Gavel, H., "A computerized optimization framework for the morphological matrix applied to aircraft conceptual design," *Computer-Aided Design*, vol. 41, pp. 187–196, 2009.

[104] Pace, D. K. and Sheehan, J., "Subject Matter Expert (SME)/Peer Use in M&S V&V," in *Foundations for V&V in the 21st Century Workshop*, October 22 2002.

[105] Peel, D. A., "Solving Problems with Mathematical Models," in *Modelling and Simulation in Practice* (Cross, M., Gibson, R. D., O'Carroll, M. J., and Wilkinson, T. S., eds.), Pentech Press, 1979.

[106] Pugh, S., *Creating innovative products using total design: the living legacy of Stuart Pugh*. Addison-Wesley Publishing Company, 1996.

[107] Raczynski, C. M., *A Methodology for Comprehensive Strategic Planning and Program Prioritization*. PhD thesis, School of Aerospace Engineering, Georgia Institute of Technology, 2008.

[108] Raczynski, C. M. and Mavris, D. N., "A Method for Strategic Technology Prioritization and Portfolio Resource Allocation," in *2011 IEEE Aerospace Conference*, 2011.

[109] Raharjo, H., *Some Further Studies on Improving QFD Methodology and Analysis*. PhD thesis, National University of Singapore, 2010.

[110] Revelle, J. B., Moran, J. W., and Cox, C. A., *The QFD Handbook*. John Wiley and Sons, 1998.

[111] Saaty, T. L., *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill, 1980.

[112] Saltmarsh, E. A., "Quantifying Arrow's Impossibility Theorem for Portfolio Management Voting Systems," AE8900 Special Problems report, School of Aerospace Engineering, Georgia Institute of Technology, 2009.

[113] Sargent, R. G., “Validation of Simulation Models: General Approach,” in *Concise Encyclopedia of Modelling and Simulation* (Atherton, D. P. and Borne, P., eds.), pp. 482–485, Pergamon Press, 1992.

[114] Schrage, D. P., “Technology for Rotorcraft Affordability Through Integrated Product/Process Development (IPPD),” in *American Helicopter Society 55th Annual Forum*, 1999.

[115] Schrage, D. P. and McCandless, W., “Integrated Product/Process Development Approach For Balancing Technology Push and Pull Between the User and Developer,” in *American Helicopter Society 68th Annual Forum*, American Helicopter Society International, Inc., May 1-3 2012.

[116] Sebastian, T. B., Crisco, J. J., Klein, P. N., and Kimia, B. B., “Constructing 2D Curve Atlases,” in *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 70–77, 2000.

[117] Shin, J.-S. and Kim, K.-J., “Effect and Choice of the Weighting Scale of QFD,” *Quality Engineering*, vol. 12, no. 3, pp. 347–356, 2000.

[118] Shuman, F. G., “History of Numerical Weather Prediction at the National Meteorological Center,” *Weather and Forecasting*, vol. 4, pp. 286–296, 1989.

[119] Sohn, S. Y., “Quality Function Deployment Applied to Local Traffic Accident Reduction,” *Accident Analysis & Prevention*, vol. 31, pp. 751–761, November 1999.

[120] Spearman, C., “The Proof and Measurement of Association between Two Things,” *American Journal of Psychology*, vol. 15, pp. 72–101, Jan 1904.

[121] Stults, I. C., *A Multi-Fidelity Analysis Selection Method Using a Constrained Discrete Optimization Formulation*. PhD thesis, School of Aerospace Engineering Georgia Institute of Technology, 2009.

[122] Supreme Court of the United States, “Federal Rules of Evidence,” 2011.

[123] Talley, D. N., *Methodology for the Conceptual Design of a Robust and Opportunistic System-of-Systems*. PhD thesis, School of Aerospace Engineering, Georgia Institute of Technology, 2008.

[124] Temponi, C., Yen, J., and Amos Tiao, W., “House of quality: A fuzzy logic-based requirements analysis,” *European Journal of Operational Research*, vol. 117, pp. 340–354, Sept. 1999.

[125] United States Code of Federal Regulations, “Protection of Human Subjects.” United States Department of Health and Human Services, January 2009. 45 CFR 46.

[126] Vansteenkiste, G. C. and Spriet, J. A., “Modelling Ill-Defined Systems,” in *Progress in Modelling and Simulation* (Cellier, F. E., ed.), Academic Press, 1982.

[127] Wang, J., "Fuzzy Outranking Approach to Prioritize Design Requirements In Quality Function Deployment.," *International Journal of Production Research*, vol. 37, no. 4, pp. 899–916, 1999.

[128] Ward, D., "Faster, Better, Cheaper Revisited: Program Management Lessons from NASA," *Defense AT&L Magazine*, pp. 48–52, March-April 2010.

[129] Waterman, D. A., *A Guide to Expert Systems*. The Teknowledge Series in Knowledge Engineering, Addison-Wesley Publishing Company, 1986.

[130] WebMD, LLC., "WebMD Symptom Checker." `http://symptoms.webmd.com/symptomchecker` Last Accessed Jul 18, 2011.

[131] Woelke, N. A., "E-mail exchange," March 2013.

[132] Wright, S. A. and Bauer, Jr., K. W., "Covalidation of Dissimilarly Structured Models," in *Proceedings of the 1997 Winter Simulation Conference* (Andradóttir, S., Healy, K. J., Withers, D. H., and Nelson, B. L., eds.), 1997.

[133] Wu, C. F. J. and Hamada, M., *Experiments: Planning, Analysis, Parameter Design and Optimization*. Wiley-Interscience, 2000.

[134] Zhou, M., "Fuzzy logic and optimization models for implementing QFD," *Computers & Industrial Engineering*, vol. 35, no. 1-2, pp. 237 – 240, 1998. Proceedings of the 23rd International Conference on Computers and Industrial Engineering.

[135] Zwicky, F., "Morphological Astronomy," *The Observatory*, vol. 68, pp. 121–143, August 1948.

[136] Zwicky, F., *Discovery, Invention, Research: Through the Morphological Approach*. The Macmillan Company, 1969.