A Standardized Model to Quantify the Financial Impact of Poor Engineering Information Quality in the Oil and Gas Industry

by

Emile Otto Coetzer

UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY

Thesis presented in fulfilment of the requirements for the degree of Master of Engineering in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof. P.J. Vlok

Co-Supervisor: Prof. C.S.L. Schutte

December 2018

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained herein is my own, original work, that I am the sole author thereof, unless explicitly stated otherwise, that reproduction thereof of the University of Stellenbosch will not infringe any third party rights to my knowledge and that I have not previously, in its entirety or in part, submitted it for obtaining any qualification.

Emile-Otto Coetzer

December 2018

Abstract

A Standardized Model to Quantify the Financial Impact of Poor Engineering Information Quality in the Oil and Gas Industry

E. O. Coetzer

Department of Industrial Engineering,

University of Stellenbosch,

Private Bag X1, Matieland 7602, South Africa

Thesis: M. Eng (Industrial)

December 2018

Industrial assets rely on thousands of data points to run safely, responsibly and profitably. The digital era has introduced the risk that control of data quality is lost. Achieving and maintaining asset data quality control is expensive. Although this issue is instinctively understood by engineers and technicians, a review of the literature indicates that the true impact of poor asset data quality is difficult to quantify. This makes it difficult to justify the expense required to rectify the deficiencies in engineering data. Consequently, the problem is often not rectified. This leads to a perpetuation of the problem and increasing risk, inefficiency and frustration. Problems surrounding engineering information quality have been implicated in several well-publicized industry disasters.

Justifying the expense is difficult because the benefits are neither immediately obvious nor able to be calculated using a defensible method. No defensible method to calculate the financial impact of engineering data quality has been found for the oil and gas industry. This research study addresses this challenge. The research objective of the present study its therefore to develop a standardized model to quantify the financial impact of poor engineering information quality in the oil and gas industry.

This study defines engineering information in the oil and gas industry as information about asset design and machinery. It is generated during design and is required throughout the asset life. The target audience is senior management in the oil and gas industry, where authority for approval for data quality initiatives is held.

A review of the literature has shown precedent in related industries, but none in the oil and gas industry. The precedent in other industries, coupled with an analysis of several potential approaches, revealed that a survey-based research design was appropriate for this problem. A survey questionnaire was therefore developed from a literature review and validated during a series of structured interviews at an operating asset. The contents of the validated survey questionnaire indicated that the financial impact of poor engineering information quality consist of the four categories of productivity loss, increased cost, reduced production and increased risk.

Using the survey questionnaire as a basis, a model was developed to calculate the cost of poor engineering information quality, both deterministically and stochastically. Following a review of commonly used numerical methods, it was concluded that Monte Carlo simulation was the most applicable approach for the stochastic model. Data collected during the survey validation structured interviews was used to populate a laboratory data set, which was used to test the model.

The construction and testing of the model enabled a case study of actual field data from another operating asset. The results of the case study were discussed and interpreted in the thesis.

The results of the model are intended to serve as inputs for senior managers to assign funding to engineering information quality improvement. In order to present the data in the most acceptable form, a review of the literature around organisation decision-making and information presentation requirements was undertaken. The review indicated that the target audience was comfortable with uncertainty but was at risk of cognitive strain. The cognitive strain could be reduced by presenting information graphically and reporting the confidence of the result. An appropriate data presentation and management report was therefore developed. This included reporting results in Pareto form. For this reason, a taxonomy was developed and validated by a series of unstructured interviews with senior managers. These Pareto results enable the prioritisation of data quality improvement drives.

Both the initial structured interviews and case study results proved the original contention that the cost of poor engineering information quality is not insignificant and presents an opportunity for improvements in the oil and gas industry that is competitive with other opportunities.

This study is a first exploration of the subject. Many opportunities for future research have been identified, including more sophisticated statistical models, exploration of causality and the mechanistic properties of poor engineering information quality.

Opsomming

‘n Gestandardiseerde Model vir die Berekening van die Finansiële Impak van Ontoereikende Kwaliteit van Ingenieurs-Inligting in die Energie-Industrie

E. O. Coetzer

Department van Bedryfsingenieurswese,

Universiteit van Stellenbosch,

Private Bag X1, Matieland, 7602, Suid-Afrika

Tesis: M.Ing (Bedryfs)

Desember 2018

Industriële aanlegte maak staat op duisende data-punte om veilig, verantwoordelik en winsgewend te kan bedryf. Die digitale era het ‘n nuwe risiko meegebring: die moontlikheid dat beheer oor die kwaliteit van die data verloor kan word. Die daarstelling en instandhouding van aanleg-data van toereikende kwaliteit is duur. Alhoewel hierdie probleem instinktief verstaan word deur ingenieurs en tegnici in die industrie, dui ‘n oorsig van die literatuur aan dat dit kompleks is om die ware impak van ontoereikende kwaliteit van aanleg-data te kwantifiseer. Dit maak dit moeilik om die onkoste te regverdig om die vereiste kwaliteit te bereik. Die probleem word gevolglik dikwels nie aangespreek nie, wat voortgesette verhoogde risiko, oneffektiwiteit en frustrasie tot gevolg het. Probleme rondom die kwaliteit van ingenieurs-inligting word aangehaal in verskeie hoogs-gepubliseerde industriële rampe.

Die regverdiging van die onkoste is moeilik omdat die voordele van data van die aangewese kwaliteit beide nie voor die hand liggend is nie, en ‘n geloofwaardige metode om dit te bereken nie beskikbaar is nie. Geen verdedigbare metode is gevind om die finansiële impak van lae-kwaliteit aanleg-data in die energie-industrie te bereken nie. Hierdie navorsing spreek hierdie leemte aan.

Ingenieurs-inligting in die energie-industrie word in hierdie studie gedefinieer as inligting wat verband hou met die ontwerp van industriële aanlegte en gepaardgaande masjinerie. Hierdie inligting word grootliks gegenereer tydens ontwerp en word benodig tydens die totale leeftyd van die aanleg. Die navorsingsdoelwit is om ‘n gestandardiseerde model te ontwikkel vir die berekening van die finansiële impak van ontoereikende ingenieurs-inligting in die energie-industrie. Die doelwitgehoor van die navorsing is senior bestuurders in die energie-industrie, waar die goedkeuring gesetel is om fondse te bewillig.

‘n Oorsig van die literatuur toon aan dat daar geen so ‘n metode in die energie-industrie bestaan nie, maar dat verwante industrieë alreeds soortgelyke studies aangepak het.

Hierdie voorbeelde, tesame met ‘n analise van verskeie potensiële navorsingsbenaderings, wys daarop dat ‘n opname-metode aangewese is vir hierdie probleem. ‘n Opname-vraelys is gevolglik ontwikkel vanuit ‘n literatuurstudie, en is bekragtig deur middel van ‘n reeks gestruktureerde onderhoude by ‘n aanleg wat tans in bedryf is. Die inhoud van die finale vraelys dui daarop dat die koste van ontoereikende kwaliteit van ingenieurs-inligting toegeskryf kan word aan vier hoof-kategorieë, naamlik verlaagde produktiwiteit, addisionele koste, verminderde produksie en verhoogde risiko.

Op grond van die vraelys is ‘n model ontwikkel om the koste van ontoereikende kwaliteit van ingenieurs-inligting te bereken, beide deterministies en stochasties. Vir die stochastiese model is ‘n Monte-Carlo-simulasie gekies, gebaseer ‘n oorsig van algemene numeriese metodes. Tipiese grootte-ordes vir elke vraag in die vraelys is ook verkry tydens die gestruktureerde onderhoude. Die model is getoets deur gebruik te maak van hierdie laboratorium-data.

Die beskikbaarheid van ‘n bewese model het ‘n gevallestudie moontlik gemaak. Werklike opname-data vanuit ‘n ander aanleg is aangewend en werklike resultate in die industrie is bereken. Die resultate van die gevallestudie word in die tesis bespreek en geinterpreteer.

Die doel van die resultate van die model is om as insette te dien sodat senior bestuurders kan evalueer of fondse bewillig moet word om die kwaliteit van ingenieurs-inligting te verbeter. Ten einde die resultate in die mees aanvaarbare manier aan te bied, is ‘n oorsig gedoen van die literatuur rondom besluitnemingsdinamika in organisasies en die formaat waarin resultate voorgelê moet moet. Die literatuuroorsig se slotsom is dat die teikengehoor gemaklik is met onsekerheid, maar grafiese aanbieding verkies en geneig is tot kognitiewe spanning. Gevolglik is ‘n bestuursverslag ontwikkel wat die aanbieding van resultate in die aangewese formaat bevat. Die bestuursverslag sluit ook die aanbieding van resultate in Pareto-formaat in. Vir hierdie doel is ‘n hierargie ontwikkel en voorgelê vir kommentaar aan ‘n aantal senior bestuurders. Hierdie Pareto-verslae maak dit moontlik om aktiwititeite rondom die verbetering van die kwaliteit van ingenieurs-inligting te kan prioritiseer.

Beide die aanvanklike onderhoude en die gevallestudie-data bevestig die aanvanklike standpunt dat die koste van ontoereikende kwaliteit van ingenieurs-inligting beduidend genoeg is om te kan kompeteer vir befondsing met ander verbeterings-inisiatiewe.

Hierdie studie is ‘n eerste verkenning van die onderwerp. Verskeie geleenthede vir verdere navorsing is ontbloot, insluitende die ontwikkeling van meer gesofistikeerde statistiese modelle en verkenning van die eienskappe en oorsake van ontoereikende kwaliteit van ingenieurs-inligting.

Acknowledgements

Our Heavenly Father, who arranged the circumstance to make this adventure possible.

My family, who remained supportive, despite.

Prof P.J. Vlok, who remained patient and uncompromising, despite.

Antonio, Colin, Dave, David, Hugo, Jamie and Tim, who coached, criticized and encouraged.

Renette and Patrick, whose editing magic repeatedly saved the day.

Table of Contents

List of Figures

List of Tables

Appendices

List of Acronyms and Abbreviations

AI	Artificial Intelligence
API	American Petroleum Institute
AOGHS	American Oil & Gas Historical Society
ARD	Asset Reference Data
BOE	Barrel Oil Equivalent
BTU	British Thermal Unit
CAD	Computer-Aided Drafting
CDF	Cumulative Distribution Function
CFTE	Total Annual Cost of an FTE
CLT	Central Limit Theorem
CoPEIQ	Cost of Poor Engineering Information Quality
EI	Engineering Information
ER	Entity – Relationship (Data Model)
EIM	Engineering Information Management
EIQ	Engineering Information Quality
EPC	Engineering, Procurement and Construction (Company)
EPRI	Electric Power Research Institute
FTE	Full-Time Equivalent
HT	Hypothesis Testing
IE	Impact Element
KM	Knowledge Management
LCG	Linear Congruent Generators
LOPC	Loss of Primary Containment
MCS	Monte Carlo Simulation
MoC	Management of Change
OGI	Oil and Gas Industry
OpCo	Operating Company within the OGI (as opposed to Vendor, EPC, Projects etc.)
P&IDs	Piping and Instrumentation Diagrams
PDF	Probability Distribution Function
PHA	Process Hazard Analysis
RBI	Risk-Based Inspection
RCM	Reliability-Centred Maintenance
RNG	Random Number Generator
SEU	Subjective Expected Utility

SIL	Safety Integrity Level
TAR	Turnaround

Chapter 1

Problem Definition and Rationale

1.1 Introduction

Complex industrial assets rely inherently on hundreds of thousands of data points required to design and run safely, environmentally responsible and profitably. Understanding how to structure, control and distribute this information dates back almost as far as engineering design itself. The arrival of the digital era, with all its advantages, has introduced a risk within one generation: information can be changed, copied and distributed so quickly and cheaply that there is a very real risk that control of it may be lost. This loss of control is not immediately evident in the cut and thrust of daily operations. Maintaining (or regaining) control is invariably an expensive proposition. Justifying the expense is difficult if the benefits are not immediately obvious or calculated using an accepted or defensible method. This research is intended to address this challenge.

1.2 Rationale of the Research

Engineering Information (EI) in the Oil and Gas Industry (OGI) is information about plant design and machinery in the form of data in databases, drawings, documents and numeric or graphical models. EI is generated during the design and construction of plants and is required throughout the plant life for a host of uses such as debottlenecking studies, optimizations, maintenance programme reviews (usually in the form of Reliability Centred Maintenance (RCM)), Risk-Based Inspection (RBI), Safety Integrity Level (SIL) analyses, Process Hazard Analysis (PHA) studies, regulatory reporting, etc). Incomplete or inaccurate EI, which by implication is associated with poor Engineering Information Quality (EIQ), has a negative impact on asset performance and risk profiles. The impact is regarded as significant in industry.

Two general examples may be:

- If a certain compressor is not shown on the asset register, its maintenance may not be scheduled, thereby increasing the risk of expensive downtime.

- If the pressure setting of a relief valve is shown incorrectly on the maintenance procedure, the risk of an explosion is increased.

Problems around EIQ have been implied in several well-publicized industry disasters. The most visible of these is arguably the issue around as-built drawings of the blow-out preventor of BP's Macondo asset in the Gulf of Mexico and the repercussions of that issue during the crucial immediate response period (BOEMRE, 2011).

The existence of this problem has been known for decades. Many attempts have been made to develop standardized data taxonomies and associated data attributes, such as standardized contents of data sheets and associated metadata. Only two examples are cited: EPISTLE (1998), ISO15926 (2003); many more exist.

Although this problem is instinctively understood by engineers and technicians, a review of the literature will indicate that the true impact of poor EIQ is difficult to quantify.

The inability to quantify the impact makes it difficult to justify the expense to rectify the deficiencies in EI. Consequently, the problem is often not rectified. This leads to a perpetuation of the problem and increasing risk, inefficiency and frustration.

Project engineering managers frequently quote the cost required to deliver additional engineering information. Although it is generally well understood that acceptance of inadequate engineering information would imply considerably more cost to update in the Operations phase, a defensible figure for the subsequent work is not immediately to hand, meaning that the decision to absorb the (unknown) additional cost is carried as a matter of course. If a defensible business case could be made, it would enable managers to weigh the benefits of EIQ equally during their decision-making.

Several estimates of the impact of this or related issues are available in the literature. A few studies in other industries have been done that quantify the effect of poor data quality or related issues. These are discussed in Section 2.1. No such calculation has however been found in the literature for the OGI.

1.2.1 Scope of this Study

The scope of this study is limited to investigating the financial impact of EIQ in the OGI and specifically from the perspective of an Operating Company (OpCo). As the thesis title suggests, the intent is to develop a model to measure the financial impact of poor EIQ in an OpCo. To further explain the scope, the component parts of the scope are defined and discussed in more detail in the following paragraphs.

1.2.2 Defining Engineering Information

ISO 14224 (2006) defines equipment data as “technical, operational and environmental parameters characterizing the design and use of an equipment unit”.

Neely et al. (2006) distinguish between configuration data and transaction data. Configuration data is associated with physical and contextual attributes and in that sense, may be deemed “reference” data which does not change frequently. Transaction data is generated during the operation of the asset and therefore changes frequently.

Vayjan et al. (2007) divide data into the categories of master data, transactional data and historical data. They deem master data to be created once, used many times and changed infrequently. The consider master data to contain the “basic characteristics of business entities”.

These definitions are interpreted in practice as that collection of information, whether in the form of data, documents or numerical or graphical models, that serves as reference information to support the daily activities of the asset. This implies that this information is stored in a library or electronic data warehouse, as opposed to the transient information contained in the transactional systems that support daily work. Clearly, this latter collection is the category of information intended for this research.

It is noted in passing that "data" will be used in the singular in this text, as a matter of convention.

1.2.2.1 Defining Engineering Information Quality

EIQ is defined by Haug et al. (2012) as "fitness for use" and by Marsh (2005) as "accurate, consistent, complete, up-to-date and readily accessible". This definition is echoed by Haug et al. (2011).

Klein (2000) refers to seven literature sources before concluding that there is no consensus on the definition of data quality, but that "accuracy, currency and completeness" are significant components of it.

Ballou and Pazer (1985) divide data quality into the dimensions of accuracy, completeness and consistency, of which accuracy is the easiest to evaluate.

Four intrinsic characteristics of data quality are proposed by Wand and Wang (1996). These are completeness, lack of unambiguity, meaningfulness and correctness.

EPISTLE (1999) defines EIQ in considerably more detail. To demonstrate the general intent, this definition is quoted verbatim below, in its entirety:

> *"Properties of information for which quality requirements should be assessed include:*
>
> - *Relevance: the usefulness of the data in the context of the business - does the information need to be retained? What activities does it support?*
> - *Clarity: the availability of a clear and shared definition for the data - do creators and users of information use the same codes and terms with the same meaning?*
> - *Accessibility: where, how and to whom the information is available or not available - is the data easily accessible?*
> - *Compatibility: the compatibility of the same type of data from different sources - if the same type of data comes from different sources, is it created in the same way? Are there multiple copies or versions of this data and if so, is there a master copy from which the others are derived?*
> - *Consistency: the consistency of data from different sources - is the information about objects consistent in terms of naming, values and relationships?*
> - *Completeness: how much of the required information is available - is the entire mandatory information supplied?*
> - *Timeliness: the availability of the data at the time required and how up-to-date that data is - is the data you require available and available when you need it?*
> - *Accuracy: how close to the truth the data is - is the accuracy of the data known and does it meet your requirements?*

- *Cost: the cost incurred in obtaining the data and making it available for use - is the information supplied in a form that means the cost of maintaining it throughout the life of the asset has been minimised?"*

The OGI usually views EIQ in more specific terms such as Data Completeness and Accuracy and takes a risk-based approach not unlike the one suggested in ISO14224(2006). This means that more important information (that is, EI in support of higher-risk equipment or activities) has a higher EIQ requirement than EI of lower importance or risk.

It may be argued that "good" EIQ is a "perfect" data set and that "poor EIQ" is therefore some measurement of the deviation from this perfect world. However, since the definition of "perfect" varies considerably across projects, regions, OpCo's and jurisdictions, this thesis will not attempt to define EIQ in exact terms. Instead, a defensible model to calculate the financial impact of what the sample population perceives as poor EIQ will be pursued. For the purposes of this thesis, therefore, EIQ is intended to mean "EI that is complete and accurate to the specification required by the OpCo and readily available to the target population in the correct repository".

1.2.2.2 Defining Operating Company

An Operating Company (OpCo) is most often deemed distinct from the value chain serving it, the latter consisting of Engineering, Procurement and Construction companies (EPCs), equipment suppliers, service companies, software vendors and consultants. Inside OpCo's there is generally an organisational division between those building new assets ("Projects") and those operating them ("Operations"). The divide often extends beyond organisation to culture, required information and what is deemed good practice. The perspective of the OpCo Operations is specifically selected because most of the cost of poor EIQ is eventually borne in the Operations and Maintenance (O&M) phase of the life of an OGI asset (Gallaher et al., 2004).

1.2.2.3 Defining Model

The title "Model" is selected because the intent of this study is to design, build and populate a model to capture research data in the appropriate form and apply statistical instruments as appropriate to calculate the financial impact. It may be argued that the title "Framework" is more appropriate; however, a review of the definitions[1] suggests that the term "Model" is closer to the intent.

In summary, therefore, the scope of this study is to develop a Model to quantify the financial impact of poor EIQ in an OpCo. This scope is selected for the following reasons:

- Haug et al. (2011) contend that a narrower scope makes it easier to estimate the financial impact of poor EIQ. Since the ultimate impact of poor EIQ is borne by the OpCo, regardless of the source of the EIQ, it makes sense to select that scope.

[1] Merriam-Webster defines "Model" as "a system of postulates, data and inferences presented as a mathematical description of an entity or state of affairs" and "Framework" as "a basic conceptual structure (as of ideas)."

- The OGI is under escalating scrutiny by society with respect to accurate and transparent accountability. (Spangler et al. 2011). This is exacerbated by globalisation and the resultant mergers and universal dependence on IT (Marsh, 2005).

- Technical staff in the Operations function of OpCo's are under increasing pressure due to the predicted imminent reduction of experienced senior technical staff (Cotton et al. (2012)).

- According to Kohli et al. (2011), OGI is deemed a "latecomer industry", meaning that it has been slow to adopt digital technologies and standards.

- An opportunity exists at the time of writing to present a case for investing in EIQ due to period of relatively low profits. This opportunity is described further in Section 1.3.3.

The scope will be explored further in the next section, where those items not in scope are discussed.

1.2.3 Concepts Out of Scope for this Study

There are several closely related subjects that will deliberately not be included in the scope of this study:

- IT strategies, interoperability architecture and related items may arguably have an impact on the effects of poor EIQ. These factors are, however, driven from a very different technical discipline base. EIQ is within the ownership domain of the Engineering disciplines and EIQ is frequently neglected during IT projects. The intent of this study is therefore to study the effects of poor EIQ only.

- This study is not intended to develop bespoke data analysis methodologies or algorithms in service of data analysis. The intent is rather to select an appropriate methodology and apply it to this study of CoPEIQ. The selection process in Chapter 4 utilises the most appropriate, readily available model for this study, based on the criteria identified.

- There is considerable precedent in the literature and many industry-level activities, to develop and integrate detailed data specification and models. Of these, ISO 15926 and its predecessors and variants, the latest of which is CFIHOS (2017), are arguably the most prevalent. This study is therefore not intended to comment on or contribute to that work, other than perhaps serve as a quantified basis for its implementation.

With the rationale for the study and being articulated and its scope being defined some detail, the research design will be discussed.

1.3 Research Problem Statement and Objectives

In this section, the proposed research design will be articulated, from which will follow the first literature study in chapter 2.

1.3.1 Problem Statement

The initial literature study will demonstrate that poor EIQ raises significant risks and unnecessary costs, but that neither a standardized model to quantify the effect, nor a defensible range of impact for any context, are available for the OGI. Indeed, in their seminal work, Haug et al. (2011) declare that, although poor data quality is problematic to many companies and is causing significant costs, "only very few studies demonstrate how to identify, categorize such costs" and "the exact extent of such costs is difficult to estimate". Chapter 2 will demonstrate this in more detail.

This research therefore sets out to develop a defensible model to quantify the financial effects of poor EIQ on an OpCo, deconstructed into the classifications which follow from the research design. The problem statement may therefore be termed as follows: "The financial impact of poor EIQ on operating companies in the oil and gas industry is substantial, but no defensible method exists to quantify that impact. The ability to justify investment in improving EIQ is therefore absent, perpetuating the problem."

1.3.2 Research Questions

Before the Research Questions are stated, the term "Impact Element" (IE) is introduced and defined as "a specific impact or consequential effect of poor EIQ on a part of an OpCo". Examples might be "time spent searching for EI" or "impacts resulting from the use poor EIQ during design".

The following Research Questions are posed for this study:

- Has a method been developed previously to measure the financial impact of poor EIQ?
- What is a sensible classification of Impact Elements that jointly constitute the financial impact of poor EIQ?
- What standardized list of aggregated dimensions is appropriate to report the results (outputs) of IE's against?
- What are the appropriate units of measure for the outputs?
- How should data of this nature be presented to management in the OGI?
- What is the appropriate model to use for analysing this data?
- What are the appropriate statistical instruments for analysing this data?

These Research Questions jointly yield the basis for meeting the Research Objectives articulated in Section 1.3.3.

1.3.3 Research Objective

The Research Questions listed in the previous section culminate in the following objective:

> To develop a standardized Model to quantify the financial impact of EIQ in the OGI, for a specific context.

This objective will serve the OGI by enabling the assessment of the value of an EI remediation effort on an equitable basis to other improvement or expansion activities in an OpCo.

1.3.4 Importance of the Research Problem

The problem of deficient EIQ is regarded as highly significant in the OGI. From an informal discussion with the CEO of the PPDM Organisation in 2016, it was possible to derive impacts in the order of a hundred billion dollars annually across the OGI, using conservative assumptions.

The intrinsic cost benefit is accentuated by four simultaneous influences on the OGI that collectively serve to raise the importance of the present study:

1. There is escalating societal pressure on the OGI to be held accountable for the way in which it manages its information. This is summarized well by Cotton et al. (2012): "In the wake of the 2010 Macondo well blowout … governments require more reporting from sites to assure safety and regulatory compliance. *It is likely this increased attention will bring further pressure on the petroleum industry to standardize data communications protocol".* (italics added). At the risk of singling out one OpCo, a similar report is cited that was published in 2011, in which BP's documentation was scrutinized in considerable detail by a US regulator, found to be deficient and publicly reported as such. BSEE (2011). Since assurance in a capital-intensive industry relies predominantly on an accurate foundation of EI, the OGI will be required to address this foundation urgently.

2. Haug et al. (2012) predict an imminent reduction of experienced senior technical staff. This implies that OpCo's would increasingly need to rely on what may be called as Knowledge Management to control their complex technical processes (Noller et al., 2012). Haug et al. (2012) explains further that increasing remote sites, harsh environments and aging workforce will "raise the sense of urgency in standards adoption".

3. The OGI is deemed a "latecomer industry" by Haug et al. (2012) in terms of addressing efficiencies in the EI domain. The impact of this on an industry basis is hundreds of millions of dollars annually. They quote a study by Kohli and Johnson (2011), which asserts that the OGI has been slow in adopting standards. (It is mentioned in passing that adoption of standards will facilitate an improvement in EIQ.) They discuss possible reasons for this, but for this research it will suffice to state that the OGI has some catching up to do which will add urgency to the recognized need for improvement of EIQ.

4. Since the dramatic collapse in OGI profitability during 2015/16 (World Bank 2015), profits in the OGI are relatively low or non-existent. There is therefore a requirement for an incisive review of internal efficiency. The commercial environment created by low profits presents an opportunity to demonstrate this need, which is often ignored when profits are high. The present study will quantify the effect of an improvement drive for EI, which will help promote the case for funding of the investment required to improve or maintain EIQ.

In summary, there is a considerable financial case for improving EIQ in the OGI, but this can only be accomplished if a robust business case can be made, which is the objective of this study. Given the state of the industry at the time of writing, the management atmosphere is receptive for improvements of this nature.

1.3.5 Use of this Research

Considering the preceding, this study will be immediately useful for the following:

- A defensible estimate of the actual financial impact of poor EIQ for a specific context. This context may be a specific EIQ decision, such as whether to invest in additional EIQ for a certain project, or a macro-decision across an asset or OpCo.
- Measurement of the real effect over time of EIQ remediation, using a longitudinal design.
- Benchmarking of the effect of poor engineering information quality between comparable populations and contexts.
- A possible correlation of EI management maturity and the financial impact of poor EIQ.

In addition, this study may be useful for the following:

- A sufficient number of studies such as this might prompt research into the root causes of the problem for specific contexts, which in turn would point to solutions to this industry-wide challenge.
- This study may form a methodological basis for expanding the scope to adjacent contexts in the OGI, such as EIQ in Major Projects or IT technology improvement drives, and beyond OGI into business cases for funding decisions for EI Standards improvement.

Finally, this study may accelerate similar work in other commodity industries, such as mining.

Summarizing, this thesis will contend to meet the immediate research objectives and is of potential value beyond the immediate context.

1.3.6 Target Audience for this Research

Given the original premise of this study, the primary target audience is senior management of the OpCo population. This is so because senior leaders are the final customers of the entire supply chain of EI where decisions are made regarding funding for EIQ remediation.

There are, however, additional potential benefits from this study, as shown in Section 1.3.4. The following populations may be therefore also be interested in this work:

- Senior EPC Management (which may benefit from this study to evaluate the need for good EIQ development during the project phase)

- EI content and software vendors (which may find these methods of value to demonstrate the contribution of their work to the reduction of the effects of poor EIQ)

- Representative industry bodies (which may use the insights gained from analyses of EIQ to prioritise their activities).

- The academic community, which might find useful precedent in this study for future analyses of a similar nature.

This study is therefore targeted at a specific audience, but is designed to enable expansion to adjacent audiences.

1.3.7 Limitations and Assumptions

There are a number of limitations of this study:

- The selected Research Method (Section 2.3) will show that this study necessarily reduces many specific, integrated individual problems to a generalised model. Its design endeavours to extract the most prevalent phenomena as accurately as possible with reasonable effort, but the results of this method will remain by nature an approximation.

- A further limitation is that every company will at every point in time have a different reality in terms of cultural, technological and process maturity, upon which a unique set of commercial circumstances will be superimposed. This study will attempt to simulate these variables in a reasonable standardized list of inputs, but again will by nature only be and approximation of each reality. (This limitation may in fact be utilised as a benefit for diagnostic insight into these differences, as is explained in Section 7.2.)

- Finally, each of the variables under scrutiny will vary over time for each context. The output of any study using the method developed here will therefore be valid only for the specific context for a finite time. While not deemed a constraint for the immediate need of the work, namely to provide a business case for funding EIQ improvements, this time limitation needs to be kept in mind for general statements within OpCo's and during longitudinal studies.

The following assumptions have been necessary for the execution of this study:

- Survey responses have been assumed to be independent (Chapter 3).
- Certain model inputs have been assumed to be constants, not conditional and normally distributed (Chapter 4).

These assumptions are reviewed and discussed in Chapter 7.

The ethical implications of this research will be discussed next.

1.3.8 Ethical Implications

Bryman et al. (2014) list the principles of ethics as harm to research subjects, informed consent, privacy and deception. Both informed consent and the likelihood of deception are addressed by the Codes of Conduct and by the permission-granting process at the OpCo's where the research is envisaged.

The only possible harm that may come to the subjects of this research could be some form of repercussion for declaring negative consequences of poor EIQ. The likelihood of this is negligible because policies prohibiting such victimization exist in all responsible OpCo's. Statistical validity prefers random selection of subjects and the prevention of response biases (Bryman et al. (2011)) favours anonymity.

The results of the application of this model for a specific context will almost certainly be required to be treated in confidence within the various OpCo's. A sanitized version of the results will be published in the thesis.

1.4 Research Design

Bryman et al. (2014) define research design as "a framework for the collection and analysis of data". Having explored the research problem and defined the attributes of the research, the basis is prepared to discuss and overview of the research design, which follows in this section.

1.4.1 Overview of the Research

Figure 1.1 is a graphical overview of the chapters of this thesis. It will be used as a frame of reference to explain the research design.

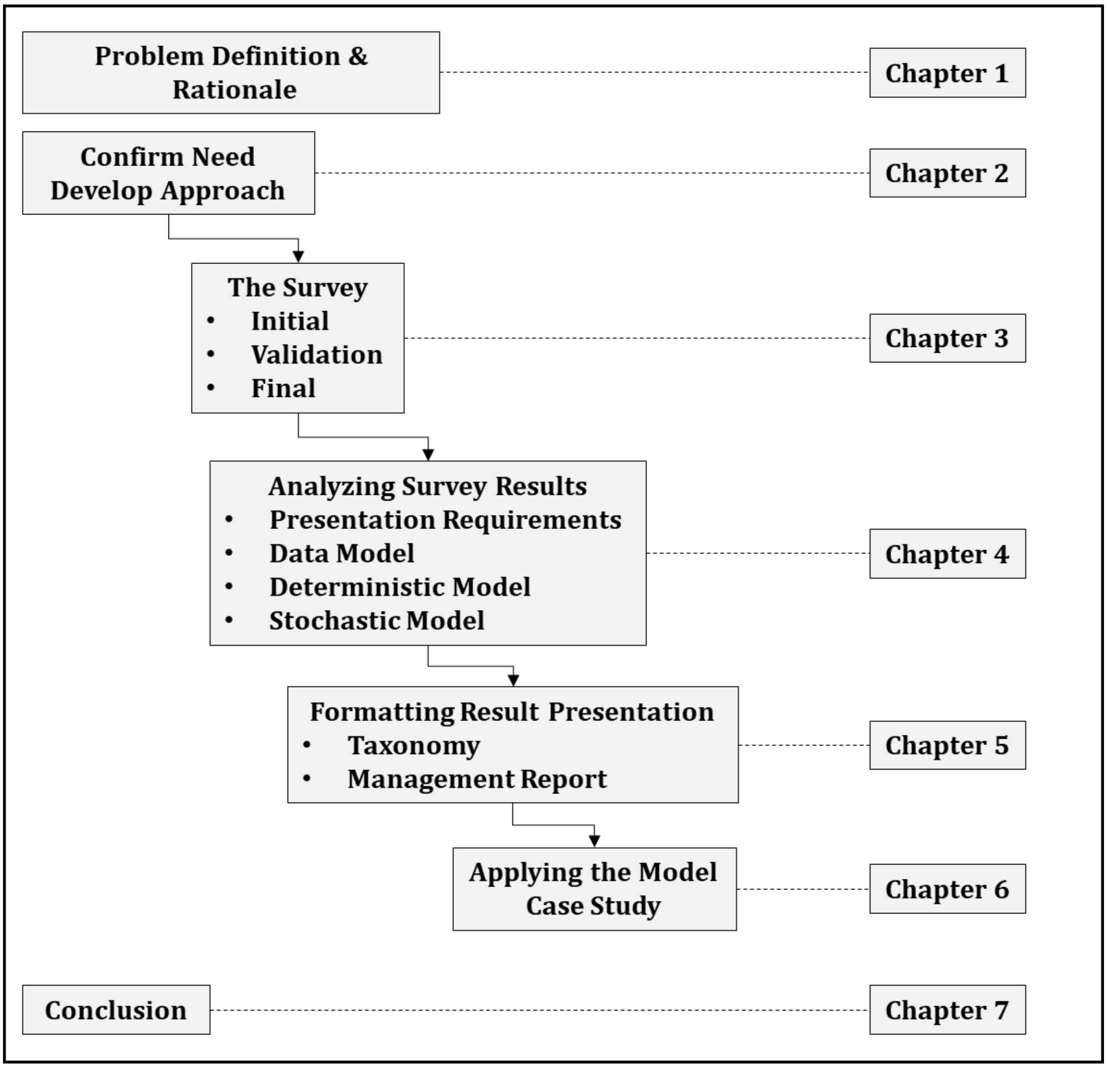


Figure 1.1 - Thesis Overview

1.4.1.1 Chapter 2 - Confirmation of the Need and the Development of the Research Approach

The During this initial phase, three objectives will be met:

- Familiarization with the extent of publications related to this subject
- Confirmation of the necessity to launch this study in the first instance, in other words, confirmation of the need for the work.
- Establishment of a platform of facts to use as basis for positioning the deliverables.

The chapter will start with an initial literature survey, which will analyse 32 sources to determine the extent to which precedent for this study exists. Details are given in Section 2.1, which will show some precedent in adjacent industries, but confirm that no defensible approach exists in the OGI. It will therefore be concluded that this study would indeed fill a need in the literature and meet a need in the OGI for which a solution does not exist.

A number of themes will subsequently be extracted from the literature, followed by a detailed review of three specific publications. From here a fundamental approach will be derived: a cross-sectional design, using a survey method to collect data and reporting results against a standardized taxonomy.

1.4.1.2 Chapter 3 – The Survey

Having, in Chapter 2, confirmed the original contention that there is a need in the literature for this research problem and having derived a fundamental approach to the problem, Chapter 3 will extract from the literature a complete list of the effects of EIQ and transform them into a survey. Gallaherpoor et al. (2004) alert the reader to the need to validate a survey. This initial survey will be validated through a series of structured interviews. The knowledge and experience gained from these structured interviews will be used to develop the final survey.

1.4.1.3 Chapter 4 – Analysing Survey Results

The data collected in the Survey must be calculated into a result that constitutes the financial impact of poor EIQ for the specific context where the Survey data was collected. The fundamental objective of this study is to enable the effects of poor EIQ to be evaluated on an equal footing with other investment opportunities; the results of this study need to be effectively presented to the target audience. Three subsequent steps are required to reach this objective:

- An understanding of the process and dynamics of corporate decision-making. The development journey of this research will therefore pause in the world of behavioural psychology to gain somewhat of an understanding about how OpCo's make decisions and, more specifically, how information in support of decisions should be presented to OpCo's to facilitate effective decisions.

- A data model in a format that is suitable for the survey data to be calculated into a result.

- An appropriate statistical model to calculate the results of the survey into an appropriate result.

From this basis, the stage will be set to present the results.

1.4.1.4 Chapter 5 – Formatting the Results Presentation

The Survey results will have been captured into a data model and then used to calculate a result. The results will need to be presented in a format useful for interpretation and decision-making. This chapter will describe the development of an initial presentation taxonomy, its validation and the development of graphical details.

1.4.1.5 Chapter 6 – Applying the Model – Case Study

With all the preceding development work complete, this chapter will describe the implementation of the model in an operating OGI asset. The chapter will cover the construction and testing of the statistical model, an initial survey by means of a survey instrument and an overview of their results achieved.

1.4.2 Technical Aspects of the Research Design

A research design is defined by Bryman et al. (2011) as "a framework for the collection and analysis of data". The following paragraphs will discuss a few technical aspects of the research design as set out by this definition.

The fundamental design for this study is cross-sectional. This is frequently the design used for social surveys and is the most prevalent design for quantitative business research. The design consists of a data collection on more than one case but at a single point in time. Data is presented in a matrix form of several Observations per Case, of which a simplified form is shown in Table 1.1.

Table 1.1 - Cross-Sectional Design

	Obs 1	**Obs 2**	**Obs 3**	**...**	**Obs n**
Case 1					
Case 2					
Case 3					
...					
Case n					

The data collection is field-based, rather than experimental. There is no attempt to infer or derive causality.

Theory

Since the objective is to quantify the impact of poor EIQ in service of rational decisions regarding EIQ approval or improvement, the output needs to be in a format that is palatable to the management of an oil and gas company. This yields the necessity of a quantitative study [2]. This study interprets a number of theories of human decision-making, as articulated in Section 2.3 and utilises the theoretical basis for the statistical model, but does not profess to develop a new theory or a new application of an existing theory.

At the time of writing, no theory upon which to base this study was evident. The work is therefore inherently deductive. There are inferred dependencies on theories related to decision-making, systems and data management; however, these appear distant from the central research problem.

Epistemology

The initial premise of this study is critical realism, since it recognises that the effects of poor EIQ are 'not spontaneously apparent' (Bryman et al. (2014)). It is envisaged that an open-ended question regarding causality be included in the research questionnaire, thereby introducing a phenomenological bias, but a derivation of specific causality is beyond the scope of this study.

[2] It is noted in passing that the likely causality of this problem is an interesting study into decision theory, behavioural economics, network theory, human irrationality and many more approaches. This research will, however, focus on a quantitative study, for which an appropriate model needs to be found.

Ontology

Given the complexity of the manifestation of the studied problem, an objectivist view is indicated, since standardized organisational process, engineering disciplines and other structural artefacts are assumed and used as the basis of analysis. This is not to say that constructionist views are dismissed; these are merely disregarded during this early investigation into the effects of poor EIQ.

Paradigm

A functionalist paradigm is adopted for this study, since the research objective is fundamentally concerned with solving a business problem.

The approach taken for this study is predominantly quantitative. There is an element of qualitative validation in the early stages of the work, where the structure for the subsequent model is derived from literature and validated during qualitative structured interviews.

Reliability

Stability is not expected since organisations are in continuous flux as market conditions, culture, change management and individual leaders' influence take effect. The results of this study are therefore likely to be valid only for a finite period and only for a specific context. Since the work is primarily intended to support business decision-making, and businesses are made at discrete points in time and based on data from a finite period, this is not seen as a concern.

Of the types of Validity listed, Face, Concurrent, Predictive and Construct Validity are addressed by the two activities listed below:

- A systematic development of a survey from literature.
- A validation of the survey by several structured interviews.

The remaining types of Validity, Convergent and Discriminant, are addressed by the retrospective structured interviews and reviews of concept correlation during the final stage of analysis.

Chapter 2

Confirmation of the Need and Development of the Research Approach

Chapter 1 has introduced the research subject of this study and provided an overview of how the subject will be approached. The approach starts with confirming the assumption that no work of this nature is found in the literature and developing the research approach. That is the subject of this chapter.

Figure 2.1 presents a graphical overview of the process steps of this chapter. It shows the context of Chapter 2 against its immediately preceding and following chapters, together with the detail steps planned within it. This pattern is repeated at the start of each subsequent chapter.

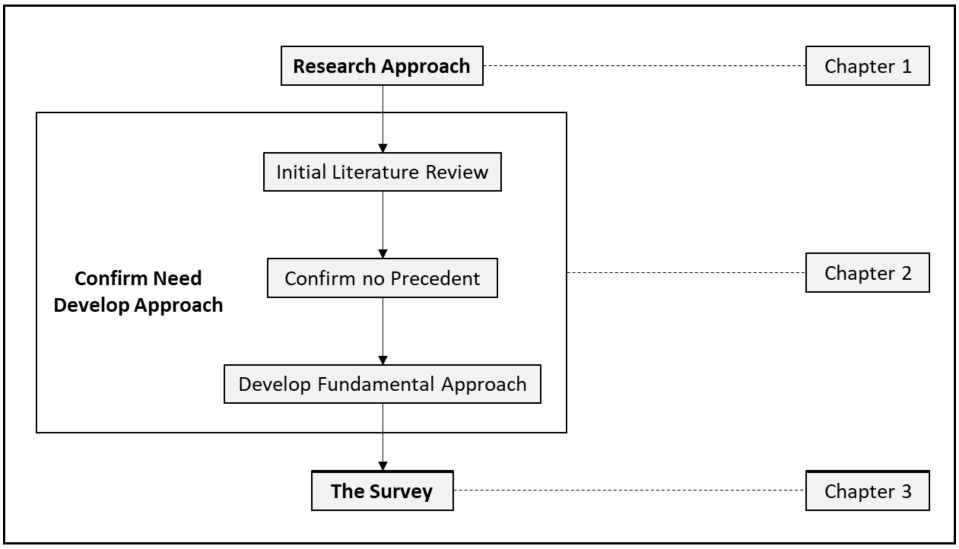


Figure 2.1 - Chapter 2 Detail

2.1 Initial Literature Review

In this first literature review, the objectives are to confirm that there is indeed a need in industry to develop a model quantifying the financial impact of poor EIQ on an OpCo and to gain an understanding of how this problem might be approached.

During this initial phase, three objectives will be met:

- Familiarization with the extent of publications related to this subject
- Confirmation of the necessity to launch this study in the first instance
- Establishing a platform of facts to use as basis for positioning the deliverable stated in Section 1.3.2.

This chapter will analyse 32 sources to meet these objectives. The review will show some precedent in adjacent industries, but confirm that no defensible approach exists in the OGI. It will therefore be concluded that this study would indeed fill a need in the literature and meet a need in the OGI for which a solution does not exist.

A number of themes will subsequently be extracted from the literature, from which a fundamental approach will be derived. Three specific studies will be reviewed in more detail, from which an approach will be derived. This approach will be a survey based on a standardized taxonomy. The collection of responses to the survey will provide the data required for this analysis.

2.1.1 Overview of the Literature

There are many instances in the literature where the negative impact of data quality in general is described. From the date ranges in the references given in this section, it is also clear that this impact has been known for decades.

The exact sources or derivation methods for these figures in the literature are often opaque or anecdotal. Both Eppler & Helfert (2004) and Kim & Choi (2003) note the apparent shortage of data quality studies of scale in the academic literature, whereas Haug et al. (2011) note that industry experts, rather than academics, provide such studies. Despite this opacity, it is evident in the literature that there are many savings to be had and that they are not inconsequential.

The search terms used in the literature review follows a circular path around the central subject. The initial search term “Engineering Information” leads to “Cost Analysis Engineering Data” and onto “Interoperability Cost Analysis Oil & Gas”, “Engineering Data Warehouse Cost Benefit Oil & Gas”, “Interoperability Engineering”, “Terotechnology”, “Configuration Management”, “Information Quality”, “Master Data Quality” and” Engineering Asset Management”. On this journey many peripheral concepts of the central theme have been uncovered, yielding an ever-widening list of impacts of poor EIQ.

The most comprehensive summary of the general impact of poor data quality found has been done by Marsh (2005). He quotes reports by industry experts, including the Gartner Group (2001), PriceWaterHouseCoopers (2002) and Eckerson (2002). The summary of these reports is as follows:

- *75% of organisations have identified costs stemming from dirty data*
- *33% of organisations have delayed or cancelled new IT systems because of poor data*
- *$611bn per year is lost in the US in poorly targeted mailings and staff overhead alone*
- *According to Gartner, bad data is the number one cause of CRM system failure*
- *Business Intelligence projects often fail due to dirty data.*
- *Customer data typically degenerates at 2% per month or 25% annually*
- *Organisations typically overestimate the quality of their data and underestimate the cost of errors*

- *Business processes, customer expectations, source systems and compliance rules are constantly changing. Data quality management systems must reflect this.*
- *Vast amounts of time and money are spent on custom coding and traditional methods – usually firefighting to dampen an immediate crisis rather than dealing with the long-term problem.*

Haug et al. (2011) state in general that data quality can be "crucial to a company's success", or, conversely, can negatively affect the efficiency of an organisation. Elsewhere in the literature, several themes have emerged. These findings, grouped per theme, are discussed in the next sections.

2.1.1.1 Interoperability

Coopers & Lybrand report on the POSC-Caesar project (1997) and state that it could take up to 2000 man-hours to transfer an instrument index from one engineering company to another. This figure is likely to vary according to the scale and metadata mismatch of a specific case, but is nevertheless substantial. Reference is made by Prawel (2003) to studies over a decade of "enormous" mistakes, inefficient time and resultant cost of interoperability problems with CAD. There is reference of broad benefits related to the adoption of communication standards for the OGI, but interestingly, *"no specific cost savings have been presented"* (Cotton et al. (2012)).

Marsh (2005) state that more than 80% of data integration budgets either exceed budget or fail. Haug et al. (2011) identify the risk of increasing data volume and complexity, leading to data silo's, which in turn lead to many different data definitions. This will in turn make interoperability very difficult.

2.1.1.2 Direct Financial Savings

This section lists several cost savings in percentile units for a variety of contexts. Redman (1998) is quoted by Haug et al. (2012) as stating estimates around 10% of cost savings have been concluded by three or more studies that are proprietary (and therefore not accessible for this analysis). Redman is also quoted by Klein (2000), in support of the contention that "errors in data can have a significant financial impact on organisations". It is stated, furthermore, that a service organisation may "informally" consume half or more of its costs due to poor data. Haug et al. (2011) quote the same studies by Redman and add that the studies presented results pointing to estimates around 10% of revenue. Arlbjørn et al., 2007) is also quoted as having shown business performance benefits due to improvements in master data.

Fouhy (1998) states that "an electronic warehouse of design data could save 10% on the life-time costs of a plant". He quotes Howard Masters of BNFL Engineering as saying that "the benefits to plant maintenance would increase plant availability by about 10%".

Ring (1997) presented a paper at the World Petroleum Congress in Beijing in support of the POSC standard. In it he states, in the context of adapting information standards, that "cost savings of 25% are targeted; time savings of 40% have been noted".

Mukhopadhyay et al. (1995) report total benefits of about $100 per vehicle at a Chrysler plant through electronic data exchange.

Coopers & Lybrand (1997) report interviews where potential savings between 15% and 30% on an engineering budget were estimated.

Haug et al. (2011) link poor data quality to increased running costs and lower performance.

2.1.1.3 Confidence in Data

Haug et al. (2012) conclude that fewer than half of companies feel confident in their EIQ and fewer than a quarter trust data delivered to them. Despite the unknown accuracy of these figures, this is a significant factor in the productivity of an OpCo, since the perception of poor EIQ is enough to prompt a considerable verification effort. Marsh (2005) reports that fewer than half of companies feel confident in their data quality and as little as 15% feel confident in data supplied to them. Haug et al. (2011) contend that poor data counteracts the building of trust in data.

2.1.1.4 Time Wasted Searching for Data

Coopers & Lybrand (1997) suggest that more than a quarter of an engineer's time is consumed looking for information. Whether this time spent included confirmation of data is not clear. By extrapolation, a reduction of 10% to 20% of offshore staff is "expected". The report also suggests orders of around 10% less rework or variation orders. This provides a first indication of what Chapter 5 will show to be a significant factor in the intended research. Haug et al. (2011) note that time is wasted detecting and correcting errors in data, and report error rates of .5% to 30% at the field level.

2.1.1.5 Higher-Quality Engineering Analyses

Both API 580 and ISO 14224 support the value of data quality for accuracy of reliability and integrity analysis, which ultimately yields reduced risk and improved uptime for an OpCo. Since the API and ISO will issue documents only after a comprehensive and structured review process, the inclusion of references to the value of good EIQ in these documents confirm that the importance of EIQ is widely recognised in the OGI. An example of a data quality issue would be a case where the material specification of a pressure vessel is recorded incorrectly in the integrity database. The Risk-Based Inspection (RBI) Study could derive an incorrect corrosion rate from this database, leading to excessive maintenance or, far worse, an unexpected Loss of Primary Containment (LOPC) event.

Eppler & Helfert (2004) identify several examples of the financial impact of poor data quality, amongst which are increased assessment costs and process failure costs. Both of these impacts directly support quality engineering analyses.

In extreme examples, the relative value and cost of EIQ is so important that the balance between getting and updating EIQ is formally optimised against the value it brings to the OpCo. For example, Walls (ca. 2003) explains how cognitive science and Bayesian statistics are both used to determine the value of initial and additional exploration data.

2.1.1.6 Human Error Reduction

In his landmark 1990 book "Human Error", Professor James Reason quotes work by J. Rasmussen and sponsored by the Institute of Nuclear Power Operators where the highest category of human performance problems is "deficient procedures or documentation". This observation is one example of a large body of knowledge amongst human factors specialists that recognize the need for ready access to accurate, complete information in a complex, high-risk industrial setting such as an OpCo. Subsequently, Reason (1999) explains that an engineering model of safety performance emphasizes the influence of the informational properties on the performance of front-line operators.

The implication of the themes listed in this section is simply that there is strong evidence of significant financial benefit potential if EIQ is addressed in the OGI. The following section discusses three specific studies that confirm this contention and adds insight into how this subject might be addressed.

2.1.2 Three Influential Studies

During the review of the literature, three specific reports were discovered that provide clear guidance about how the financial impact of poor EIQ in the OGI might be quantified. These three reports are significant in the sense that they share a common architecture, and are deemed relevant since they describe studies in adjacent industries that high capital intensity and reliance on machinery with the OGI. They are reviewed in some detail in this section. Conclusions are drawn from these studies after the overview.

2.1.2.1 NIST Study

The US National Institute of Standards and Technology (NIST) published a "Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry"in 2004. As part of the rationale for the study, some sources of poor interoperability were identified. Amongst these were, curiously, anecdotal reference to a study performed in the late 1980's by an OpCo, which concluded that a consistent data structure would yield a saving of 11 to 14 per cent of Operations and Maintenance costs. The NIST study correctly related these costs to a lack of data standards, which is a requirement for good EIQ.

Their approach was to conduct informal conversations at the outset, to gain an initial overview of the subject. Three categories of cost were defined: avoidance, mitigation and delay. Upon this basis they derived a cost by comparing the current state of their survey subjects to a perfected state by means of a hypothetical counterfactual scenario. Significantly for this study, their focus was "on the changes in business activities and costs associated with data availability — *holding data quality constant*". (italics added). They organised data collection by the following life-cycle phases, identified as appropriate for their industry:

- Planning
- Engineering and Design
- Construction
- Operations and Maintenance

For each of the cost categories a number of standardized cost components were derived. A standardized group of stakeholders were also identified. By means of personal and telephone interviews and an internet survey of more than 100 interviewees across 70 companies, data was collected and organised as shown in Figure 2.2.

This presentation format for a complex result set is potentially attractive for the intended audience of senior OpCo leaders, which is generally accepted to be predominantly graphical thinkers.

One way of describing the general approach of the NIST Study is to say that the study used an architecture of disaggregation of effects, followed by a survey to collect data against the architecture and a subsequent aggregation of results.

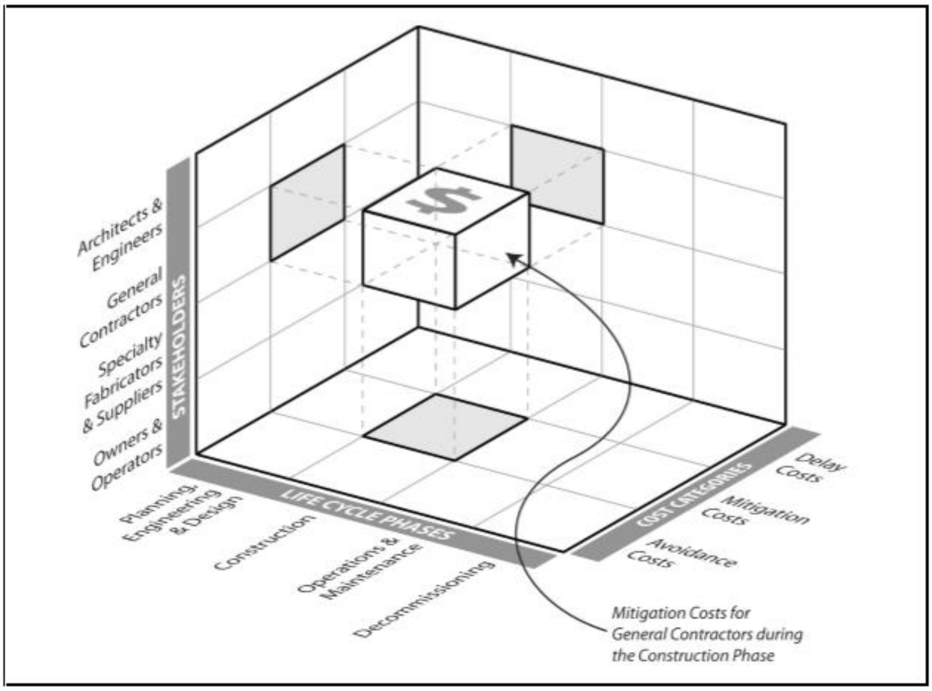


Figure 2.2 - Organisation of the NIST Research – NIST Report (2004)

2.1.2.2 ORCHID Report

The European Committee for Standardization (CEN) published the ORCHID Workshop Agreement CWA 16180-1 in September 2010. In it, they articulated the consensus regarding the “CEN ORCHID Roadmap Standardising Information Across the Plant Engineering Supply Chain - Part 1: Direction and Framework”. From this report a few conceptual framework perspectives are apparent.

Five dimensions of information maturity were defined:

- Business Processes
- Strategic Alignment
- People and Organisation
- Plant Lifecycle Information
- ICT Technology and Infrastructure

Several business processes were identified and classified by the authors where information flow was deemed to be important:

- Processes related to generating EI
- Processes related to classifying EI
- Processes of interchanging EI outside the company

The Orchid Report also, therefore, applied disaggregation of effects.

2.1.2.3 EPRI Study

The Electric Power Research Institute published a report entitled "Data-Centric Configuration Management for Efficiency and Cost Reduction- An Economic Basis for Implementation" in December 2014. The objective of this benchmarking study was to quantify the benefits of such a configuration management system. The approach was to develop an investment model along stochastic grounds that an asset in the process industry could use to determine the benefit of data-centric projects. The project looked at different assets, where the transition occurred at different entry points in the lifecycle. This is particularly advantageous for this present study, since many OpCo's are in very different places in this domain, as stated by Grant (2013). The study identified six different end states as a logical progression along a continuum of a "fully integrated data-centric Configuration Management Information System". These end states are:

- Electronic Document Centralization
- Critical Documents Cross References to Plant Tags
- Data Centralized
- Object-Relationship Model
- 2D/3D Model integration
- 2D/3D Model Analytical Tool Integration

Two classes of savings were defined: "hard savings", which were equal to reduction in time required to perform tasks and "soft savings", which were equal to the number of indirect benefits around efficiency gain. Hard savings were defined according to the following categories:

- Engineering Programmes
- Systems Engineering
- Design Changes
- Engineering Evaluation
- Procurement Engineering
- Work Planning
- Outage Planning

Savings and costs were modelled to be input using a few probability distribution options and a Monte Carlo Simulation (MCS) was used to calculate net present value for the investment decision option.

The EPRI study therefore also demonstrates the architecture of disaggregation/aggregation, this time using a MCS.

2.1.3 Discussion and Conclusions

The following general conclusions may be drawn from this first literature survey.

- The impact of improved EIQ is consistently positive; no cases were found where the opposite was true.

- Impacts are manifested in many forms. This has a direct implication for the methodology required to quantify the effect, i.e. that it should be multi-faceted. Some of these forms or effects may be grouped by business process, asset design phase, engineering discipline, or some similar taxonomy. For this sample of the literature, most of these forms were related to cost savings in general, as opposed to, for example, enhanced morale.

- The central theme from the literature review was, however, that the losses due to poor EIQ and related subjects are widespread and substantial. This study is therefore deemed a significant opportunity to reduce the effects of a problem which is shown to be non-trivial.

This initial literature study demonstrates both the need for and the absence of, a robust method or precedent for quantifying the effects of poor EIQ. Most of the literature reviewed relies on metrics like "% time savings" or "% savings per capital expenditure", quoting "estimated" and "reported" without substantiating the results (for instance Schenk (1985) and CEN (2009)). Neither Schenk nor CEN provide the exact requirements of this study, but they provided useful perspectives and elements in the eventual design of this research.

The three specific studies reviewed in Section 2.1.2 did however employ systematic analysis of comparable problems. The following conclusions have been drawn from these sources:

- Various methods have been applied to address a comparable problem, but they share a common architecture.

- The applied architecture consists of a disaggregation of some structural element like process, activity, maturity state, life cycle stage, stakeholder or similar. Data is then collected at the disaggregated level, often by means of a survey, after which the results are aggregated via an arithmetical or statistical instrument of some form, and presented to the stakeholders.

Many of the sources cited may be criticized as not being able to stand up to academic scrutiny. Indeed, Haug et al. (2012) declared that "in contrast to the sparse information about the costs of poor quality data found in academic journal papers, many industry experts provide such information". Their inclusion is justified on the basis that they provided peripheral context "from the trenches" on what appeared to be an elusive subject in the academic literature. This point is supported by Nickerson et al. (2014), who suggest that "we want to develop useful taxonomies, but not necessarily 'best' or 'correct' ones, as these cannot be defined and, in fact, may be moving targets that could change over time".

In addition to the conclusions drawn in the preceding paragraphs, another perspective emerged from many conversations within the OGI on the subject of calculating CoPEIQ during the course of the research. This perspective is shown in Table 2.1.

Table 2.1- Options for Calculating the Cost of Poor Engineering Information Quality

Option	Title	Cost	Time to Result	Accuracy	Credibility	Comment
A	Detail Analysis	Very High	Very long	High	Medium	High potential for bias
B	Existing Loss Data	Low	Long	Partial	Medium	Data seldom available
C	Survey	Low	Short	Adequate	High	Questionnaire contents important
D	Anecdotal	Low	Short	Low	Medium	Appropriate for certain contexts

A few comments on Table 2.1 are in order:

- By Option A "Detail Analysis" is meant a detailed work study across the various organisational entities and work processes within the scope of a particular study. Apart from the huge cost and time for such a study, it introduces the possibility of the Hawthorne effect and bias by analysts.
- Option B "Existing Loss Data" is to codify losses related to CoPEIQ into the existing loss management system, and require these losses to be logged as part of the loss management process. Not many assets in the OGI have such data available. Even where the data is available, only a partial result will be achieved and a lot of data will be needed to make an accurate calculation.
- Option C "Survey" has many advantages, but does demand a well-designed survey questionnaire to capture the salient data points.
- Option D "Anecdotal" is appropriate within a corporate culture where "war stories" carry more weight than rational fact. This subject is discussed at length in Section 4.1.

The fundamental research method was derived from a combination of the common architecture derived from the three influential studies and Option C in Table 2.1. It is described in Section 2.3.

Before concluding this section on the literature review, the subject of literature review per se is briefly discussed, using Petticrew and Roberts (2006) as a reference. They list the following types of literature reviews:

- Systematic
- Narrative
- Conceptual
- Traditional
- Critical
- State of the Art

The authors provide a definition for each item in the above list.

Based on the definitions provided by Petticrew and Roberts (2006), this study is interpreted as being a "Critical Review". This is defined as a "term sometimes used to describe a literature review that assesses a theory or hypothesis by critically examining the methods and results of the primary studies, often with a wealth of background and contextual material, though not using the formalised approach of a systematic review".

2.2 No Precedent for This Study

The literature review in Section 2.1 demonstrated that precedent for this type of study has been found in the US construction and nuclear power industries. That the OGI recognizes EIQ as a problem is evident by the many references to the problem and its effects and the significant investment already done in standardizing approaches, e.g. EPISTLE and CFIHOS. However, no standardised method was found specific to the OGI.

2.3 The Fundamental Research Method

This section describes the Research Method adopted after the analysis described in the preceding section of this chapter was performed. The description of the Method is presented here at a conceptual level; more details are provided in Chapter 3.

Bryman et al. (2011) contrast the terms "Research Design" and "Research Method" as follows: Research Design provides a "framework for the collection and analysis of data", whereas Research Method is a "technique for collecting data". The former of these was discussed in some detail in Section 1.4.2; the latter is discussed in this section. The conclusions drawn in Section 2.1.2 are interpreted in this section to develop the Research Method. The approach has been summarized previously in Section 1.4.1.1; in this section detail is added.

Table 2.1 suggests that a survey method is low-cost, quick, reasonably accurate and credible. Section 2.1.2 demonstrated how three similar studies had been conducted. The fact that the three so-called "influential studies" used methods of collecting perception, such as surveys, further supports the suggestion that a survey method is appropriate for this study. Since these three studies jointly constitute models for the research discussed in this thesis, some common characteristics are derived to serve as a precedent for this study. These characteristics collectively yield the general approach adopted in this study. This is shown graphically in Figure 2.3 and subsequently discussed.

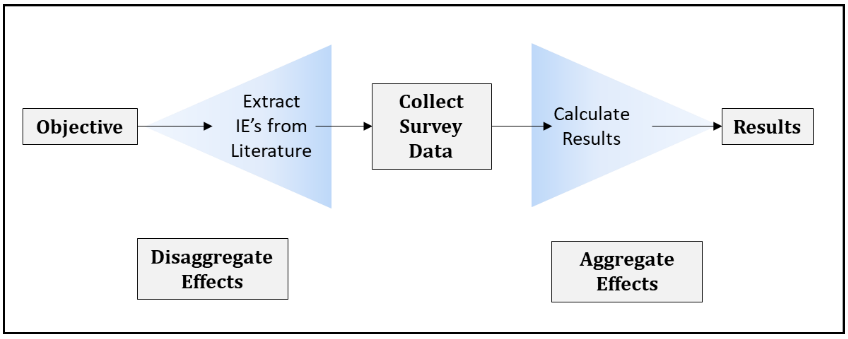


Figure 2.3 - Fundamental Research Method

According to this logic, the first activity of the Research Method is to acquire the individual effects of poor EIQ on an OpCo. For this purpose, the concept of an IE was introduced.

The literature review described in this chapter provide a comprehensive and available source of IE's. This is the same approach taken by Eppler & Helfert (2004).

The NIST report shows the necessity of validating a structure of this nature. (Gallaher et al., 2004). The next step is therefore to collect initial survey data by means of structured interviews. There are two objectives with these interviews, namely to validate the survey questions for completeness and mutual exclusivity and to collect a first baseline result, which in turn will assist in validating the fundamental concept. The baseline result was an indicative order of magnitude (or "baseline") for each of the IE's for the specific context in which the structured interviews were conducted. This baseline is useful for scaling issues and provided a first confirmation of how a survey might be received in the OGI. Its primary objective, however, is a *quid pro quo* in the form of immediate feedback to the OpCo where the structured interviews are conducted.

Having validated the IE's during the structured interviews, an updated list of IE's is transformed into a final Survey for which the data collection method is to be an internet-based self-completion survey. This data collection method is selected for its practicality across continents and time zones, and for its prevalence and familiarity in the OGI. Pending appropriate permissions, the primary elements of a large-scale survey is prepared.

The development of the final Survey is discussed in detail in Chapter 3.

It was hence possible to start the design of a statistical model that would aggregate the results of the Survey in a deterministic manner. An understanding of the decision-making dynamics of OpCo's is required and specifically the requirements of information presentation for this context. Once the deterministic algorithms are specified, it is possible to develop an appropriate stochastic model. These steps are discussed in Chapter 4.

The Research Method, in summary, is therefore to disaggregate the impact of poor EIQ into granular elements, validate the list and its presentation by several structured interviews, followed by data collection at scale by means of an internet-based survey. The method is concluded by developing a model to aggregate the results (Chapter 4), an analysis of results presentation (Chapter 5) and the application of the entire research design in an actual setting (Chapter 6).

Chapter 3

The Survey

Chapter 2, having described the Research Method at a superficial level, has set the stage for the development of the Survey in detail. That is the objective of this chapter.

Figure 3.1 provides an overview of the development of the Survey; descriptions follow subsequently. As was the case for Figure 2.1, the format is to refer to previous and subsequent chapters and show the details within this chapter. This is done to assist the reader in following the logic of the thesis.

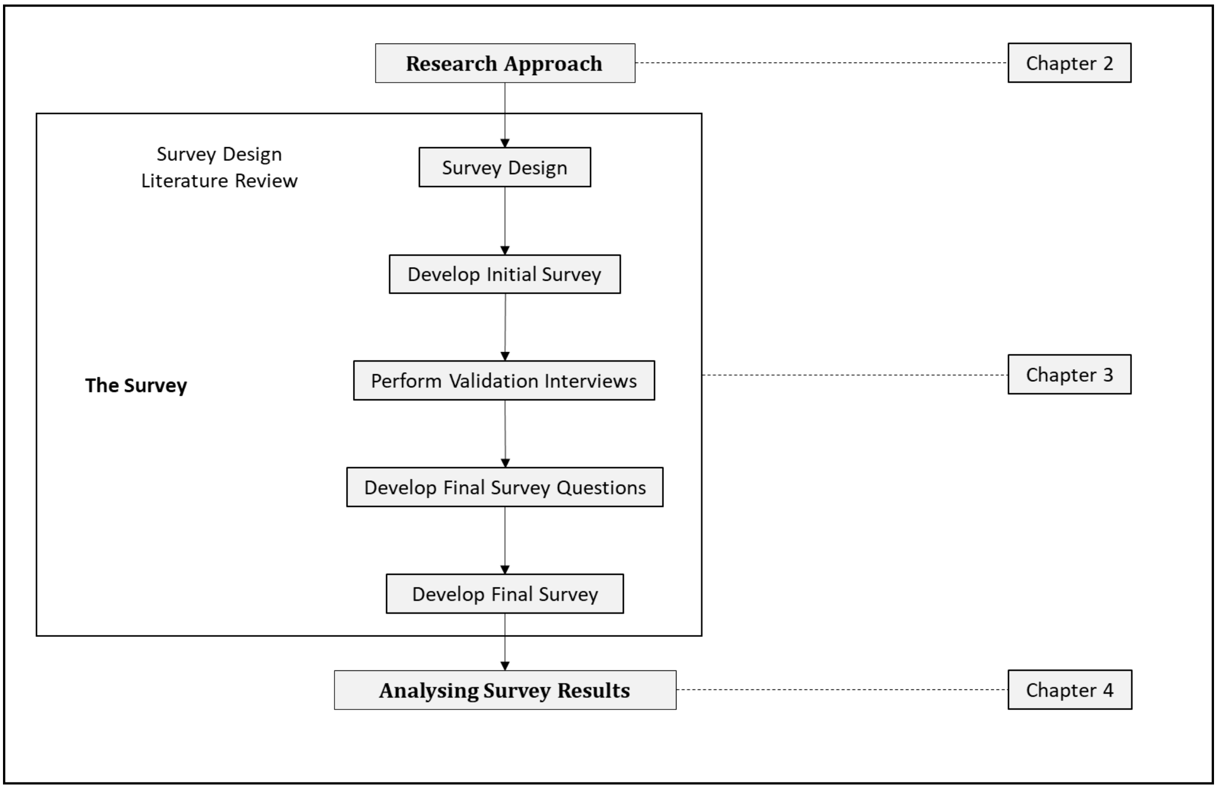


Figure 3.1 - Survey Detail

3.1 Literature Review: Survey Design

In this section, an overview of the literature is provided, which then forms the basis for the fundamental design aspects of both the initial and final surveys. After a summary of the reviewed literature, the selected design inputs are summarized.

Gackowski (2009) states rather directly that "Reluctant, disinterested respondents rarely challenge questions", before distinguishing between necessary and nice-to-have qualities. He calls for "organised, focused and succinct" questions, stated within an appropriate context. He then proposes that the order of questions should correlate inversely with their relative importance.

Survey responses have been known to be affected by questionnaire design (Peytchev et al. 2006). Design issues may be 1) the use of pages or scrolling, 2) the use of automated skip logic or 3) the use of hyperlinks. Reminder emails may increase response rates. Scrolling reveals the length of the survey but may increase the completion time due to the need for navigation. The study however showed no significant difference between scrolling or paging, nor between the use of mandating responses or not.

In their study of satisficing and how to reduce it, Hamby et al. (2016), defined satisficing after Simon's (1956) satisficing theory, as "the tendency to seek quick, 'good enough' answers" instead of taking the time and cognitive effort to produce an optimal survey response. They assert that survey design has a "huge effect" on satisficing behaviour. Quoting Scharz & Strack (1985) and Tourengeau & Rasinskki (1988), they explain four steps required to develop a survey response:

- Understand and interpret the question
- Search memory for relevance
- Integrate the information to form a decision
- Determine where that decisions fits into the range of possible answers.

With this insight, it is easy to understand how likely respondents are to satisfice and how this relates to cognitive effort, survey length, complexity and presentation, including linguistic characteristics (Krosnich et al., 1996). They also note how people become tired towards the end of a survey. They note that five or seven-category scales are most frequently used.

Axxin et al. (2011) support Gackowski's (2009) notion regarding reluctance towards surveys, and then proceed to explore the notion of "responsive survey design" together with the effects of pre-notification, incentives and alternative modes of data collection.

Lauer et al. (2013) propose a "Janus-Faced Approach" to survey design, where the experience of the respondent and the requirements of the researcher are addressed simultaneously.

Lauer et al. (2013) also mention "survey fatigue" and suggest ways to combat it:

- They quote survey research by Galesic & Bonjal (2009) that suggest that survey completion rates drop off rapidly after 20 minutes
- They include what they call "the fine print" (ethical information) in smaller font.
- They include a progress indicator that provides feedback to the respondent about how far along respondents are in the survey.
- They remind researchers that the order of the questions need not align with the data models or analysis sequence.
- They feel that SurveyMonkey and similar platforms limit researchers to a simplified survey.

Jooste (2014) quotes the web survey implementation framework proposed by Belfo & Sousa (2011). It is included in Figure 3.2.

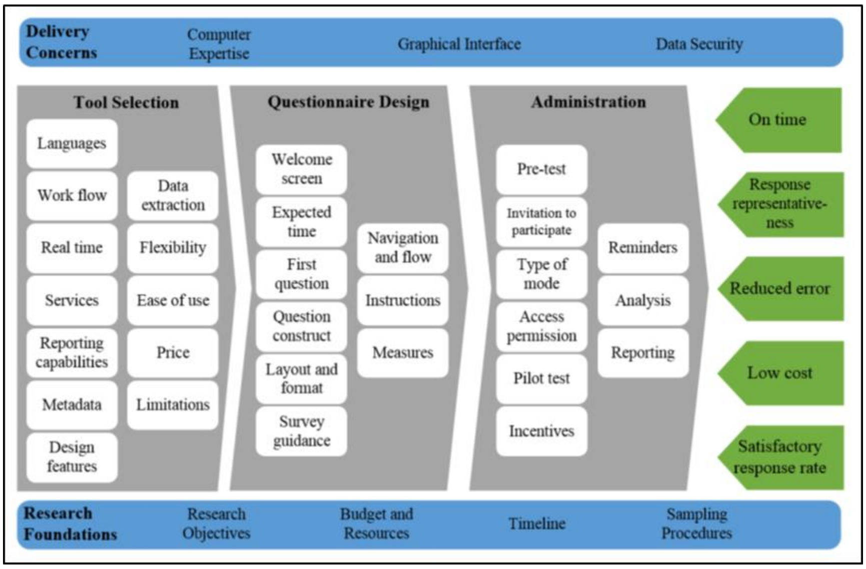


Figure 3.2 - Web Survey Implementation Framework (adopted from Belfo & Sousa (2011))

The framework proposed by Belfo & Sousa (2011)) is considered very comprehensive and includes many of the aspects covered in other parts of the literature survey in a structured and holistic framework.

This brief review of the literature provides the following conclusions with respect to the Survey Design:

- The fundamental objective of the survey design is to present the question in the most palatable manner possible for the respondents. Put another way, the survey design needs to be a balance between maximizing the probability of good respondent data quality and achieving the scope and granularity of data required for good analysis.

- The survey design needs to minimise satisficing and fatigue and use a format and word choice familiar to the respondent to minimise cognitive effort.

- The implication is that the survey needs to be as short as possible, contain focused and clear questions, require no more than 20 minutes to complete, and be presented in a way that minimises cognitive effort.

- The framework by Belfo & Sousa (2011)) is used to inform both general and specific survey design questions. Fundamental survey design questions are listed below. The same framework is referred to later, in Sections 3.2 and 3.3, to make specific design decisions.

Based on the conclusions, the following fundamental survey design decisions were made for this study:

1. Since English is almost universally spoken in the OGI, it was selected as a default language. Future translations are obviously not excluded for specific contexts.

2. The price of the selected survey tool should ideally be zero or minimal.

3. Expected survey time is limited to 15-20 minutes.

4. The construct of the questions is intended to be succinct and to the point, using specific phrases consistently that were defined in the survey introduction.

5. Incentives were never considered, since their efficacy was instinctively accepted to be zero (which was later confirmed by Hamby et al. (2016)) In addition, incentives may raise several ethical objections within the constraints of an OGI. Instead, since participation in the survey indirectly contributes to the performance of an OGI asset, it is considered part of a Continuous Improvement drive and therefore deemed part of the contribution of OpCo staff members of to the performance of an OpCo. In the Final Survey, this point was made by the introductory email explained in Section 3.6.

6. To compensate for survey fatigue, questions are ordered in their perceived order of importance. That order was significantly modified in the Final Survey.

7. Linguistics. Since OpCo's frequently use specific phrases for certain concepts, there is a real risk of using vernacular that is only understood in one Opco or even one asset. To standardize the survey questions to the greatest extent possible, some effort was made to use naturalized and generally accepted language in the OGI.

Based on these fundamental survey design principles, the Initial and Final Surveys are described in the following sections.

3.2 Initial Survey

This section describes the development of the Initial Survey. Several general design observations are first described, with reference to the framework in Figure 3.2, after which the process is described in some detail.

The following general design considerations are relevant:

- Ease of use is not an issue, since the interviewer can explain and clarify each question and personally capture the data.

- Survey guidance is provided via the Survey shown in Table 3.2, which is also projected onto a screen during the structured interview meetings.

- Ranges and end rates are not specified in the initial survey. In fact, one objective of the initial survey is to gain an understanding of the orders of magnitude being claimed by respondents.

- Units of measure are likewise not specified. A wide range of units of measure were reported during the structured interviews for the time elements of the questionnaire. From this list of responses two options were selected for the final survey. This is shown in Section 3.3.3.

3.2.1 The Structure of the Initial Survey

The Initial Survey was done in a structured interview format at an operating plant of a major OpCo. This meant that the objectives, ethical constraints and definitions of terms could be explained by the interviewer. Further, the reference data required for the calculation of CoPEIQ, (which is called Asset Reference Data in Section 3.4) was known to the interviewer, thereby obviating the need to formally record it. Consequently, the Initial Survey could be simpler and less formalized than the Final Survey.

The Initial Survey was structured in five parts as shown in Figure 3.3.

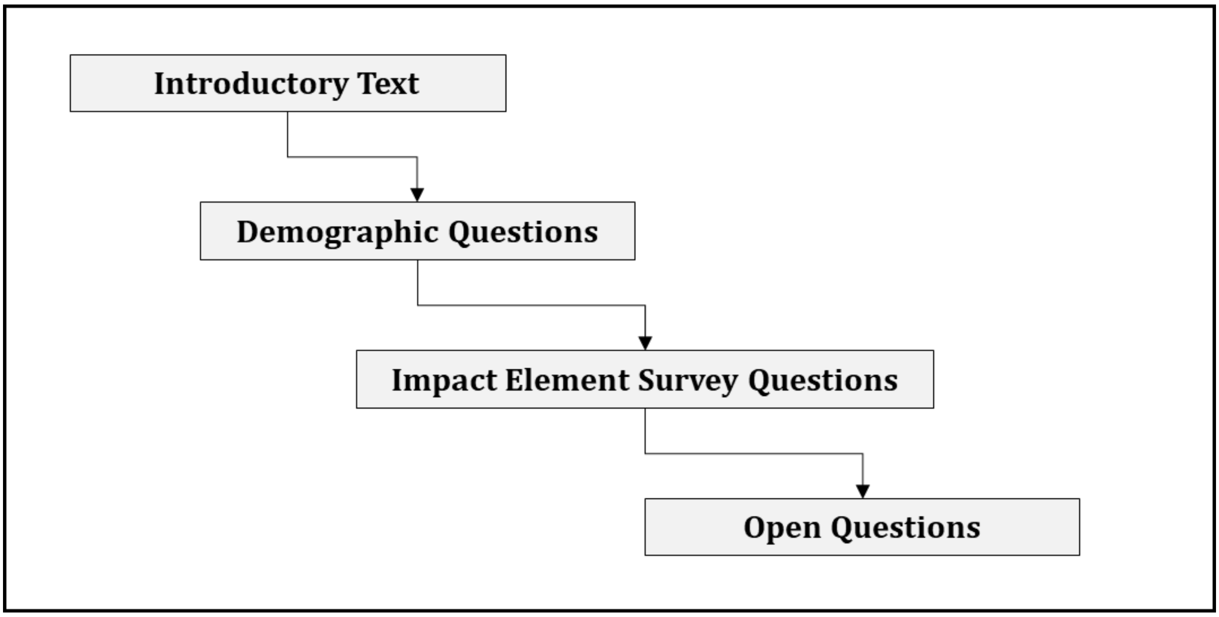


Figure 3.3 - Structure of the Initial Survey

The following explanations are relevant for Figure 3.3:

- Introductory Text was developed to provide a consistent verbal introduction to the Survey. It includes the objective, definition of terms and assurance of confidentiality.
- It has been demonstrated in Chapter 1 that this research aims to fill a real need in the OGI. The bias of the study is therefore practical, as opposed to academic. Academic prudence, however, demands that the effort expended to collect data should be used to the maximum practical extent. Accordingly, data collected in the Demographic Questions are not used in this study; instead the intent is to collect the data for secondary research at a future date. This is explored further in Chapter 7.
- IE Survey Questions captures the direct CoPEIQ was captured. The development of the questions is described in Section 3.3.2.
- Free-Form Text Questions have two objectives:
    - Questionnaire Validation – here the intent is to learn from the respondents about the structure, contents and length of the survey for future iterations, as well how respondents generally perceive the Survey.

- Open Questions test for the perceived cause and remedy of the existence of CoPEIQ and are intended to provide a database of future thematic research into the phenomenon, subject of course to a sample of valid size and variability.

The points above constitute the design of the initial survey. After validation of the survey, a few modifiers were applied, based on several learnings that are summarized in the next section.

3.2.2 Development of the Initial Survey Questions

3.2.2.1 Extract Potential Impact Elements from Literature

As in the case of Eppler & Helfert (2004), the literature review provided a comprehensive and readily available source of IE's. During the initial literature review, 32 sources were consulted, from which 196 potential IE's were extracted. The extraction of IE's from the literature was done with reference to the definition of an IE given in section 1.3.2. The subjects of the literature covered a range of themes explained in Section 2.1, but the text provided several clues of IE's. Two examples are:

- "instruments for decision support" (Haider (2008))

- "forced to change from one vendor's equipment to another "(Cotton et al (2012))

In the latter example, the IE was later classified to "IT" as shown in Figure 3.4 and therefore not included in the survey.

The complete list of extracted IE's is included in Appendix A. They are of a variety of levels, types and classes and the next challenge was to arrange them into a practical list of questions, suitable for the purposes of this thesis.

3.2.2.2 Derive Initial Impact Elements List

With the initial list of IE's extracted, the process of deriving the IE List can start. This list eventually evolves to the Initial Survey. The process is described below.

Classification

The extracted list of IE's are classified according to a few dimensions. The dimensions are developed by engineering judgement based on domain knowledge and are intended to provide a starting point for the survey iterations. The initial dimensions used are as follows:

- Key business function accountability
- Cost category
- Tangibility or intangibility
- Asset lifecycle phase
- Engineering discipline

For this analysis, it is useful to keep in mind how Cost categories were defined by NIST (2004):

- "Avoidance costs are related to the ex-ante (based on forecasts rather than actual results) activities stakeholders undertake to prevent or minimise the impact of … problems before they occur.
- Mitigation Costs stem from ex-post (based on actual results rather than forecasts) activities responding to interoperability problems. Most mitigation costs result from electronic or paper files that had to be re-entered manually into multiple systems and from searching paper archives.
- Delay costs arise from interoperability problems during completion of a project or the length of time a facility is not in normal operation".

Filter for Scope

The IE's are subsequently reviewed for scope, based on the scope defined in Section 1.2.1. For this step the categorizations of "Key Business Function Accountability" and "Asset Lifecycle Phase" are useful. The following interpretation was applied: since the research subject is "Engineering Information", all EI related to the following broad organisational constructs is in scope: Engineering, Planning, and Maintenance personnel in the Operate phase. The classification "key business accountability" is subsequently used to exclude the following accountability groups: Projects, Supply Chain Management, Finance, HSE and IT. The following accountability groups remain in scope: Operations and Maintenance, Planning of any form in Operations, Process Safety and Engineering support in the Operations phase.

The remaining IE's deemed out of scope are simply removed from the list.

A summary of the functional scope distinction is shown in Table 3.1.

Table 3.1 - Functional Scope Division of this study

Functions in Scope for this Thesis	Functions out of Scope for this Thesis
Engineering (in the Operate Phase) Corporate Risk Process Safety Planning Maintenance	Capital Projects Finance Supply Chain Management Human Resources Health, Safety and Environment Operations Information Management and Technology

The NIST reporting structured shown in Figure 2.2 shows that there are additional benefits for peripheral stakeholders such as equipment suppliers, engineering contractors and solution providers. As a result of this step, these benefits are not included in this study and are therefore deemed potentially additional benefit.

After reducing the IE's to this Scope, 185 IE's remained from the original 196.

Remove Intangibles

Those IEs with the attribute Intangible are removed next. This is done because the benefits of good EIQ on these IE's would be impractical to determine, or be it that they are instinctively significant. Many of these are in fact able to be measured, for example human error rate, however for the purposes of this study, the measurement of these Intangibles is deemed too specialized and complex.

The Intangible Benefits removed from the list were:

- Improved standardisation
- Improved ability to integrate all workflows
- Improved concurrent engineering
- Improved decision quality
- Improved ability to prioritize work
- Improved Situational Awareness
- Improved working relationships
- Improved collaboration
- Reduced human error rate
- Improved transfer of tacit to explicit knowledge
- Improved efficiency culture
- Improved utilisation of intellectual capital
- Reduced likelihood to divert from standard procedures
- Improved relationships with external stakeholders

This left 171 remaining IE's. These intangibles may be presented as additional benefits to any cost-benefit report using the process described here.

Remove Duplicates

Several duplications are subsequently removed.

One example is "Reduction in headcount due to efficiency" is represented by "Work Process Efficiency". At this stage, the remaining list of IEs' to be included totals 115.

Final Check

The final step in the rationalization of the initial survey is to combine the remaining IE's until the ending condition "every condition is unique and not repeated" (Nickerson et al. (2014)) is met. An example of a combination is that the empirical element "Detailed work processes are hindered by lack of definition" is met by both "Lack of common interpretation" and "Additional time spent reviewing EI Standards"; they are therefore combined.

Of the original 196 extracted IE's, 51 are included in the Initial Survey.

3.2.2.3 Convert to Survey Questions

The preceding sections yielded a list of IE's ready for inclusion in the Initial Survey. Two steps remained to prepare the Initial Survey:

Sort by Respondent Group

The IE list was sorted into the following sequence to enable the construction of a filtered questionnaire:

- Questions relevant to all subjects

- Questions relevant to subjects working in Engineering
- Questions relevant to subjects working in Operations and Maintenance
- Questions relevant to subjects working in a corporate function

Convert text to Question

The final step for the Initial Questionnaire was to develop each IE into a sentence suitable for the intended audience. Two examples are given, for the IE's "Potential production loss" and "Regulatory Response".

The following questions were developed for these IE's respectively:

- Estimate the potential loss of production
- Estimate the additional time spent responding to regulatory query

These steps concluded the development of the Initial Survey question list.

3.2.3 The Initial Survey

This section describes the initial survey in some detail, based on the development of the survey questions described in Section 3.2. Before proceeding, however, a number of concepts are introduced which will assist in explaining some of the questions and variables being used in the subsequent narrative.

- Barrel Oil Equivalent (BOE) is widely used in the OGI as a generalized term, based on energy content, to report production of various fractions of oil and gas in the various processes. A BOE is based on a barrel of oil, which in turn is 42 US gallons of oil. This size of barrel was decided upon in 1872 as being "about as much as a man could reasonably wrestle". (aoghs.org, accessed May 5, 2017). The US IRS (accessed April 19, 2017): Section 29(d)(5) and (6) states that the term "barrel-of-oil equivalent" (BOE) with respect to any fuel generally means that amount of the fuel which has a BTU content of 5.8 million. (Note that Imperial units are being used. This is because the majority of the OGI conventionally uses Imperial units, despite the drive towards SI Units internationally.)

- Full Time Equivalent (FTE) is defined as 'the hours worked by one employee on a full-time basis'. The concept is used to convert the hours worked by several part-time employees into the hours worked by full-time employees. (www.accountingtools.com accessed May 27, 2017). FTE is frequently used in the OGI as a basis of calculating workload or adjusting plans to level resource requirement. It is noted in passing that in the OGI, FTE is frequently used regardless of whether the human resource pool are employees or contractors.

- Asset Rated Production (ARP) is colloquially known as "nameplate" production rate. ARP is, for this study, defined as the rate used in the business plan for year in which the survey is undertaken. The term "nameplate" is more accurately described as the rated production of the asset in its original or modified design basis.

- Plant Availability (Aplt): This measure is simply intended to calculate the time during which an operating asset is not producing, expressed as a percentage of total calendar time.

- Process Safety Incident (PSI): Process Safety is defined by the American Institute of Chemical Engineers as “a disciplined framework for managing the integrity of operating systems and processes handling hazardous substances by applying good design principles, engineering and operating practices. It deals with the prevention and control of incidents that have the potential to release hazardous materials or energy. Such incidents can cause toxic effects, fire, or explosion and could ultimately result in serious injuries, property damage, lost production and environmental impact”. (www.aiche.org/ccps accessed May 27, 2017). Accordingly, the API defines a Process Safety Incident as “an unplanned or uncontrolled Loss of Primary Containment (LOPC) of any material including non-toxic and non-flammable materials (e.g. steam, hot condensate, nitrogen, compressed CO_2 or compressed air) from a process, or an undesired event or condition that, under slightly different circumstances, could have resulted in a LOPC of a material”. (www. api.org accessed May 27, 2017). The Cost of a Process Safety Incident (PSI) includes direct costs like “cost of repairs or replacement, clean-up, material disposal, environmental remediation and emergency response”.

Figure 3.4 shows the wider context of the cost of poor EIQ. This figure relates to the scope discussion in Section 1.2.1. It is clear that there are many benefits of good EIQ to a wider stakeholder group, but the focus of this study is on the Operations phase of a particular asset. It is a graphical representation of the evolution of the survey and shows how the additional benefits of EIQ in Intangibles and for Capital Projects, third parties, IT, and ultimately the Asset. The dotted line box, containing “Asset”, represents the scope of this study.

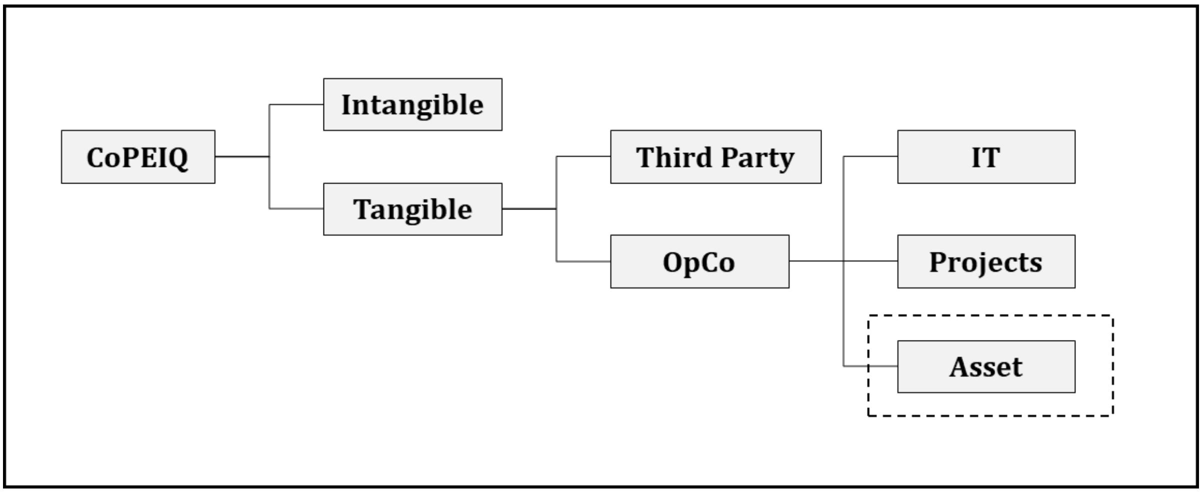


Figure 3.4 - The Context of the Survey

Within this context, the outcome of the preceding sections is recorded in this section, where the contents of the Initial Survey are reported as follows:

- Table 3.2 shows the introductory text to the Initial Survey.
- Table 3.3 shown the demographic questionnaire in the Initial Survey.
- Table 3.4 shows the IE Questionnaire of the Initial Survey.
- Table 3.5 shows the free-form text questions and closing statement of the Initial Survey.

Table 3.2 - Introductory Text to the Initial Survey

Thanks for making the time to discuss this subject. Your experience is invaluable in this research. This research is intended to develop a method to estimate the effect of poor Engineering Information Quality. This will in turn enable improvement initiatives in this domain to compete with other improvement projects for funding.
The questionnaire has two objectives:
- Collect data.
- Confirm that the contents of the questionnaire is complete and relevant.

You will come across the phrase Engineering Information Quality (EIQ) in the questionnaire. EIQ is defined as documents, drawings, data in models and databases that are related to equipment design, operating context, maintenance, performance and condition, and that is complete, accurate, in the right system, coded correctly and readily available to a user.

You will also come across the phrase “Additional Time” in the questionnaire. Please read this phrase as “the unnecessary extra time needed for the specific activity due to poor EIQ only”. This does not imply that the activity will take zero time in the ideal state, and is not intended to capture unnecessary time due to other factors, such as poor access to the plant or non-compliance to work processes, etc.

For the data to be analysed, the questions are presented in a structured format. We request that you follow along in this structure. At the end of the question set there is opportunity for free-form text comment.

Your contribution is confidential. Demographic data is collected for analysis only.

The questionnaire is divided into four parts:
-Demographics data. This data is captured for future analysis of the causes from a cultural and demographic perspective.
- The questionnaire. Please note that there are ‘filter’ sections so that you only answer the sections relevant to you
- Comments on the contents and format of the questionnaire itself.
- Open-ended questions where your general comments are requested.

Please answer frankly, using your best judgement. The objective is not to be exactly accurate, but rather to provide a reasonable and realistic estimate. If you cannot easily estimate a response, please ignore it and move onto the next one.

Table 3.3 - Demographic Questionnaire in the Initial Survey

A reminder that this data is captured for analysis, and will be kept anonymous and not be used for any other reason.	
Please complete the following data	
Your Asset	
Date	
Your Discipline	
What is your highest academic qualification?	High School
	Diploma
	Bachelor’s Degree

	Master's Degree
	MBA
	PhD
How many years of professional experience do you have?	0-5 years
	5-10 years
	10-20 years
	20-30 years
	30+ years
Industry in which most of your professional time has been spent?	O&G Upstream
	O&G Downstream
	O&G LNG
	O&G Pipelines
	Mining
	Oilsands
	Other
Current role	Technical specialist
	Supervisor
	Manager
	Trainer/coach
	Other

Table 3.4 - IE Questionnaire of the Initial Survey.

Please answer the following questions:
Please estimate the additional time preparing for PMs
Please estimate the additional time spent to prepare/message corporate KPIs
Please estimate the additional time spent preparing Regulatory Report
Please estimate the annual production loss due to inaccurate engineering analyses
Please estimate the additional time spent optimizing resource
Please estimate the additional time spent responding to regulatory query
Please estimate the additional time required to support Decision Quality
How much longer does it take to respond to significant change in circumstance
Regulatory Penalty: Estimate the Incremental Likelihood times the Likely Cost
Please estimate idle resource time
Please estimate the time spent searching paper archives
Please estimate the additional time spent searching electronic archives
Please estimate the additional time spent verifying EI is correct
Please estimate the additional time spent resolving interpretation differences
Please estimate the additional time spent updating EI
Please estimate the additional time spent transferring EI
Please estimate the additional time spent reworking EI
Please estimate the additional time spent creating rogue databases
Please estimate the additional time spent revising EI
Please estimate the additional time spent reviewing EI Standards
Please estimate the time spent clarifying misunderstanding
Please estimate the Effect of misunderstanding
Please refer to one event that you know of, and estimate the cost where the company's knowledge was not fully or properly used
If you are working in Engineering, please answer the following additional questions:
Please estimate the additional time spent on MoC Process
Please estimate the additional time accepting Design EI
Please estimate the additional time accepting Construction EI
Please estimate the additional time accepting Commissioning EI
Please estimate the time spent recreating EI not delivered by Projects
Please estimate the loss of potential production
Please estimate the likelihood of professional error due to poor EIQ X likely cost to company

Please estimate the time spend manually re-entering Engineering EI
Please estimate the time spent asking for redundant information from vendors
Please estimate the time of Idle engineering resources due to delayed in EI
Please estimate the additional time spent approving mapping between system
Please estimate the time spent approving EI conflicts between systems
Please estimate the cost of redundant spares and materials
If you are working in Process Safety, please answer the following additional question:
Please estimate the Probability of a Process Safety Incident X Likely Consequence
If you are working in Operations and Maintenance, please answer the following additional questions:
Please estimate the additional time to call of contracts
Please estimate the additional time spent by TAR team (re) creating Work Packages
Please estimate the Additional Planned Downtime
Please estimate the additional Unplanned Downtime
Please estimate the increase in Availability
Please estimate the maintenance cost reduction
Please estimate the reduction in Capex
Please estimate the additional time required to predict asset remaining life
If you are working in Engineering Information Management, please answer the following additional questions:
Please estimate the additional time spent verifying that all stakeholders have the same EI
Please estimate the additional data format changes due to externally imposed changes
Please estimate the time spent identifying conflicts between systems
Please estimate the time spent converting data formats between systems
If you are working in a corporate function, please answer the following additional questions:
Please estimate the additional time spent assessing asset performance
Please estimate the Asset Performance deficit due to sub-optimization

Table 3.5 - Free-Form Text Questions and Closing Statement of the Initial Survey

Questionnaire Validation
After having seen the questionnaire, please answer the following questions:
Does the structure make sense?
Are there any unnecessary Impact Elements?
Are there any Impact Elements not included?
Is the format elegant and easy to understand?
Open Questions
This section seeks your input in a broader sense. Please speak freely and honestly.
Is there any important element of the effect of poor EIQ that has not been mentioned?
What, in your opinion, is the single cause of poor EIQ in industry?
What, in your opinion, is the single thing required to rectify the situation?
Close
Many thanks again for your time. Your contribution is invaluable to make our industry better.

The tables presented here describes the development of the contents of the Initial Survey. A short section follows to describe how the initial survey was implemented.

3.2.4 Mechanics of Data Collection for the Initial Survey

As was explained in Section 1.4.1, the initial survey was done to validate the survey questionnaire contents and understand more about the general subject. The data collection consisted of structured interviews with a few senior leaders at an operating

OpCo asset. The structured interview approach is deemed by Bryman et al. (2011) to be appropriate for business contexts and was selected because of the exploratory nature of the first surveys. The target audience was a convenience sample at the time.

The mechanics of data collection for the Initial Survey were as follows:

1. Invitation to participate was extended in person, after the support of the initiative was stated in a meeting chaired by the senior technical manager. The key technical and operational management of the following disciplines were interviewed per discipline for about thirty minutes each: Mechanical, Electrical, Instrumentation, Reliability, Process Engineering, Process Safety, Maintenance Planning, Maintenance Management.

2. Reminders were issued through the scheduling of meetings in the OpCo electronic calendars.

3. After the conclusion of the structured interviews, the data was analysed and the results reported to the respondents and the sponsoring senior technical manager. The results were used as the basis for a data remediation initiative at the OpCo.

The development and implementation of the initial survey yielded, in addition to the specific business case for the asset where it was conducted, a broad range of new information. These are described in the next section.

3.3 Validation of the Initial Survey

Upon completion of the validation structured interviews, the format and content of the initial survey was reviewed, based on the feedback received. This section summarizes the feedback received and the consequent changes.

3.3.1 General Changes

The general feedback was that interviewees struggled to answer the questions, largely due to a low awareness of the hidden value in most assets in this domain. Responses ranged from “that is too broad to answer” to “I don’t know the answer to that”. It was found that “direct” questions could be answered with relative ease, while “indirect” questions were more difficult to answer. As a result, the feedback was integrated into the questionnaire, thereby readying it for data collection. During this step, “Direct” questions were listed first and indirect questions towards the end of each section, to ease a respondent into the subject.

Another insight gained during these interviews was that business cases are evaluated quite conservatively by the target audience. This might indicate the case for more stochastic evaluations, as is discussed in Chapter 4.

As is shown in Table 3.5, four additional open-ended questions were asked during the initial survey. These were:

- Does the structure (of the survey) make sense?
- Are there any unnecessary IE’s?

- Are there any required IE's not included?
- Is the format elegant and easy to understand?

These questions, together with some informal feedback, provided the following points of learning that were incorporated into the final survey.

1. The objective of the research was not instinctively obvious and had to be explained verbally on a few occasions.

2. Some of the questions were difficult to answer and required considerable cognitive effort. In the light of points about fatigue and satisficing made by Hamby et al. (2016) and Lauer et al. (2013), it was therefore particularly important to minimise questions and phrase them as succinctly and clearly as possible. The introduction to the final survey was consequently somewhat expanded, including the addition of a few definitions.

3. Some of the questions overlapped in intent and were not mutually exclusive. This was, in retrospect, a predictable error for an immature survey and resulted in the reduction of the number of questions by about 50% in the final survey.

4. Presentation of the final results to senior management required more than a single figure; the ability to Pareto-rank the results from several perspectives, in service of pointing to actionable and prioritized remediation effort was also required.

3.3.2 Units of Measure

During the validation structured interviews, no units of measure were specified. Instead, it was intended to test the units of measure most frequently used by respondents. Regrettably, no pattern emerged and a plethora of units of measure were reported. These were:

- $/year
- % Availability
- % of Engineering Budget
- % BOE/day
- % Planned Downtime
- % Unplanned Downtime
- Average hours/day
- Average hours/project
- Hours/incident
- Probability X likely impact in $/year
- Work days/year
- Dollars per month
- Dollars per year

The variety of units of measure encountered add significant complexity to the model and in essence measure the same variables in different ways. The units of measure were therefore reduced to the most fundamental variables. In the case of time, hours/day was selected because it is the simplest of the range of time units encountered. As a result, the final survey contains only the following units of measure:

- Hours/day
- % (of Production)
- % (Change in Availability)
- Cost (in USD)

This reduced list of units of measure would simplify the deterministic model (Section 4.4) to a considerable degree.

3.3.3 Alternative Contribution

The NIST study deemed the aggregated wasted time to be equal to unnecessary headcount and the cost of the "extra" personnel was simply added to the total cost of the study. This was also the unspoken position taken during the initial survey. Upon review of the results, however, a challenge was posed regarding possible alternative deployment of "extra" personnel who would be released upon an improvement in EIQ. This line of logic posed the challenge "what is the potential contribution to value of an individual engineer, planner or maintenance leader" (which, it will be recalled, constituted the target population of this study). No precedent in this regard could be found in the literature. It is conceivable that studies may have been done in this regard within the OGI; however, these studies were necessarily not accessible for this research. It was subsequently suggested to leave the choice of invoking this option and the population of the input data, to the senior leader who would endorse a survey as envisaged in this study. It will be remembered from Section 3.3 that primary inputs are required by the senior leader endorsing a survey at a specific asset. Table 3.6 was consequently added to the list of initial "Asset Reference Data".

Table 3.6 - Alternative Contribution

Role	No of FTE in Role	Risk Reduction	Added Production	Reduced Cost
		[USD]/ person/year	[% nameplate boe/d]	[USD]/ person/year
Technician				
Engineer-in-Training				
Engineer				
Senior Engineer				
Planner				
Supervisor				
Manager				
Senior/General Manager				
Weighted Average				

These points of learning were included in the Final Survey described in the next section.

3.4 Final Survey

The learnings from both the literature survey and the validation structured interviews were combined to produce the final survey design, which is described in this section.

It will be recalled that general survey design principles were derived in Section 3.1 and specific design decisions were made in Section 3.2. These decisions remained generally valid, except for the specific changes and modifications enumerated in this section.

- Data collection was intended to be done by means of SurveyMonkey. It is described by Jooste (2014) as being self-administered and able to: maintain respondent anonymity, facilitate questionnaire design, collect data real time, email-enabled and export data for analysis. The OpCo, however, specified against SurveyMonkey. Instead, an internally approved survey tool was used. Given the requirement for a relatively large population of respondents, the need for low cost and the transcontinental nature of the surveys, there was little option but an email link to an online survey.

- Ease of use. Considerable effort was expended to minimise the number of questions, simplify and clarify the language and design the graphical interface to be as simple as possible. For example, the "Fine Print" section was presented in a smaller font as suggested by Lauer et al. (2013).

- Ranges and end rates. According to Hamburg (1974), the selection of the number of classes in an interval and their relative size is "essentially arbitrary". Hines & Montgomery (1980) add that class intervals should ideally be equally spaced and depend on the amount of scatter or dispersion of the data. Interval width is closely related to unit of measure.

Intervals and ranges were originally considered as shown in Table 3.7. However, upon review, ranges and end dates were not used in the final survey, since it was considered simpler for respondents to submit data in free form.

Table 3.7 - Units of Measure and Intervals Considered for the Final Survey

Class of UoM	UoM	Range	Interval width	Comment	Assumptions
Time or Additional Time Spent	[hrs/day]	0 to 10	1	A compromise between office/onshore hours and shift/offshore hours	Assume all onshore/office 8 hours/day and shift/offshore 12 hours/day; average 10 hours/day
Cost	[USD]	Free text		USD X 1000	
Production	[%]	0 to 24	2	% of rated production	
Availability	[%]	0 to 24	2	% change in Availability	

These modifications were included in the Final Survey.

3.4.1 The Structure of the Final Survey

Given the lessons learned from Section 3.3 and the change in data collection method, the Final Survey required a slightly different structure. This is shown in Figure 3.5 and subsequently described.

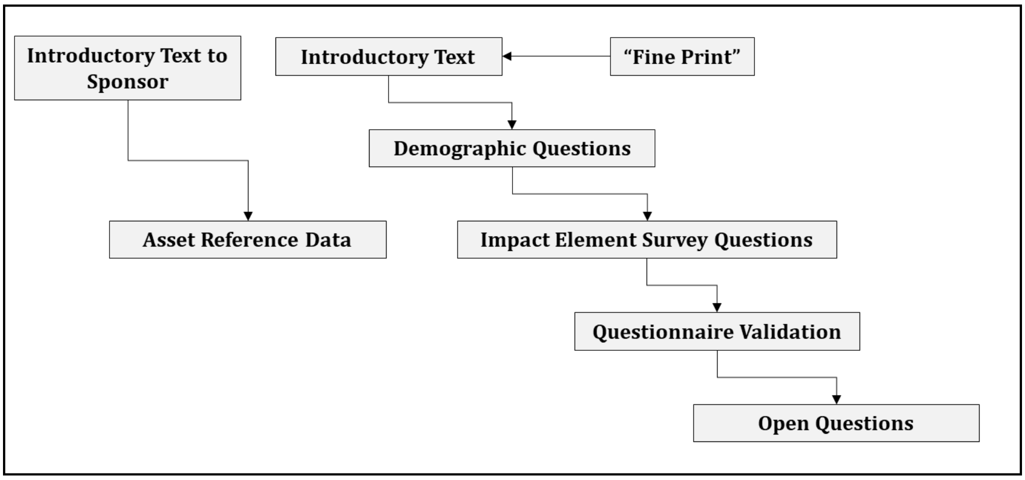


Figure 3.5 - The Structure of the Final Survey

Asset Reference Data (ARD) was not known, as was the case during the initial survey. It was therefore necessary to include a one-off data collection step to gather Asset Reference Data. The adopted procedure was to gather ARD from the sponsoring manager during the planning phase of a survey. The contents are shown in Table 3.8. The Asset Reference Data included an optional section on Alternative Contribution.

Introductory Guidance needed to be expanded, since the advantage of personal explanation during a structured interview had been forfeited. It was necessary to expand the introduction and add several definitions.

The "Fine Print" section was added.

Demographic Questions were left largely unchanged.

IE Survey Questions were simplified and shortened.

Open Questions were reduced to the following:

- In your opinion, what is the root cause of poor EIQ?
- In your opinion, what is the one thing that will solve the problem of poor EIQ?

These changes concluded the development of the Final Survey.

3.4.2 The Final Survey

All the preceding steps in this chapter culminated in the Final Survey, which is reported in this section in tabular form, as follows:

- Table 3.8 shows the Asset Reference Data for the Final Survey

- Table 3.9 shows the introductory text to the Final Survey
- Table 3.10 shows the text for the Fine Print section
- Table 3.11 shows the Demographic Questionnaire in the Final Survey
- Table 3.12 shows the IE Questionnaire of the Final Survey
- Table 3.13 shows the free-form text questions and closing statement of the Final Survey

Table 3.8 - Asset Reference Data for the Final Survey

Thank you for sponsoring a survey to establish the Cost of Poor Engineering Information Quality at your asset. We are confident that the results of this Survey will provide you with the ability to make informed decisions in this domain.

We require a small number of specific inputs to perform the calculations in the model. These are listed below.

- What is the nameplate production of your asset in [BOE/d]?
- What is the average annual total cost per Full Time Equivalent [USD/year]?
- What is the budget price of product for this year? [USD/BOE]
- What is the budget cost of production for this year? [USD/BOE]
- What is the budget plant availability for this year? [%]
- What is assumed cost of a PSI incident for your asset? [USD]

In addition, the survey model provides you with a choice regarding the response to reduced personnel requirements as the time wasted by poor Engineering Information Quality is reduced. The default approach is to deem increased efficiency equivalent to reduced headcount. There is, however, the option of "Alternative Contribution", whereby personnel coming available due to increased efficiency may be redeployed and add "alternative" value. If this is a feasible alternative for you, we request that you complete the table below. It is recognized that these figures can never be more than approximations. Please note that the required units of measure are given in [square brackets].

Table 3.9 - Introductory Text to the Final Survey

Thank you for making the time to take this survey. This survey tests the impact of poor quality Engineering Information on your day-to-day work. Improved Engineering Information Quality will help to simplify your job and help your asset to perform better. Obtaining this survey data is therefore important for you and the company.

The survey shouldn't take more than 20 minutes to complete.

There is some required information in the "Fine Print" box for your attention.

Please answer honestly using your best immediate estimate. We are asking for your experience, not an exact calculation.

Please answer the questions for your current or, if you are new in this role, your most recent role.

By Engineering Information Quality (EIQ) we mean all Documents, Drawings and Data related to equipment that you would refer to while performing your work. In other words, EIQ means reference information about equipment that is used repeatedly, not day-to-day transactional data.

At the end of the question set there is room to provide free-form text comment.

The questionnaire is divided into three parts:

- Non-confidential data about you and your job. We need that for future understanding of training needs etc.
- The Questionnaire. Please note that there are 'filter' sections per job group; you need only to answer the sections relevant to you.
- Open-ended questions, where you can add general comment.

You will come across the phrase "additional time spent" in the questionnaire. We mean by that "unnecessary extra time needed for the specific activity due to poor EIQ". You will also come across the phrase "time spent" in the questionnaire. We mean by that "unnecessary time needed for something that should be automatic and take no time at all".
For the "time" category of question, please respond in [hours/day].
For the "production" category of questions, please respond in [%] terms
For the "cost" category of questions, please respond in [USD].

Table 3.10 - Text for the Fine-Print Section

Your responses will be used to determine if there is a business case to improve engineering information in an asset.
Your contribution is confidential.
Demographic data is collected to analysis only and your identity is not retained.
Your specific responses will not be known to the survey analysts or your management.
Participation in this survey is entirely voluntary.
There is no obligation to complete the questionnaire.
This survey has been developed as part of a Master's Degree in Industrial Engineering.

Table 3.11 - Demographic Questionnaire in the Final Survey

DEMOGRAPHIC QUESTIONS	
A reminder that this data is captured for analysis only, will be kept anonymous and not be used for any other reason. We will use this data to better understand training and support needs.	
Please complete the following data:	
What is your current asset?	
What is your primary discipline?	Electrical
	Electronic
	Mechanical
	Chemical
	Civil/Structural
	Mining
	Geology
	Other (please specify)
What is your highest qualification?	School
	Diploma
	Bachelor's Degree
	Master's Degree
	MBA
	PhD
	Other (please specify)
After completing your studies, how many years of professional experience do you have?	0-5 years
	5-10 years
	10-20 years
	20-30 years
	30+ years
In which part of the industry have you worked the longest?	O&G Upstream
	O&G Downstream
	O&G LNG
	Mining
	Other (please specify)
What is your current role?	Technician or Technical specialist
	Engineer-in-Training
	Engineer
	Senior Engineer
	Planner

	Supervisor
	Manager
	Senior/General Manager
	Other (please specify)
In which team/department are you currently working?	Operations
	Electrical
	Mechanical (Machinery)
	Mechanical (Integrity)
	Instrumentation/Control/Automation
	Maintenance
	Maintenance Planning
	Turnaround
	Process Engineering
	Process Safety
	Corporate Planning/Analysis/Strategy
	Projects
	Engineering Information/Technical Documentation
	Other (please specify)

Table 3.12 – Impact Element Questionnaire of the Final Survey

Question	UoM
Please estimate the Additional Time spent looking for Engineering Information (EI)	[hrs/day]
Please estimate the Additional Time spent verifying or re-entering EI	[hrs/day]
Please estimate the Additional Time repeating processes	[hrs/day]
Please estimate the Time spent to validate/prepare corporate KPIs due to poor EIQ	[hrs/day]
Please estimate the Time spent clarifying misunderstandings surrounding EI & EIQ	[hrs/day]
Please estimate the annual Financial impact of misunderstanding	[USD]
Please estimate the Additional Time spent reviewing EI Standards	[hrs/day]
Please estimate the Time spent creating and maintaining unofficial databases	[hrs/day]
Please estimate the Time spent resolving EI conflicts between databases	[hrs/day]
Please estimate the Additional time spent optimizing the budget and production plan	[hrs/day]
Please estimate the Production loss due to the wrong data being reported	[%]
Please estimate the Additional time spent preparing Regulatory Reports	[hrs/day]
Please estimate the Additional time spent responding to Regulatory Queries	[hrs/day]
Please estimate the Likely Cost of a regulatory penalty due to Poor EIQ	[USD]
Please estimate the additional Time required to prepare Decision Review Board Support information	[hrs/day]
If you are working in Engineering, please answer the following additional questions:	
Please estimate the Time spent recreating EI not delivered from Projects	[hrs/day]
Please estimate the Time spent obtaining lost EI from Vendors	[hrs/day]
Please estimate the Idle/non-productive Engineering resource time due to EIQ	[hrs/day]
Please estimate the Additional Time spent updating EI in the MoC Process	[hrs/day]
Please estimate the Reduced production due to poor design	[%]
Please estimate the Additional Time accepting/verifying/approving EI from Projects	[hrs/day]
Please estimate the Additional Time spent approving mapping EI between systems	[hrs/day]
Please estimate the Cost of redundant scrapped material	[USD]
Please estimate the Cost of redundant procurement	[USD]
Please estimate the Cost of redundant construction	[USD]
Please estimate the Likely Cost of professional error due to EIQ	[USD]
If you are working in Operations and Maintenanace, please answer the following additional questions:	
Please estimate the Additional Time (re) creating Maintenance PMs	[hrs/day]
Please estimate the Idle/non-productive time in Maintenance resource time due to EIQ	[hrs/day]

Question	UoM
Please estimate the Additional Time to call of contracts/ mobilize vendors	[hrs/day]
Please estimate the Additional Time spent by TAR team (re) creating Work Packages	[hrs/day]
Please estimate the Additional Time spent optimizing maintenance resource	[hrs/day]
Please estimate the Additional Time required to optimize asset operation	[hrs/day]
Please estimate the Increase in Availability if EIQ was better	[%]
Please estimate the Cost of performing unnecessary inspection/ maintenance due to poor EIQ	[USD]
Please estimate the Cost of redundant spares in warehouse	[USD]
Please estimate the Cost of spares expediting ("hot shot costs") & management	[USD]
Please estimate the Production loss due to Asset sub-optimization	[%]
If you are working in Process Safety, please answer the following additional questions:	
Please estimate the Likelihood of a Process Safety Incident due to poor EIQ	[%]
If you are working in a corporate function, please answer the following additional questions:	
Please estimate the Additional Time spent assessing asset performance	[hrs/day]
If you are working in Engineering Information Management, please answer the following additional questions:	
Please estimate the Time spent identifying conflicts between systems	[hrs/day]
Please estimate the Additional Time verifying all that stakeholders have the same EI	[hrs/day]
Please estimate the Additional Time spent mapping between systems	[hrs/day]

Table 3.13 - Free-Form Text questions and Closing Statement of the Final Survey

Free-Form Text Questions
In your opinion, what is the root cause for poor EIQ?
In your opinion, what is the one thing that will solve the problem of poor EIQ?
Closing Statement
Thank you again for your time. We will analyse your data with others in your team and report the results to your management. We hope that there is a strong business case for improving your data quality.

The contents of the survey have been finalised; the next section discusses how data was collected.

3.4.3 Mechanics of Data Collection for the Final Survey

Since the Final Survey was intended for a much larger audience and was to be done remotely, more effort was required to mobilize data collection.

In addition to Gackowski (2009) and similar material elsewhere in the literature, some effort needed to be made to overcome the reluctance to participate. Numerous introductory presentations were made across a multinational OpCo to obtain permission to run surveys at specific assets. The objections raised during these presentations were valuable in the sense that they enabled the presentation contents to pre-empt and prepare for objections and enabled further refinement of the method. Once permission was granted, the process to activate the data collection was specific to each asset, but the following general process steps were followed:

1. An introductory email was sent to a specific population by the leader granting the permission. In this email, the objective and mechanics of the survey were

explained, together with the sponsoring manager's stated support and the value it could bring to the asset and the corporation.

2. Next, the electronic link to the survey was sent to the target population by email, reiterating the value and the mechanics. A completion date, normally two weeks after the sent date, was also included in this email.

3. Two subsequent reminder emails were sent to the population, one at the one-week mark and one two days before survey closure.

4. Upon survey closure, a closing email was sent to the sample population and a different email sent to the sponsoring manager. In the email, the sponsoring manager was thanked for permitting the survey, the response rate was reported, and a date committed at which time the results would be reported back.

This chapter has described the full evolution of the Survey used to collect data. The next chapter will describe how the results of the collected data were to be analysed and presented.

Chapter 4

Analysing Survey Results

This chapter defines the methodology and conventions used to analyse the results. Consistent with the convention applied in Figures 2.1 and 3.1, the sections of the chapter proceed along the logic outlined in Figure 4.1.

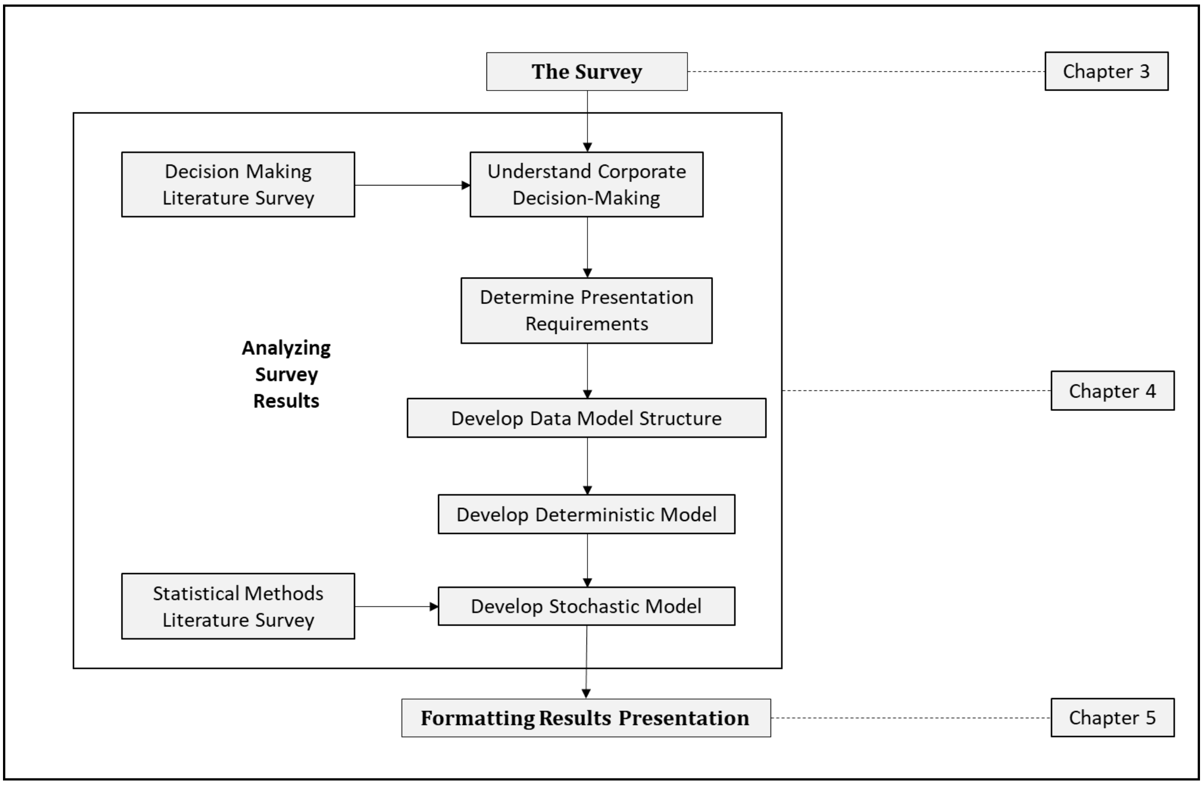

The Survey
Chapter 3
Decision Making Literature Survey
Understand Corporate Decision-Making
Determine Presentation Requirements
Analyzing Survey Results
Develop Data Model Structure
Chapter 4
Develop Deterministic Model
Statistical Methods Literature Survey
Develop Stochastic Model
Formatting Results Presentation
Chapter 5

Figure 4.1 - Analysing Results Detail

As shown in Figure 4.1, the description of the CoPEIQ model is interrupted by a review of corporate decision-making, which in turn provides the format in which model results are presented. This insight is necessary to define the outputs of the model, which is subsequently described.

4.1 Understanding Corporate Decision-Making

In this section, the objective was to gain an overview of the current thinking with respect to corporate decision-making and specifically the information presentation requirements during the corporate decision-making process. The former objective was intended to provide context for the latter and the latter, in turn, supported of understanding the presentation requirements of the stochastic model that is to be selected in Section 4.5.

It is noted at the outset that the subject of corporate decision-making constitutes a very large body of knowledge. (A Google Scholar search conducted on 31/01/2017 for the text "Corporate Decision Making" yielded 1.85 million results). The intent in this section is not to research this subject in depth. Instead, the objective is to gain enough insight into it to provide context for the design of the statistical model and the subsequent format of results presentation.

4.1.1 Literature Review: Corporate Decision-Making

The sample of literature reviewed for this section has exposed several themes in the literature. These themes are illuminated in the sections below.

4.1.1.1 The Relationship between Decision Quality and EIQ

It may generally be said that quality decisions are at least partly dependent on accurate reporting metrics, which in turn rely on good EIQ. For example, it would be difficult to optimise the resource loading of an integrity management programme for a refinery if the foundational asset register was incomplete or the equipment class attributes used for searches and filters were assigned incorrectly. Indeed, Tam and Price declare that “The availability of useful data is paramount to making the best decision in asset management”. (Tam and Price 2008). This is confirmed by Haider (2005), who explored the cycle of learning, optimization and change, and concluded that a crucial factor in this cycle is completeness of asset foundational data. Haug et al. (2011) confirm that decision-making processes are rendered inefficient by poor data quality.

4.1.1.2 Rationality of Corporate Decisions

The extent to which decisions in corporations are rational is a prevalent theme in this body of literature and has been for decades.

Prof. James Reason is generally accepted as the leading authority on the subject of human error and how it impacts organisational performance. Although his book “Human Error” (1990) is now somewhat dated, it provides a useful and concise summary of the evolution of understanding of human judgement and decision-making. He explains that in the period preceding the 1970’s, it was assumed that decision-making was rationalist, applying Bayesian paradigms and "assumes that people always know what they want and choose the optimal course of action for getting it". The Subjective Expected Utility (SEU) Theory assumes that decision makers have a clear definition of their utility function, understand their options exhaustively, can understand the probability distributions of the outcome of every option and will choose the option that maximises their ‘subjective expected utility’.

Reason proceeds to quote Fischoff (1975 and 1978) as concluding that, in fact, decisions are made by significantly less rational processes and that these processes are further exacerbated by "hindsight bias". He then quotes Simon's (1975) "bounded rationality" and concludes that humans are likely to engage in "satisficing behaviour, the tendency to settle for satisfactory rather than optimal courses of action". Since decisions in the OGI are often made by groups, it is of note for this study that bounded rationality is equally true for individual and collective decisions.

As if to reinforce the preceding, Reason then quotes Tversky & Kahneman: "man is apparently not a conservative Bayesian[3]: he is not a Bayesian at all". Man, instead, utilises a relatively short list of heuristics to simplify judgemental operations.

Reason concludes this discourse by describing man's propensity to minimise cognitive strain by applying strategies like persistence-forecasting, or using cues that have proved successful in previous experience. This avoidance of cognitive strain, termed 'reluctant rationality', "direct our thoughts along well-trodden rather than new pathways". By way of example, he quotes the failed decision-making processes preceding the Bay of Pigs event to demonstrate how groupthink can suppress rational decision-making.

Thawesaengskulthai & Tannock (2008) set out to develop a decision aid for selecting improvement methodologies like Six Sigma and the like. They start with a review of the thinking around decision-making and touch about how rational it is in organisations. They quote four studies to conclude that there is a push towards more rational decision-making, but acknowledge the influence of irrationality in this subject. They quote Charlesworth (2000) and Weiler (2004) that conclude that decisions are influenced by perceptions about what constitutes best practice and what benefits may be achieved by it, as suggested by "friends, colleagues, gurus and practitioner publications". Significantly, quoting Clark and Greatbatch (2004), they state that management ideas may have become popular because of their perceived beneficial effects, "not because they actually work".

Summarising the results of interviews with forty-four Thai thought leaders, they conclude that decisions in this domain could be grouped into four views: current management thinking (fashion), achieved benefits, strategic fit and fit into the organisation.

Focusing then specifically on the presentation of results, they quote Hill (1995, 2000), Platts (1990) and Akao (1990) to support the point that pictorial performance profiles have been used in various applications and "should have credibility with senior managers".

Zastron (2016) consulted twenty-six managers and concluded that 68% of them felt that opinions, rather than factual data, drove decisions, even in the presence of such data. Zastron quotes McAfee and Brynjolfsson (2012), who concluded that decision-makers "may be more interested in using their intuition than the facts provided by data." In fact, they proceed to conclude that most executives do indeed prefer intuition and judgement over factual data.

In summary, these sources appear to conclude in general that decisions are biased towards irrationality and simplifying heuristics, or influenced by external perceptions.

[3] Bayesian Decision Theory is derived from the original work of The Rev. Thomas Bayes (1702-1761) and has been described by Hamburg (1974) as a method to update previously judged or assumed probabilities as new or additional data becomes available.

There are, however, several authors who take a contrary view; some of these are discussed below.

Hamburg (1974) explains that Bayesian Decision Theory has been developed in the period since 1945 to "solve problems involving decision making under uncertainty". He notes that Bayes' Theorem is used in business applications to supplement the judgements and intuitions of business leaders by emerging empirical data to update probabilities. Hamburg also introduces the subject of Hypothesis Testing by stating it may be used to control the risks of poor decisions under uncertainty.

Cabanous et al. (2010) seek to influence organisations towards a more rational paradigm by means of "performativity", which is described as several processes intended to cause theory to influence reality until it ultimately becomes reality. Indeed, they quote the contention that "the rational man can be brought into being within organisations" (Feraro et al. (2005)) and sets that as a contradiction to the "reviewed wisdom". The difficulty of this objective is however acknowledged and it is concluded that rationality is a "fragile product". They quote Keeney (1982) and list the axioms of Decision Analysis as being the Generation of Alternatives, Identification of Consequences, Quantification of Judgement and Preference, Comparisons of Alternatives, Transitivity of Preferences and Substitution Consequences.

The relationship to the current research is confirmation of the "lasting debate on the inherent rationality vs irrationality of organisations and their decisions."

It is evident that a continuing effort to influence organisations towards rational decisions exists and that decision aids are used in that drive, but its elusive nature is acknowledged, and the objective of rational decisions is far from achieved.

4.1.1.3 The Influence of Context

The extent to which the inherent rationality of decisions is influenced by context is explored in the next theme in the literature. Three contextual factors are explored: The impact of the availability of expertise for decision support may or may not influence decisions; how interruptions increase the cognitive strain on decision-makers; and the business context limiting optimal decisions.

Farrington-Darby et al. (2006) published a review of expertise and offered some explanations for the range of points of views of the subject, with a focus on ergonomic work in complex and dynamic contexts. They conclude with a proposal to a more "naturalistic" approach to work design. Ghosh et al. (1993) conducted a laboratory experiment to test the extent to which the structure of uncertainty, and specifically the centre and range of the probability distribution, is a determinant of choice. The work focused on the likelihood of tax evasion given the possibility of an audit and develops a relationship between ambiguity and the value of risk. The relevance to this study is the impact of the presentation of information on the decision.

Farrington-Darby et al. (2006) quote prior research (Gosh et al., 1992 and Mackay et al. 1992) that concludes that individuals' response to uncertainty is determined to some extent by the knowledge of the event in question by that individual.

Significantly to this study, they conclude that "most decision models and empirical studies … continue to assume that probability theory provides an adequate characterization of all the decision-maker's uncertainty". They proceed to reference, like Reason, the SEU, but then mention the work of Ellsberg (1961) that suggests that decisions are likely biased towards clear, rather than ambiguous, events. It appears from this that the ambiguity of a probability, as opposed to the actual probability itself, may be a determinant of decisions. Put another way, SEU assumes that decisions are influenced solely by the probability of an outcome, whereas Ellsberg demonstrates that decisions are more likely to be driven by which of these outcomes decision-makers know most about. Since one objective of the present study is to determine the most suitable way to present information to decision-makers, it is significantly to note that "the structure of uncertainty" is defined by the authors as "the mean and range of the uncertainty".

Farrington-Darby et al. (2006) contend that optimal decisions can actually be made, but that the evaluation of expert decisions is frequently compared with what statistical models and theory recommend. From this basis they propose that "Naturalistic Decision Making" (NDM) settings contain elements of what might very well be a real-life context for OGI decisions: poorly structured problems, fast-changing contexts with high uncertainty, competing objectives, control loops, very high stakes, time pressure, multiple stakeholders with divergent agendas.

In their study of the identification of domain experts, Malhotra et al. (2005) contend that decision-makers depend on their personal experience to make the final decision, because, despite the assistance of decision support tools, they cannot model "the complexities of the real world". Their work concludes that the breadth of experience of domain experts is the best indicator of expertise, and that the number of years of experience is "imperfect", depth of experience is a poor indicator and judgements about other peoples' expertise could be "substantially incorrect". Of particular interest for this study, is their extensive use of performance profile graphical presentation mentioned by Thawesaengskulthai & Tannock (2008).

In their investigation into the influence that interruptions have on various types of decision tasks and the extent to which information presentation formats can reduce this influence, Speier et al. (2003) provide useful insight into the cognitive workload during decisions. After confirming from the literature that decision performance is indeed influenced by information presentation (quoting DeSantis (1984) and Tan & Benbasat (1990) and Vessey (1991)), they proceed to model interruptions.

The discourse of Speier et al. (2003) about the interaction of interruptions and presentation format is worth significant attention: They contend that "a widely shared belief" exists that the effectiveness of a presentation format is determined by the task performed. Citing Vessey's (1991) Theory of Cognitive Fit, they explain that an alignment of presented information and the processing required to most easily complete the task results in a "cognitive fit". As a result, decision-making is facilitated, since the processes of decision-making and problem-solving align. If this cognitive fit does not occur, greater cognitive effort is required to digest the data, leading to reduced performance in the decision-making process.

From this analysis, they conclude that analytical processes, such as calculations, align best with discrete sets of symbols, while perceptual processes like visual comparisons align best with showing relationships between sets of symbols. The cognitive effort is exacerbated by environmental stresses, such as interruptions and time pressure. In their analysis, they also include expertise as a factor.

The conclusion of Speier et al. (2003) with respect to information presentation is remarkably concise: "decision-makers who are interrupted when solving complex symbolic problems are better supported by graphs than by tables". This conclusion is of interest to this study.

The notions of cognitive load and cognitive fit are explored in some detail by Zastron (2016). He concludes that the "difficulty associated with interpreting information is influenced by the format in which [it] is conveyed". He suggests that cognitive strain could be reduced by presenting information in a format that is both familiar to the recipient and supportive of the task at hand.

The notion of cognitive effort is explained in more detail by Toker et al. (2012), as quoted by Zastron (2016). They identify two types of factors that influence the ability to absorb visual information. Short-term factors include attention of the user and cognitive load of the communication. Long-term factors are experience, expertise and the recipient's inherent cognitive ability.

Zastron (2016) also quotes the finding by the Institute of Asset Management (IAM's) that decisions within Asset Management are regularly constrained by business content such as regulations, resources and budget limits. This finding infers that decisions, particularly with respect to funding (such as in the case explored in this thesis), are often a balancing act between conflicting priorities and specific imperatives.

In the ebb and flow of an operating asset, it can be confidently assumed that actual or purported expertise will be readily to hand, that interruptions and other interfering factors will exist, and that funding constraint will be a ubiquitous reality.

4.1.1.4 Uncertainty and Preferences in Decisions

The final references to literature serve as illumination of the likely intellectual preferences decision-makers will follow, namely that they are inclined to follow their own preferences and tend to prefer the best-understood option, uncertain though it may be.

Olson et al. (1995) investigated the extent to which the factors underlying theory, preference information elicitation and alternative structure influence the results of different decision aids. They conclude that it is important to consider the accuracy of information showing the preference of the decision-maker. They quote five sources that agree that one rationally "best" decision does not exist in an environment of multiple attributes. The preferences of the decision-maker, which are implicit, will therefore likely drive the decision taken. It follows that the design of a decision support system should invest effort in understanding this implicit preference set, using sensitivity analysis.

The results of the study of Olson et al. (1995) indicate that "naive users prefer simple systems" and that the preferences of the decision-maker appear to be more important than the underlying model itself. It follows that the inputs provided to the decision-maker should be simplified and "more natural" to the decision-maker to improve accuracy. Importantly for this study, they conclude that "exact numerical data for complex concepts is … not necessary".

This is supported by Hamburg (1974), who observes that in the context of business and social sciences, statistical relationships, rather than exact relationships, prevail.

4.1.1.5 Formats for the Presentation of Results

Several indicators of how to present information have been gained from the literature consulted above. These are briefly discussed below.

The insights acquired about the propensity for reluctant rationality around complex decisions are useful for this study. They indicate the necessity to present results in the most distilled and palatable format possible, to minimise cognitive strain. (Reason (1990) and Speier et al. (2003)).

Gosh et al. (1993) show the importance of a decision-maker to understand "the structure of uncertainty" and then proceed to define this as "the mean and range of the uncertainty".

Fess (1991) implies that a statement of probability serves as assurance that the reported results are reliable.

Olson et al. (1995) conclude that precise numerical presentation for complex concepts is "not necessary".

Speier et al. (2003) are confident presentation format should align with the task being performed. They conclude that graphs are better supporters than tables of complex, symbolic problems in the context of interruption.

Thawesaengskulthai & Tannock (2008) contend that senior managers should find pictorial performance profiles credible. This format is used extensively by Malhotra et al. (2005).

Bryman et al. (2011) suggest that bar charts, pie charts and histograms are often utilised for the display of quantitative results because of ease of interpretation. They suggest that bar charts or pie charts are appropriate for nominal or ordinal variables and histograms for interval or ratio variables. They also propose a so-called 'boxplot' to combine the demonstration of central tendency and dispersion in one graphic.

Zastron (2016) perform a comprehensive review of information presentation. He distinguishes between exploratory graphics, which are intended to convey the general patterns and attributes of a large dataset and presentation (or explanatory) graphics, which should be well-defined, of good quality and presented with the target audience in mind. He distinguishes between text, tables and graphs. The applicability of each category is summarized as follows:

- Text is appropriate for fewer than five data points, low interaction, slope, or shape.
- Tables are appropriate for more than five data points, specific numeric values and accurate representation.

- Graphics should be used for relationships, trends and cognitive support such as enhanced recognition.

Finally, Zastron (2016) presents an exhaustive analysis of graphical presentations formats. He reiterates the convention that the horizontal axis is usually used to present independent variables (those under the control of the researcher) whereas the vertical axis is used for dependent variables, or those variables under investigation by the researcher. He discusses, like Bryman et al. (2011) nominal, ordinal, interval and continuous variables and offers a useful method to calculate the number of bins for interval attributes developed by Scott (2009). He then discusses the attributes, advantages and likely application of various graphical information presentation formats. These are summarized in Table 4.1.

Table 4.1- Graphical Presentation Format Comparison

Graphic Format	Advantages	Disadvantages	Applicable; Comment
Bar graph	Simultaneous absolute and relative data Effective for small sizes/large differences	Care indicated since definitions differ Not ideal for proportional comparison	Compare values of many entities at one point in time
Box and whisker plots	Simple, effective comparison of distributions of datasets	No indication of independent variable distribution	Identify anomalies in dependent variables
Columns graphs	Show change over time		Show differences between groups of independent variables
Histograms	Show irregularities in data	Sensitive to interval selection	Continuous data
Line graphs		Only for ordered (ordinal or interval) data	Show change over time
Area graphs		Not good for specific values Does not reflect crossing series	Show trends, relationships, specific emphasis
Pie chart	Outperform column graphs for part-to-whole relationships	Multiple, adjacent graphs not effective	Depict proportions 3D graphs less effective
Radar graphs	Compare many variables simultaneously	Not good for subtle differences	Show relative strengths/weaknesses between two groups, or between one group and a reference
Point graphs	Events without set independent variables		Show approximate relationships

4.1.2 Conclusions from the Literature

The exploration of the literature in the preceding sections has yielded important clues about the original purpose of this section, namely to arrive at an understanding of the information presentation features that OpCo's need to support decision processes.

The literature review in the preceding section may usefully be summarized as follows:

- Based on this sample of the literature, the debate about decision rationality in organisations does not appear to be settled. Efforts do exist to influence or drive organisations towards rationality, but the difficulties are acknowledged.

- Decision-makers experience cognitive strain when faced with increasing complexity, external pressure and interruptions. They are inclined to rationalize towards a 'reluctant rationality' to reduce cognitive strain.

- The preference of the individual decision-maker ultimately carries the day and an understanding of an uncertain option, rather than the quantification of the uncertainty, is likely to drive towards that option.

- The presentation of the information does assist in decision-making and graphical formats are preferred. Results that are uncertain in nature should be reported by in stochastic terms.

From these insights it may be gathered that the actual magnitude of uncertainty is less important than its attributes, implying that a conclusion should be reported with its mean, range (standard distribution) and its confidence. In addition, presentation should be in an appropriate graphical form, for which a few conventions exist.

These insights are subsequently interpreted for the purposes of this research.

The present study is inherently rationalist. Its very intent is to determine the financial impact of poor EIQ in a credible, repeatable format. The objective is to present that impact to decision-makers in a way that will of enable them to evaluate the merits of an EIQ remediation drive in a comparable form to that of other proposed initiatives presented to them. The extent to which a specific OpCo chooses to attempt rational decision-making is therefore left for that OpCo to decide; instead this study seeks to present information that will enable a rational decision.

The complexity of the decision process, including the multitude of factors affecting it and the exacerbating factors increasing the cognitive strain of the decision-maker, is useful context for this study. It points to the responsibility of this study to present its conclusions in a way that enables the most rational decision with the least possible cognitive strain.

4.2 Determine Presentation Requirements

From the conclusions drawn in Section 4.1.2, it follows that this research should present its outcomes as follows:

- Results should make it possible for decision-makers to understand the structure and magnitude of the results' uncertainty. This means in practice that the centricity, dispersion and confidence level should be reported. These attributes are more important than the exact results.

- Results should be reported in the most appropriate graphical form.

The understanding about the dynamics of corporate decision-making gained in Section 4.1 yielded a specification of presentational requirements. The preparation for the development of the model is therefore nearly complete; a brief discussion on data structure follows in Section 4.3, after which the model will be described in Sections 4.4 and 4.5.

4.3 Develop the Model Data Structure

This section describes the structure of the CoPEIQ model. It is done to provide the reader with a conceptual overview of the details described in Sections 4.4. and 4.5. Some of the concepts mentioned will only be formally identified and discussed in Sections 4.4 and 4.5.

The section starts with a brief look at data model theory, after which it will be applied to this study.

4.3.1 General Data Model Concepts

In this section, the subject of data models is reviewed briefly, based primarily on Hoberman et al. (2009). This is done as a preliminary investigation, or "sanity check", with the objective of determining the need for its application during the design of the CoPEIQ database, which is addressed in subsequent sections of this chapter.

At the outset it is noted that Hoberman et al. (2009) declare that there is no consensus on the terminology used in the industry. This concurs with observations made in the OGI during this research. The observations in this section are therefore introductory and conceptual, rather than definitive.

Hoberman et al. (2009) define a data model as "a visual representation of the people, places and things of interest to a business". They proceed to explain that data models are usually presented in a hierarchy of increasing detail, as shown in Table 4.2.

Table 4.2 - Levels of Data Model

Data Model Level	Title	Summary Objective
1	Very High Level ("Conceptual")	Align on Scope and Common Meaning
2	High Level	Gather requirements
3	Logical /Entity-Relationship (ER) Model	Incorporate Business Rules
4	Physical Model	Technical details of the data structure, design specification of a database

The levels described in Tale 4.2 can generally be described as follows:

- Level 1 is appropriate for designing macro-level data and system architectures for integrated corporations. It may describe, for instance, how personnel data, including recruitment, competency and remuneration details, are kept separate from equipment, production and financial data. By implication, several databases or systems are involved and need to either interact or be deliberately kept apart in specified ways.

- Level 2 is frequently used to gather more detailed requirement for the macro-entities defined in Level 1. One example might the requirement to transmit remuneration details from the personnel database to the annual budget system.

- Level 3, an entity–relationship (ER) model, describes information entities and their relationships, whether potential or actual. Hoberman et al. (2009) deem this “an ideal choice” when a physical database is intended. It is also one of the most common architectures. An example might be the need to transmit the pressure setting of a safety valve from the design database to the maintenance procedure system.

- Level 4, the physical model, contains technical details required by programmers to execute the architecture and functionality required by the preceding levels. For instance, the tag number of a certain equipment class may be in the form NN-AA-NN-NN in SQL format (where “N” is a number and “A” a letter.)

It may be seen from the preceding that these levels represent a generalized classification of decreasing scope and increasing detail, that jointly could constitute a design of the entire information management system of a corporation.

4.3.2 Application to this Study

Upon review of the subject, as described above, it is concluded that a body of knowledge does exist to design an information management system and its constituent database parts and that this body of knowledge is predominantly intended for the design of complex IT systems and their interactions. Since the database required for this study is a single database, for which no interaction with other databases is envisaged, level 3, the so-called ER model, is of the most interest for this study.

Level 3 is concerned with defining entities and their relationships. Figure 4.2 shows the information entities envisaged for this model and their interaction during the calculations relating to CoPEIQ. These entities and their respective constituent parts are defined and discussed in Section 4.4 and 4.5.

Level 4, or the physical model, is of lesser interest for this study, as will become clear in Section 4.3.3.

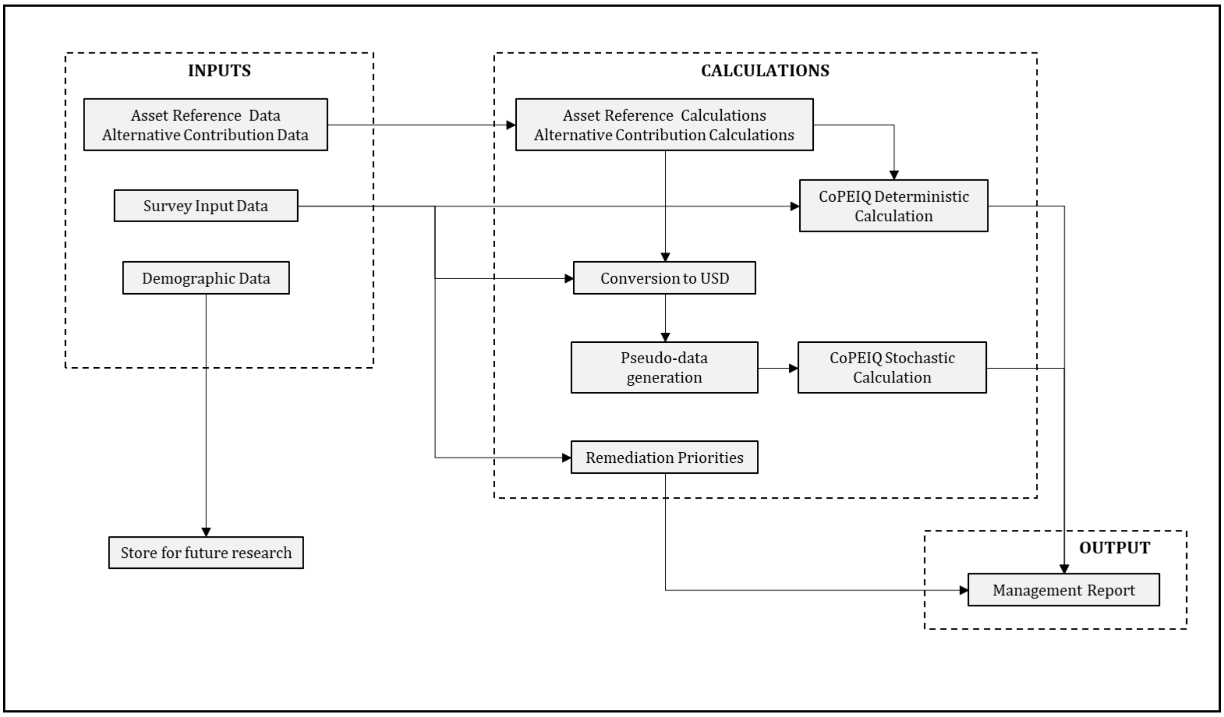

INPUTS
Asset Reference Data
Alternative Contribution Data
Survey Input Data
Demographic Data
CALCULATIONS
Asset Reference Calculations
Alternative Contribution Calculations
CoPEIQ Deterministic Calculation
Conversion to USD
Pseudo-data generation
CoPEIQ Stochastic Calculation
Remediation Priorities
Store for future research
OUTPUT
Management Report

Figure 4.2 - Simplified Entity-Relationship Model for the CoPEIQ Models

The high-level structure of the CoPEIQ models having been designed, a final step remains before the detail of the models is described. The selection of appropriate software and the structure of the software is discussed in Section 4.3.3.

4.3.3 The Structure of the CoPEIQ Model

In this section, the software selection is discussed, together with a brief discussion of the software structure.

During the literature survey and discussions of this study in the OGI, several options were discovered for models of this nature. Most are of a commercial nature, either highly specialized or as simpler add-ons to Microsoft Excel. It was decided, however, to simply use Microsoft Excel, for the following reasons:

- As will be shown in Sections 4.4 and 4.5, the CoPEIQ models are mathematically simple enough to be well within Excel's standard functionality.
- The value of this study to the OGI is maximised by the work being made as accessible as possible.
- The academic value for future research is also maximised by the work being made as repeatable as possible.

As a result, the models were built in one Microsoft Excel spreadsheet. The worksheet structure follows the logic and sequence shown in Figure 4.2 and will be described in detail in Sections 4.4 and 4.5.

4.4 Develop the Deterministic Model

This section describes the development of the deterministic model. The deterministic model is primarily intended to be used as the repeating algorithm for the stochastic model discussed in Section 4.5. It does however, also provide the means for a simplified calculation of CoPEIQ where a quick order of magnitude is required or where a survey is not practical or permitted.

The deterministic model is developed using the following steps: several new concepts are introduced; the component parts of the model are described; the detail calculations are given. The section concludes with an example calculation in Microsoft Excel. A brief pause is however required first to further explore introductory concepts.

4.4.1 Introductory Concepts

In this section, the concepts defined in Section 3.2.3 are discussed in further detail with specific reference to its use in the deterministic model.

4.4.1.1 Barrel Oil Equivalent

The cost of producing one BOE and the price for which a BOE is sold will form the basis of calculating revenue in the deterministic model.

The calculation of revenue for an OGI asset is in fact much more complex than using BOE. It is determined by the fraction of oil versus gas, the additional cost of produced water, whether stimulation of the geological formation is used or not, the tax and royalty regime in force where the asset is located, the rules of Capex vs Opex applied by the OpCo and a plethora of other factors. These variables are out the scope of this thesis, instead the generalized variable BOE is used, because it is assumed that each asset and OpCo will have discounted these various factors in both their cost definitions and economic evaluations. If comparisons between assets or OpCos are to be done, however, these variables become important.

4.4.1.2 Full Time Equivalent and Cost of Full Time Equivalent

The cost of one FTE (CFTE) is defined for this thesis as being the full annual cost of employment or engagement of an individual. It is intended to include such items as salaries or rates, additional benefits, overheads for safety equipment and training. If the resource is deployed offshore or at a remote “camp” location, it is intended to include the cost of the camp or installation, flights or other means of transport and all related incidental costs. Moreover, the deterministic model assumes one figure for all levels of staff. These generalizations are customary in the OGI for calculations of the nature described in this thesis.

4.4.1.3 Plant Asset Rated Production

The actual production on a daily or annualized basis will vary considerably due to a variety of operational or economic factors and is, like PSI and FTE, the subject of considerable debate and variation in industry. For an economic evaluation such as the subject of this study, terms such as “nameplate” and “production” are frequently used interchangeably and refer to a constant rate which is most frequently the production rate budgeted for the year.

4.4.1.4 The Cost of a Process Safety Incident

The Cost of a PSI extends beyond direct cost to "such as business opportunity, business interruption and feedstock/product losses, loss of profits due to equipment outages, costs of obtaining or operating temporary facilities, or costs of obtaining replacement products to meet customer demand". The cost of a PSI could extend far beyond these elements. As an extreme example, USA Today reported on July 14, 2016 that the Deepwater Horizon example cost BP USD 62B, including fines, tax impacts and loss of share price.

These variables indicate the complexity of declaring or estimated the cost of a PSI. Like calculating the profitability of an OGI asset, these complexities are beyond the scope of this study. Instead, the intent with this concept is to use the figure that the asset deems reasonable for the purposes of its risk assessments. In the most optimistic case a figure will have been calculated during a risk assessment of the asset. More realistically, the figure used will be an informed judgement by the sponsoring manager.

4.4.1.5 Plant Availability

In practice plant availability is subdivided in various ways across the OGI and may overlap with terms like reliability, utilisation and planned versus unplanned downtime. These terms and their relationship to each other are frequently defined differently across various OpCo's. For the purposes of this study, the intent is not to review or conclude a specific definition; instead the intent is for the asset under review to state what the total downtime is that is expected for the year under review - not unlike the definition of a PSI.

Figure 4.3 displays a schematic indication of how the Deterministic Model is constructed. It consists of several groups which are closely related to the survey structure explained in Section 3.4.2. These are discussed in Section 4.4.2.

4.4.2 Component Parts of the Deterministic Model

This section will describe the component parts of the deterministic model. The parts and their interaction are shown in Figure 4.3 and subsequently described in terms of inputs, calculations and outputs respectively.

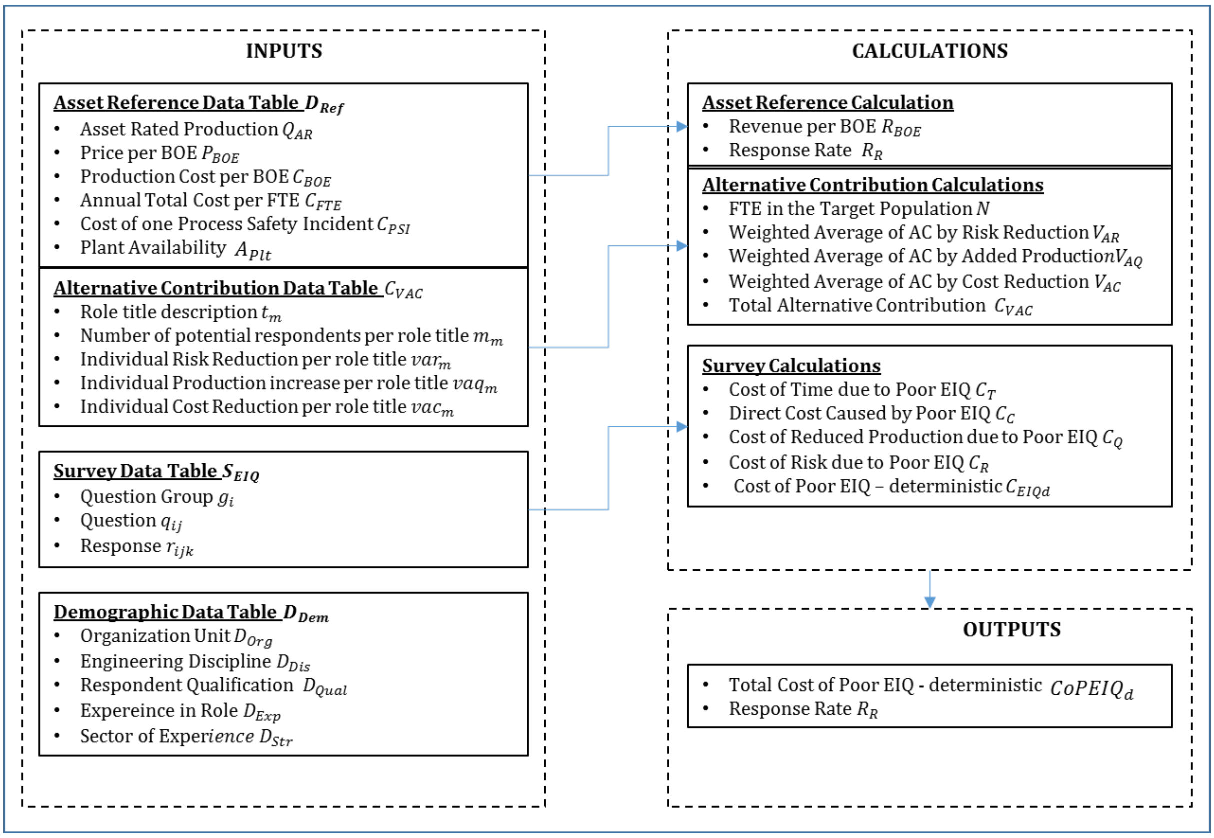

INPUTS
Asset Reference Data Table D_{Ref}
• Asset Rated Production Q_{AR}
• Price per BOE P_{BOE}
• Production Cost per BOE C_{BOE}
• Annual Total Cost per FTE C_{FTE}
• Cost of one Process Safety Incident C_{PSI}
• Plant Availability A_{Plt}
Alternative Contribution Data Table C_{VAC}
• Role title description t_m
• Number of potential respondents per role title m_m
• Individual Risk Reduction per role title var_m
• Individual Production increase per role title vaq_m
• Individual Cost Reduction per role title vac_m
Survey Data Table S_{EIQ}
• Question Group g_i
• Question q_{ij}
• Response r_{ijk}
Demographic Data Table D_{Dem}
• Organization Unit D_{Org}
• Engineering Discipline D_{Dis}
• Respondent Qualification D_{Qual}
• Expereince in Role D_{Exp}
• Sector of Experience D_{Str}
CALCULATIONS
Asset Reference Calculation
• Revenue per BOE R_{BOE}
• Response Rate R_R
Alternative Contribution Calculations
• FTE in the Target Population N
• Weighted Average of AC by Risk Reduction V_{AR}
• Weighted Average of AC by Added ProductionV_{AQ}
• Weighted Average of AC by Cost Reduction V_{AC}
• Total Alternative Contribution C_{VAC}
Survey Calculations
• Cost of Time due to Poor EIQ C_T
• Direct Cost Caused by Poor EIQ C_C
• Cost of Reduced Production due to Poor EIQ C_Q
• Cost of Risk due to Poor EIQ C_R
• Cost of Poor EIQ – deterministic C_{EIQd}
OUTPUTS
• Total Cost of Poor EIQ - deterministic $CoPEIQ_d$
• Response Rate R_R

Figure 4.3 - Deterministic Model Detail

4.4.2.1 Inputs to the Deterministic Model

The four input tables of the deterministic model are described in this section. The descriptions simultaneously constitute the variable declaration of the deterministic model:

Asset Reference Data Table

The Asset Reference Data Table D_{Ref} table is provided by the sponsoring manager prior to the survey being conducted at an asset. For the purposes of this study, D_{Ref}provides actual or assumed constant values used in the subsequent calculations. D_{Ref} consists of the following detail elements:

- Q_{AR} = Asset Rated Production in the UoM [boe/d]

- P_{BOE} = Price per BOE in force at the asset at the time of the survey in the UoM [USD/bbl]

- C_{BOE} = Production Cost per BOE in force at the asset at the time of the survey in the UoM [USD/bbl]

- C_{FTE} = Annualized, average total cost of employment for the target population of the survey in [CUR p.a.]. Include all costs like offshore/camp living, employee benefits or contractor gross rate.

- C_{PSI} = OpCo budget cost for one PSI [USD]

- A_{Plt} = Plant Availability for the year, expressed as a [%]

Alternative Contribution Table

The Alternative Contribution Table C_{VAC} is the second input table. The reasoning for Alternative Contribution has already been described in Section 3.3.3. The variables associated with this table are defined here as Table 4.3, this time containing variable which will be subsequently defined.

Table 4.3 - Alternative Contribution Variables

Role Title	No of FTE per role title in the target population	Risk Reduction	Production Increase	Cost Reduction
t_m	m_m	var_m	vaq_m	vac_m
Weighted Average	N	V_{AR}	V_{AQ}	V_{AC}

The variables in Table 4.3 are defined as follows:

- t_m = The role title description for m different roles as shown in Table 3.6
- m_m = The number of potential respondents in the target population in each role t_m
- var_m = Individual risk reduction per role t_m
- vaq_m = Individual production increase per role t_m
- vac_m = Individual cost reduction per role t_m
- N = Total number of FTE in the sample population. The intent of N is the number of FTE (staff or contractor) in routine service (i.e. special, short-term project excluded) in the roles in scope, namely engineering support in the asset, maintenance and TAR planning, maintenance management and corporate planners involved in asset evaluation or performance. In practice N would be the total number of potential respondents nominated by the asset.

In passing it may be noted that the input variables in the preceding tables may be deemed variables, given the complexities associated with their calculation. For the purposes of this study, however, these variables are deemed constants that serve as inputs into the calculation of the core value, namely the cost of poor EIQ. Testing the variation of these inputs on CoPEIQ might be the subject of a sensitivity analysis in future work.

Survey Data Table

The Survey Data Table is the third input table. The composition of the survey table has implications for the subsequent calculations and reporting. For this reason, some detail in this regard is warranted.

The survey table S_{EIQ} may be seen as a collection of four survey question groups g_i where $i \in \{C, T, R, Q\}$, namely Cost, Time, Risk and Production. Each question group g_i consists of several questions q_j where $j \in \mathbb{R}$. The survey table therefore contains four different types of questions. In the next section it will be shown that each question group g_i uses a different formula to calculate its contribution to CoPEIQ. Each question q_j, in turn, will receive several responses $r_k, k \in \mathbb{R}$.

The Survey Data Table can also be described as follows:

- Survey table S_{EIQ} contains four:
  - Survey question groups g_i; $i \in \{C, T, R, Q\}$; each containing several:
    - Questions q_{ij}; $j \in \mathbb{R}$; each attracting several:
      - Individual responses r_{ijk} where $k \in \mathbb{R}$.

Accordingly, Table 3.12 is re-written as Table 4.4 to demonstrate the variables:

Table 4.4 - Survey Question Variables

Question Group g_i	Question No q_{ij}	Question	UoM	Responses r_{ijk}
$i = T$	j=1	Please estimate the Additional Time spent looking for Engineering Information (EI)	[hrs/day]	$k = 1,2,3,\ldots$
T	2	Please estimate the Additional Time spent verifying or re-entering EI	[hrs/day]	
T	3	Please estimate the Additional Time repeating processes	[hrs/day]	
T	4	Please estimate the Time spent to validate/prepare corporate KPIs due to poor EIQ	[hrs/day]	
T	5	Please estimate the Time spent clarifying misunderstandings surrounding EI & EIQ	[hrs/day]	
C	1	Please estimate the annual Financial impact of misunderstanding	[USD]	
T	6	Please estimate the Additional Time spent reviewing EI Standards	[hrs/day]	
T	7	Please estimate the Time spent creating and maintaining unofficial databases	[hrs/day]	
T	8	Please estimate the Time spent resolving EI conflicts between databases	[[hrs/day]	
T	9	Please estimate the Additional time spent optimizing the budget and production plan	[hrs/day]	
Q	1	Please estimate the Production loss due to the wrong data being reported	[%]	
T	10	Please estimate the Additional time spent preparing Regulatory Reports	[hrs/day]	
T	11	Please estimate the Additional time spent responding to Regulatory Queries	[hrs/day]	
C	2	Please estimate the Likely Cost of a regulatory penalty due to Poor EIQ	[USD]	
T	12	Please estimate the additional Time required to prepare Decision Review Board Support information	[hrs/day]	

Question Group g_i	Question No q_{ij}	Question	UoM	Responses r_{ijk}
T	13	Please estimate the Time spent recreating EI not delivered from Projects	[hrs/day]	
T	14	Please estimate the Time spent obtaining lost EI from Vendors	[hrs/day]	
T	15	Please estimate the Idle/non-productive Engineering resource time due to EIQ	[hrs/day]	
T	16	Please estimate the Additional Time spent updating EI in the MoC Process	[hrs/day]	
Q	2	Please estimate the Reduced production due to poor design	[%]	
T	17	Please estimate the Additional Time accepting/verifying/approving EI from Projects	[hrs/day]	
T	18	Please estimate the Additional Time spent approving mapping EI between systems	[hrs/day]	
C	3	Please estimate the Cost of redundant scrapped material	[USD]	
C	4	Please estimate the Cost of redundant procurement	[USD]	
C	5	Please estimate the Cost of redundant construction	[USD]	
C	6	Please estimate the Likely Cost of professional error due to EIQ	[USD]	
T	19	Please estimate the Additional Time (re) creating Maintenance PMs	[hrs/day]	
T	20	Please estimate the Idle/non-productive time in Maintenance resource time due to EIQ	[hrs/day]	
T	21	Please estimate the Additional Time to call of contracts/ mobilize vendors	[hrs/day]	
T	22	Please estimate the Additional Time spent by TAR team (re) creating Work Packages	[hrs/day]	
T	23	Please estimate the Additional Time spent optimizing maintenance resource	[hrs/day]	
T	24	Please estimate the Additional Time required to optimize asset operation	[hrs/day]	
Q	3	Please estimate the Increase in Availability if EIQ was better	[%]	
C	7	Please estimate the Cost of performing unnecessary inspection/maintenance due to poor EIQ	[USD]	
C	8	Please estimate the Cost of redundant spares in warehouse	[USD]	
C	9	Please estimate the Cost of spares expediting ("hot shot costs") & management	[USD]	
Q	4	Please estimate the Production loss due to Asset sub-optimization	[%]	
R	1	Please estimate the Likelihood of a Process Safety Incident due to poor EIQ	[USD]	
T	25	Please estimate the Additional Time spent assessing asset performance	[hrs/day]	
T	26	Please estimate the Time spent identifying conflicts between systems	[hrs/day]	
T	27	Please estimate the Additional Time verifying all stakeholders have the same EI	[hrs/day]	
T	28	Please estimate the Additional Time spent mapping between systems	[hrs/day]	

Table 4.4 shows that the four question groups g_i ; $i \in \{C, T, R, Q\}$ contain 28, 4, 9 and 1 q_{ij}questions respectively, or:

- $g_T \in \{1,2,3, \dots, 28\}$
- $g_Q \in \{1,2,3, 4\}$
- $g_C \in \{1,2,3, \dots, 9\}$
- $g_R \in \{1\}$

It follows that the questions q_{ij} are the independent variables and the responses rij_k are the dependent variables.

Demographic Data Table

The D_{Dem} table is collected in the survey as shown in Table 3.11. No variables in the D_{Dem} table will be used in the deterministic model; they are collected for future secondary research, as discussed in Section 3.2.1. D_{Dem} contains the following variables:

- D_{Org} = Organisational unit, or "team" or "department" where the respondent is deployed at the time of the survey, or, if the respondent had been deployed there for less than three months, the previous unit. This time restriction is included to obtain the most realistic estimate from the respondent. This variable is used to drive the filtered questions shown in Table 3.11.

- D_{Dis} = Engineering Discipline. D_{Dis} is collected to serve as the basis for an alternative view during the reporting of results. Section 5.1.2 explains the need for reporting results in more scrutiny and from different viewpoints; D_{Dis} is used for one such analysis.

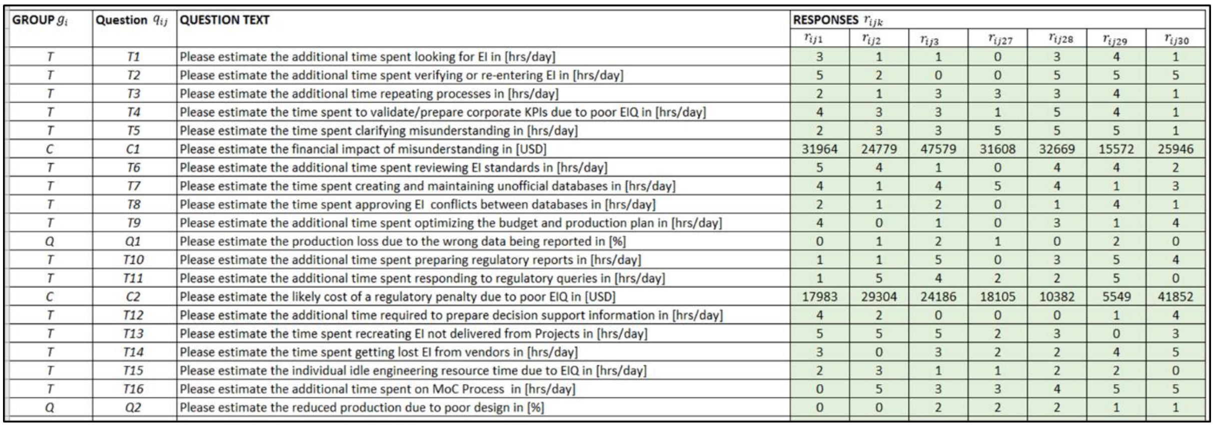
- D_{Qual} = Highest academic qualification of the respondent. D_{Qual} is intended to collect data over many surveys to seek further insight into the problem of EIQ by secondary research.

- D_{Exp} = Experience in the current role or, if less than three months in role, in previous role in [yrs]. This variable is likewise intended for secondary research and specifically to test for what may be termed the Data Island Hypothesis, which is discussed in Section 7.3.

- D_{Str} = Sector where Respondent has spent most of the Respondent's working career. Like D_{Qual} , it is intended to collect data over many surveys to seek further insight into the problem of EIQ by secondary research.

Input Data

Since the development of the model was done in a laboratory environment, a sample set of "laboratory" data was generated using Excel's NORM.INV function. For this sample set, 30 respondents were assumed and orders of magnitude for the mean and standard deviation of the Excel command were assumed, based on the insight gained during the validation structured interviews that have been described in Section 3.3. A screen shot of the 'Survey Data" worksheet is shown in Figure 4.4. The columns containing respondents 4 to 26 have been hidden for the sake of visual expediency. For the same reason, the figure only displays only the first 20 questions.



GROUP g_i	Question q_{ij}	QUESTION TEXT	RESPONSES r_{ijk}						
			r_{ij1}	r_{ij2}	r_{ij3}	r_{ij27}	r_{ij28}	r_{ij29}	r_{ij30}
T	*T1*	Please estimate the additional time spent looking for EI in [hrs/day]	3	1	1	0	3	4	1
T	*T2*	Please estimate the additional time spent verifying or re-entering EI in [hrs/day]	5	2	0	0	5	5	5
T	*T3*	Please estimate the additional time repeating processes in [hrs/day]	2	1	3	3	3	4	1
T	*T4*	Please estimate the time spent to validate/prepare corporate KPIs due to poor EIQ in [hrs/day]	4	3	3	1	5	4	1
T	*T5*	Please estimate the time spent clarifying misunderstanding in [hrs/day]	2	3	3	5	5	5	1
C	*C1*	Please estimate the financial impact of misunderstanding in [USD]	31964	24779	47579	31608	32669	15572	25946
T	*T6*	Please estimate the additional time spent reviewing EI standards in [hrs/day]	5	4	1	0	4	4	2
T	*T7*	Please estimate the time spent creating and maintaining unofficial databases in [hrs/day]	4	1	4	5	4	1	3
T	*T8*	Please estimate the time spent approving EI conflicts between databases in [hrs/day]	2	1	2	0	1	4	1
T	*T9*	Please estimate the additional time spent optimizing the budget and production plan in [hrs/day]	4	0	1	0	3	1	4
Q	*Q1*	Please estimate the production loss due to the wrong data being reported in [%]	0	1	2	1	0	2	0
T	*T10*	Please estimate the additional time spent preparing regulatory reports in [hrs/day]	1	1	5	0	3	5	4
T	*T11*	Please estimate the additional time spent responding to regulatory queries in [hrs/day]	1	5	4	2	2	5	0
C	*C2*	Please estimate the likely cost of a regulatory penalty due to poor EIQ in [USD]	17983	29304	24186	18105	10382	5549	41852
T	*T12*	Please estimate the additional time required to prepare decision support information in [hrs/day]	4	2	0	0	0	1	4
T	*T13*	Please estimate the time spent recreating EI not delivered from Projects in [hrs/day]	5	5	5	2	3	0	3
T	*T14*	Please estimate the time spent getting lost EI from vendors in [hrs/day]	3	0	3	2	2	4	5
T	*T15*	Please estimate the individual idle engineering resource time due to EIQ in [hrs/day]	2	3	1	1	2	2	0
T	*T16*	Please estimate the additional time spent on MoC Process in [hrs/day]	0	5	3	3	4	5	5
Q	*Q2*	Please estimate the reduced production due to poor design in [%]	0	0	2	2	2	1	1

Figure 4.4 - Laboratory Survey Data

The laboratory data is the final input to the deterministic model. Its calculations are described in the next section.

4.4.2.2 Calculations of the Deterministic Model

From the inputs defined in the previous section and again with reference to Figure 4.3, the deterministic model contains the calculations per individual table as shown below.

Asset Reference Data

Asset Reference Data D_{Ref} yields only one calculation:

- $R_{BOE} = P_{BOE} - C_{BOE}$

Where

- R_{BOE} = Revenue per BOE in force at the asset at the time of the survey in the UoM [USD/bbl], or, in other words, the revenue per barrel oil equivalent is the price minus the cost to produce it.

- R_R = Response Rate. This measure is used to determine the percentage of potential to actual respondents.

Alternative Contribution

Alternative Contribution C_{VAC} yields the following calculations:

- $N = \sum m_m$

- $V_{AR} = \sum(m_m . var_m) / N$

- $V_{AC} = \sum(m_m . vac_m) / N$

- $V_{AQ} = (\sum(m_m . vaq_m) \,/\, N) \,.Q_{AR}. \; 365$
- $C_{VAC} \;= V_{AR} + V_{AQ} + V_{AC}$

Where

- V_{AR} = Weighted Average of Alternative Contribution by Risk Reduction [USD]
- V_{AQ} = Weighted Average of Alternative Contribution by Added Production [USD]
- V_{AC} = Weighted Average of Alternative Contribution by Cost Reduction [USD]
- C_{VAC} = Total Alternative Contribution [USD]

Put another, the total alternative contribution is the weighted average of the constituent alternative contributions for risk reduction, production increase and cost reduction respectively.

The Survey Table

The Survey Table calculations require the definition of several additional variables:

- n = number of responses r to individual questions in the Survey. (For the deterministic model, n is used to calculate averages. This requirement will be unnecessary for the stochastic model, since the responses will be used to generate input distributions. Details are given in Section 4.5.2.)
- IE_G = Impact Element Group, of which:

  IE_T = Average Response in the Subgroup Impact Element 'Time", in the UoM [hrs/day]

  IE_C = Average Response in the Subgroup Impact Element "Cost" in the UoM [USD]

  IE_Q = Average Response in the Subgroup Impact Element "Production" in the UoM [% of APR]

  IE_R = Average Response in the Subgroup Impact Element "Risk" in the UoM [USD].

- C_{EIQd} = Cost of Poor EIQ - deterministic, of which:

  C_T = Cost of Time due to poor EIQ in [USD]

  C_C = Cost directly caused by poor EIQ in [USD]

  C_Q = Cost of reduced Production due to poor EIQ in [USD]

  C_R = Cost of Risk due to poor EIQ in [USD]

- $CoPEIQ_d$ = Total Cost of Poor EIQ - deterministic

One approach to the calculations required for the survey table is shown below. The same result may be achieved several ways; this approach has been selected to demonstrate the principle as visibly as possible.

Then:

- $IE_T = (\sum_{r=1}^{k} \sum_{j=1}^{28} r_{Tjk})/28\,k$

  And similarly:

- $IE_Q = (\sum_{r=1}^{k} \sum_{j=1}^{4} r_{Qjk})\,/4k$

- $IE_C = (\sum_{r=1}^{k} \sum_{j=1}^{9} C)/9k$

- $IE_R = (\sum_{r=1}^{k} r_{Rjk})/1k$

Similarly:

- C_T = The sum of the average responses in the Impact Elements Group "Time", normalized to an annual cost, or:

$$C_T = \frac{IE_T.N.C_{FTE}}{10}$$

  (assuming 10 hours work per day).

- C_c = The sum of the average responses in the IE group "Cost", or

$$C_C = IE_C$$

- C_Q = The sum of the average responses in the Impact Elements Group "Production", normalized to an annual cost, or

$$C_Q = \frac{IE_Q}{100}.Q_{AR}.\ A_{Plt}.R_{BOE}.365.$$

- C_R = The sum of the average responses in the Impact Elements Group "Risk", normalized to an annual cost, or

$$C_R = IE_{Rr}$$

- $C_{EIQd} = C_T + C_C + C_Q + C_R$

- $CoPEIQ_d = C_{EIQd} + C_{VAC}$

The calculations shown in this section form the basis of the deterministic model and are included in the stochastic model discussed in Section 4.5.

4.4.2.3 Outputs of the Deterministic Model.

Based on the calculation in Section 4.4.3 and again with reference to Figure 4.3, the following outputs are expected to be normally reported from the deterministic model:

- $CoPEIQ_d$
- Response Rate R_R

Consistent with the conclusions drawn in Sections 4.4, additional outputs in graphical form are required. These are discussed in Chapter 5.

4.4.3 Demonstrating the Deterministic Model

The preceding description of the deterministic model enables the construction of the model in a relatively simple spreadsheet. Such a spreadsheet was developed and populated with hypothetical data of typical orders of magnitude to illustrate the working of the formulae. Subsequent researchers are encouraged to replicate these calculations; to this end two views of the illustrative spreadsheet are included in this section.

To assist future users of this spreadsheet in following the logic, worksheets title tabs, cells and data tables have been coloured to broadly follow the following conventions:

- Green - inputs
- Yellow - interim steps
- Red or Orange - outputs

This convention is evident in the following figures. It may be noted that output data of one step will often become input data to the next.

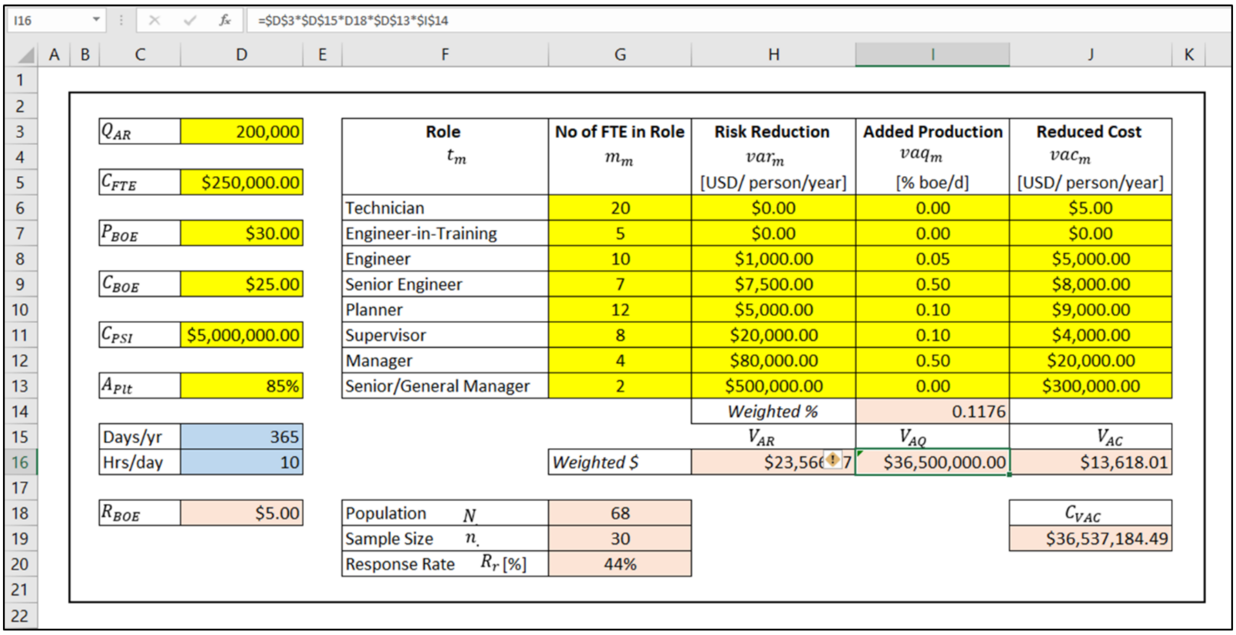
Since the spreadsheet is used to develop the model, inputs have been populated with amounts that are hypothetical, but typical for the type of asset where this study was undertaken, as had become evident during the validation structured interviews described in Section 3.3.

Following the logic shown in Figure 4.2, the spreadsheet contains several worksheets. The first of these relate to the deterministic model:

- The worksheet entitled “Dref_Calc” contains the Asset Reference data and is an extension of Table 3.8.
- The survey questions shown in Table 3.12 are used to show collected data and subsequent calculations in the worksheet called “CoPEIQ_Det”.

Two screen shots showing the salient points of the spreadsheet are included as figures.

In Figure 4.5, the Asset Reference Data D_{Ref} is shown in the worksheet “Dref_Calc”. The cursor is on cell I16 to show the formula used to calculate the cost of alternative value for added production V_{AQ} . These are typical figures for the class of asset where the structured interviews were done .



I16 =$D$3*$D$15*D18*$D$13*$I$14

Q_{AR}	200,000
C_{FTE}	$250,000.00
P_{BOE}	$30.00
C_{BOE}	$25.00
C_{PSI}	$5,000,000.00
A_{Plt}	85%
Days/yr	365
Hrs/day	10
R_{BOE}	$5.00

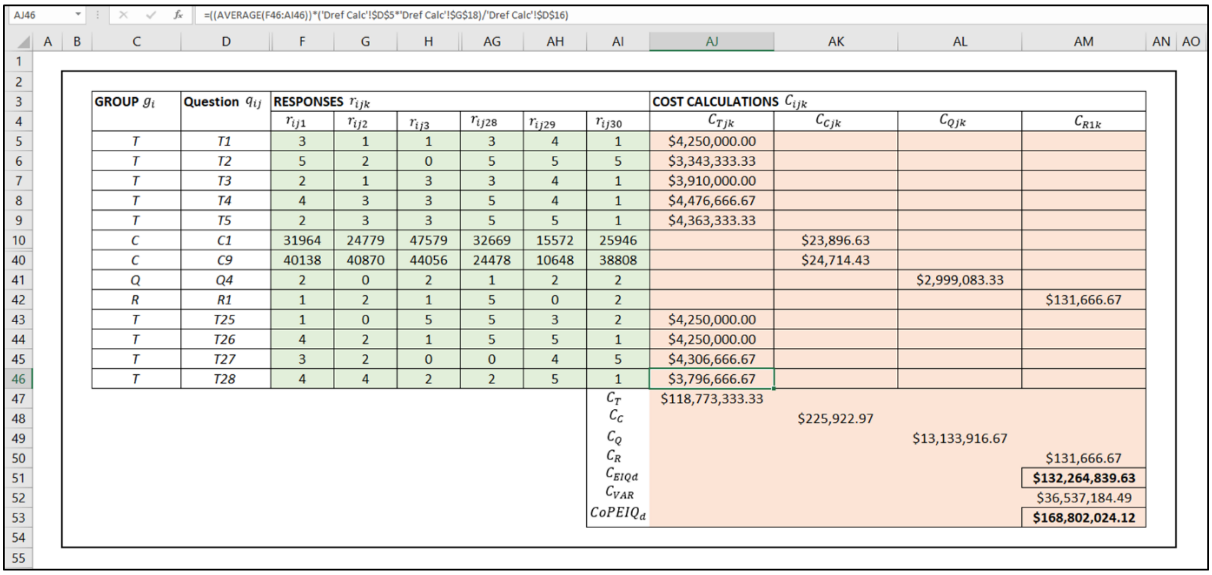
Role t_m	No of FTE in Role m_m	Risk Reduction var_m [USD/ person/year]	Added Production vaq_m [% boe/d]	Reduced Cost vac_m [USD/ person/year]
Technician	20	$0.00	0.00	$5.00
Engineer-in-Training	5	$0.00	0.00	$0.00
Engineer	10	$1,000.00	0.05	$5,000.00
Senior Engineer	7	$7,500.00	0.50	$8,000.00
Planner	12	$5,000.00	0.10	$9,000.00
Supervisor	8	$20,000.00	0.10	$4,000.00
Manager	4	$80,000.00	0.50	$20,000.00
Senior/General Manager	2	$500,000.00	0.00	$300,000.00
		Weighted %	0.1176	
		V_{AR}	V_{AQ}	V_{AC}
	Weighted $	$23,56[illegible]7	$36,500,000.00	$13,618.01

Population	N	68
Sample Size	n	30
Response Rate	R_r [%]	44%

C_{VAC}
$36,537,184.49

Figure 4.5 - Asset Reference Data

Figure 4.6 demonstrates the calculation of C_{T28k} from the worksheet CoPEIQ_Det, with input data (Columns F to AI, with columns I to AF hidden). The cursor is on cell AJ46 to demonstrate the use of the formula to calculate C_T . The intermediate result C_{EIQd} and final result $CoPEIQ_d$ are also shown.



AJ46 =((AVERAGE(F46:AI46))*('Dref Calc'!$D$5*'Dref Calc'!$G$18)/'Dref Calc'!$D$16)

GROUP g_i	Question q_{ij}	RESPONSES r_{ijk}						COST CALCULATIONS C_{ijk}			
		r_{ij1}	r_{ij2}	r_{ij3}	r_{ij28}	r_{ij29}	r_{ij30}	C_{Tjk}	C_{Cjk}	C_{Qjk}	C_{R1k}
T	*T1*	3	1	1	3	4	1	$4,250,000.00			
T	*T2*	5	2	0	5	5	5	$3,343,333.33			
T	*T3*	2	1	3	3	4	1	$3,910,000.00			
T	*T4*	4	3	3	5	4	1	$4,476,666.67			
T	*T5*	2	3	3	5	5	1	$4,363,333.33			
C	*C1*	31964	24779	47579	32669	15572	25946		$23,896.63		
C	*C9*	40138	40870	44056	24478	10648	38808		$24,714.43		
Q	*Q4*	2	0	2	1	2	2			$2,999,083.33	
R	*R1*	1	2	1	5	0	2				$131,666.67
T	*T25*	1	0	5	5	3	2	$4,250,000.00			
T	*T26*	4	2	1	5	5	1	$4,250,000.00			
T	*T27*	3	2	0	0	4	5	$4,306,666.67			
T	*T28*	4	4	2	2	5	1	$3,796,666.67			
							C_T	$118,773,333.33			
							C_C		$225,922.97		
							C_Q			$13,133,916.67	
							C_R				$131,666.67
							C_{EIQd}				**$132,264,839.63**
							C_{VAR}				$36,537,184.49
							$CoPEIQ_d$				**$168,802,024.12**

Figure 4.6 - Calculating the Sum of Responses for Time C_T

After defining some introductory concepts, this section has developed the deterministic model to calculate the Cost of poor EIQ deterministically. Variables have been defined, formulae stated and a spreadsheet populated with the typical figures to demonstrate the deterministic model. This section now forms the basis for the stochastic model described in Section 4.5.

4.5 Develop the Stochastic Model

In the preceding chapters a survey has been developed, it was determined how results should be presented and how to calculate CoPEIQ deterministically. The inherent limitations of this research were discussed in Section 1.3.6. These limitations, combined with the inherent variability of inputs in the dynamic environment of an OpCo indicate that a deterministic calculation of CoPEIQ is sub-optimal. A brief review of the literature regarding stochastic versus deterministic models yields a number of insights; these are briefly reviewed before proceeding with the development of the stochastic model.

Haney (2016) divides quantitative models into deterministic and stochastic models. He submits that:

- Deterministic models assume that there is certainty about the inputs of the model and that the solution is done on analytical grounds and yields one answer. Some insight about the variability of inputs is possible by means of sensitivity analysis, but this becomes problematic when many inputs vary simultaneously.

- Stochastic models, by contrast, deal with uncertain inputs. The models apply numerical techniques and yields a range of possible answers. This range is characterized in some form and presented to the decision-maker to provide insight into the decision.

Hines & Montgomery (1980) characterizes stochastic processes as being related to observations related to time, and as being physical processes that are "controlled by random mechanisms". They define stochastic processes as "sequences of random variables $\{X_t\}$" where $t \in T$ is a time of sequence index. X_t may be discrete or continuous.

From these descriptions a stochastic model is more appropriate for this research. This section therefore proceeds with an overview of statistical methods appropriate for the characteristics of this study. These characteristics are: many stochastic inputs, a reasonably complex model and a target audience familiar with stochastic results. These characteristics are particularly suited for MCS, as will be shown in Section 4.5.2. The next section, however, will provide an overview of statistical techniques. This is done to provide some assurance that no obvious methodology candidates are eliminated.

4.5.1 Literature Review: Statistical Methods

Fess (1991) contends that there is not one statistical technique that provides answers to all questions, adding that "even experts sometimes disagree". It follows that the most appropriate statistical instruments can be selected only by means of a literature review and some form of selection criteria.

This section contains a general discussion taken from the literature of several statistical and quantitative techniques. The intent is a 'sanity check' to confirm the contention that a Monte Carlo method is appropriate for this research. It is not contended that the list in this section is comprehensive; instead it provides a cursory review of applicability.

It is reiterated that this research does not intend to develop a bespoke analysis model; instead the intent is to select an appropriate model and apply the collected data to it, in direct service of an important business problem.

The section will begin with a discussion of general statistical concepts and proceed to more advanced stochastic methods.

Statistics is defined by Hamburg (1974) as "a body of theory and methodology for drawing inferences and making decisions under conditions of uncertainty". They support a "logical, objective" approach to decision-making. Given, therefore, what has been learned in Section 4.1, a statistical approach is well suited.

Hamburg (1974) distinguished between descriptive and inferential statistics. This study is clearly in the latter category, where the phenomenon under study is not just described, but specific characteristics about it are inferred and used as rational decision support in a real-world setting.

4.5.1.1 General Statistical Principles Applicable to this Study

This sub-section specifically interprets selected foundational statistical concepts that are relevant for this study. The remarks refer throughout to the definitions and interpretations of Hamburg (1974). If additional interpretations are included, they will be specified.

In Section 4.2 it was concluded that "results should make it possible for decision-makers to understand the structure and magnitude of the results' uncertainty. This means in practice that the centricity, dispersion and confidence level should be reported". Earlier in this chapter, it was concluded that results should be presented in a stochastic form.

Probability

The distinction between classical, relative and subjective probabilities is deemed inclusive in this study, since the perception of respondents is being tested.

For this study, an Event is defined as "a circumstance when an IE 'activates' to the extent to have an effect on CoPEIQ". Events are assumed to be not conditional, since it is unlikely that an event where an IE is activated will result in that specific EIQ element to be repaired; it is just not in the nature of dealing with day-to-day OpCo realities to pause every time an EI instance is discovered. In the absence of a deliberate EIQ initiative, that IE will remain latent until the next time it is called upon – hence "with replacement".

Probability Distributions

Thomopoulos (2013) describes that variables that can only take on a specified list of values are deemed discrete, whereas continuous variables can take on any value in a specified range. Rubinstein & Kroese (2017) characterize this difference by means of the state variable: it either changes "continuously over time" or "instantaneously at discrete points in time". Based on these definitions, only continuous distributions are deemed appropriate for this study.

The field data collected for this study may potentially be distributed by any of the accepted continuous distributions and potentially by a different distribution for every question and every survey. Since the intent of the work is to provide a defensible range of results to the central question of CoPEIQ, however, the cost and complexity of formal distribution analysis is not deemed justified. Coupled with the expected sample size, the data analysed for this study is assumed to be normally distributed.

Sample Size

If the population is normally distributed, the sampling distribution is in itself normally distributed. Its standard deviation is equal to the population standard deviation divided by the square root of the sample size. This means that increasing the sample size will have a diminishing return with respect to its value in estimating the population mean.

If, conceivably, a population is not normally distributed, the sample distribution may be deemed "approximately normal" for almost all distributions, provided "sufficiently large samples" are available, based on the Central Limit Theorem.

Given the advantages of assuming a normal sampling distribution, the questions beg what constitutes "sufficiently large". Hamburg (1974) puts it a "about 10 to 20". Hines & Montgomery (1980) feel that this is "not an easy question to answer" since it depends on the characteristics of the population and what is deemed acceptable.

A stratified sampling strategy is applied for mutually exclusive subgroups. Since the population in this study may not be mutually exclusive (i.e. an individual may be deployed as both a discipline engineer and a planner), a stratified strategy is avoided.

Based on the guidance from these sources, a minimum sample of 20 respondents was requested for the field study and on that basis sample distribution may be assumed to be normal. The results will be shown in Chapter 6.

Confidence Levels

Hamburg (1974) defines an interval estimate as "a statement of two values between which we have some confidence the parameter lies". For the current study, the sample size for each survey instance will be unknown until after survey results have been collected and are, in any event, a given with no opportunity for variation. The appropriate approach is therefore in three parts:

- Determine the sample size after survey results have been received.
- Calculate the confidence level by normal distribution parameters.
- Report the results with the confidence indicated by the data.

This sub-section interpreted some fundamental statistical concepts and how they relate to the current work. The next section takes a similar general approach to what may be termed "advanced numerical methods".

4.5.1.2 Advanced Numerical Methods

In this sub-section, an overview of numerical methods is given, with the objective of ensuring that no obvious candidate for this study is missed. The discussion does not purport to be comprehensive or complete; rather it is intended to be a "quick check" before proceeding to the MCS.

Hines & Montgomery (1980) describe stochastic processes as processes related to time within physical processes where random mechanisms control the process.

In the absence of a formal classification system, the subjects that were reviewed will simply be listed in alphabetical order. Each method will end with short discussion about the applicability to this research.

The following techniques have been reviewed:

- Artificial Intelligence
- Bayesian Decision Theory
- Decision Theory
- Expert Systems
- Fuzzy Logic
- Neural Networks
- Machine Learning
- Markov Chains

Artificial Intelligence

Artificial Intelligence (AI) is defined in computer science as studying what is termed "intelligence agents". These are (mostly computerized) devices that respond to their environments with actions intended to achieve objectives with the best possible chance of success. (Russel & Norvig (2003). These actions may be reasoning, learning, collecting knowledge, developing perception and the ability to manipulate objects. This thinking represents a synergy of many disciplines, amongst others mathematics, psychology, philosophy and computer science.

This wide definition implies that AI is often used as an umbrella term, or parent discipline, for a variety of more specific techniques and terms like neural networks.

Russel & Norvig (2003) and McCorduck (2004) describe several innovations in statistics in AI, or be it that these are not without criticism. (Katz (2012))

Specifically, for problems of uncertainty, tools have been developed from probability theory and stochastics. These include Bayesian networks for reasoning, learning and perception. (Russel & Norvig (2003))

Buchanan (2006) published "A (Very) Brief History of Artificial Intelligence (AI)" in 2006. He submits that knowledge representation and inference are areas where much learning is still required. Knowledge management, in turn, may be defined as using knowledge to the maximum effect to achieve organisational objectives.

Russel & Norvig (2009) contend that AI is "relevant to any intellectual task". For example, it has been used for predicting judicial decisions (Nikolaus et al. (2016)), but there is concern about the ultimate philosophical and ethical implications of AI, even to the point of existential risk (Hawking 2015).

AI is too wide a general a concept to determine specific applicability for this research, except to say that the stochastic requirement of the work implies the use of some form or AI. It may perhaps be argued that a comprehensive database of survey results may be an input into a future knowledge database. Testing this idea is left to future research.

Bayesian Decision Theory

Hamburg et al. (1974) explain Bayes' Theorem as "a means of revising prior probabilities of events based on the observations of additional information". For the simplified example of two events A_1, A_2 given an additional constraint B, the probability P of A_1, given B is

$$P(A_1|B) = \frac{P(A_1)(P(B|A_1)}{P(A_1)(P(B|A_1) + P(A_2)(P(B|A_2)}$$

This way, new experimental evidence can weigh existing probabilities.

This approach is being used extensively in business applications, where existing judgements and perceptions serve as initial data and are appended and updated as new data, whether empirical or otherwise, is added.

This approach may very well be of value for subsequent research into the CoPEIQ phenomenon. Since CoPEIQ is a new concept of which relatively little is known, as has been shown in Section 1.3.3, this study intends to collect the initial "intuitions" and "subjective judgements" regarding CoPEIQ. Once the subject is understood better, Bayesian Decision Theory is a likely candidate to append the understanding of the subject as more information is added. At the time of this research it is therefore not (yet) relevant for the problem addressed in this study.

Decision Functions

Hines & Montgomery (1980) define a decision function d in the decisions space a as

$$a = d(x_1, x_2, x_3, \dots x_n)$$

where x_i are the sample observations. They proceed to define a loss function $l(a, \theta)$ where θ is an unknown parameter being sought. A risk function $R(d, \theta)$ is then defined and is solved so that the optimal decision is one where R is a minimum.

CoPEIQ might be theoretically be modelled as a Loss function. However, since the typical level of investment sought for data remediation projects is unlikely to justify decision modelling at this level and decisions are frequently made on non-rational grounds (Section 4.1), this logic is not pursued further in this study.

Expert Systems

Jackson (1998) defines an expert system as the emulation by a computer of the ability by a human expert to make decisions. It is deemed to have originated at the Stanford Heuristic Programming Project under the leadership of the oft-quoted Edward Feigenbaum. Such a system consists of two parts, an inference engine and a knowledge base. The knowledge base "represents facts about the world" and the inference engine is an "automated reasoning system". Their relative value is frequently quoted to have been stated by Feigenbaum: "intelligent systems derive their power from the knowledge they possess rather than from the … inference schemes they use". They are in use by many leading business software systems and are frequently associated with business process automation and integration. They are frequently used in the context of Bayesian Networks, where existing data is supplemented by new knowledge. Constantinou et al. (2016) proposed such a method.

Hayes-Roth et al. (1983) developed 10 categories of application for expert systems. These were Interpretation, Prediction, Diagnosis, Design, Planning, Monitoring, Debugging, Repair, Instruction and Control. A major disadvantage in the literature is the problem of getting to the expert knowledge needed to populate the knowledge database.

The only category which may be of potential value for the current research is Prediction, or, as stated by Hayes-Roth et al. (1983), "Inferring likely consequences of given situations". This option is, however, discounted on the basis that there is adequate anecdotal "expert opinions" in the literature, which does not address the fundamental requirement of this research as stated in Section 1.3. To capture such opinion in an expert system is not likely to address the problem.

Fuzzy Logic

In the pioneering paper "Fuzzy Algorithms" Zadeh (1968) explains that problems for which precise algorithmic solutions are applicable are in fact "quite limited". Realistic problems are complex and either not able to be solved by algorithms, or may be "computationally infeasible". As a result, he introduces "fuzzy algorithms", where uncertainty of exact independent variables is dealt with by means of grades of memberships of intervals. He compares fuzzy algorithms with a heuristic programme by concluding that "a heuristic programme is a nonfuzzy approximation, expressed in computer language, to a fuzzy algorithm. He proceeds to predict that the notion of fuzzy algorithms may prove useful to areas such as control, pattern recognition, AI and decision-making involving uncertain data.

Nearly forty years later this prediction appears to be coming true. Novak (2005) discusses the applicability of "fuzzy sets" and contrasts it to classical set theory. He starts by distinguishing between "uncertainty" and "vagueness". Uncertainty is concerned with groupings of objects which have not (yet) actualized, while vagueness relates to the actual existence of groupings of which the boundaries are not clearly defined or the membership of the grouping at the boundaries is not clearly defined. The distinction is then between actuality and potentiality. Novak argues that the maturing fuzzy theory is appropriate for the vagueness case. He uses the example that classical set theory is "hardly" applicable to an area such as evaluating linguistic expressions, and proceeds to state that fuzzy sets are useful for challenges like AI, robotics and computer science. Conversely, areas of uncertainty (or potentiality) are appropriately addressed by mathematical models, "especially probability theory".

According to the analysis of Van den Honert (2014), fuzzy logic is powerful "to characterize vagueness and fuzziness in humanistic systems. Fuzzy systems are primarily used in deductive reasoning ".

The problem of determining CoPEIQ is in the "uncertainty" category and therefore, according to Novak, better dealt with using a stochastic approach. It may be argued in concept that the survey results are in the "vagueness" category, however the required result accuracy is not high enough (Section 4.2) to warrant the investment in fuzzy models for survey results.

Neural Networks

From within the context of accounting methodologies for exploration, Spear & Leis (1997) demonstrate that neural networks significantly outperform traditional statistical methods in terms of overall error rates. Neural networks intend to emulate the human brain for problem solving. This is done by using observations to "infer a function". The utility of neural networks is therefore in complex data sets and tasks. The following are broad categories of tasks where neural networks are applied: function approximation, time series prediction, regression analysis, pattern recognition, classification, novelty detection, decision-making of sequential nature, data clustering and filtering, robotics and prosthesis manipulations and advanced control.

According to Van den Honert (2014), Neural Networks "are tremendously powerful for tasks such as information processing, learning and adaption".

The central problem if CoPEIQ is to determine the cost impact of poor EIQ in a specific context, at a specific time, but considering a wide a variable range of inputs. There is no requirement for the current study for sequential work, causality or system learning. As the awareness of CoPEIQ increases in the OGI and the requirement to model dominant specific processes or scenario's, there may be case to warrant a Neural Network model. It is, however, for the present work, discounted.

Machine Learning

The original intent of machine learning was to develop systems with enough intellect to perform complex asks, i.e. getting and using knowledge in the form of "rich relational structures". Gaining this knowledge was achieved through heuristic and multi-step processes. There are a few changes in the science, notably the arrival of pattern recognition and mathematical modelling. (Langley (2011)). Mannila (1996) suggests that is it is closely related to computational statistics and mathematical optimization and that machine learning attempts to develop ways to learn things that humans find difficult.

Machine learning may very well be of value for the automation of some repetitive tasks related to CoPEIQ, such as, for instance, validating certain elements of EI against specific formatting requirements. Given the research objectives of this current work, however, machine learning is too detailed.

Markov Chains

Markov chains are deemed a "special type" of stochastic process and are said to be useful where probabilities are required for several states that related to each other in terms of time. Discrete-time Markov chains are used in the case where time is discrete. A "Markovian property" exists if the probability of an event at time $t+1$ (given only the outcome at time t) is equal to the same probability at time $t+1$ (given the entire history of the system). In simpler form yet, the probability of an event at time $t+1$ does not depend on history prior to time t. (Hines & Montgomery (1980)).

The analysis of van den Honert (2014) concluded that Markov chains are applicable for the classification of patterns, "where each pattern is viewed as a sequence of states".

The notion of time-dependency is not a factor in the current study, since the CoPEIQ measurement is concerned with the state in an OpCo at one point in time. It is a good candidate to model specific processes related to CoPEIQ. For example, it may be applied quite successfully to understand the time and quality tolerances for developing EI content during a major capital projects where hundreds of engineers are developing content across several continents. The objectives of this current research are too general for this level of detail. This technique is therefore discounted for the purposes of this thesis.

Monte Carlo Simulation

MCS is described by Mooney (1997) as "a flexible and powerful tool" and an empirical technique "using random samples from known populations of simulated data". If many samples are drawn from a population, then their relative frequency distribution is an indication of the population's density function. For practical reasons the "many samples" are generated artificially. From within the context of this section, MCS may be viewed as a specific simulation technique.

According to Kroese et al. (2014), MCS are used, in the main, for the following three classes of calculation:

- Optimization
- Numerical integration
- Drawing from a probability distribution

MCS is suitable where input uncertainty is high, such as business risk.

These assessments demonstrate the efficacy of MCS for the current work, where the intent is to derive a stochastic result for a very uncertain problem, of which the causality or exact mechanisms are not proven but are instinctively understood to be complex.

MCS is discussed in more detail in Section 4.5.2.

Queueing Theory

Queueing systems are studies of queues or waiting dynamics from a mathematical perspective. They are described in terms of input processes, arrivals, queues, service facilities and departures. (Hines & Montgomery (1980)). As is the case with other specific, time-based or sequential techniques, there may be a case for the analysis of specific processes, such as EI acceptance or review; however, the current, initial research objectives are too general for this technique.

Simulation

Sawilowsky (2003) quotes several authors to define a simulation as either mimicking important elements of a phenomenon, "a representation … in simplified form to study its behaviour" or "forming an abstract model from a real situation to understand the impact of modifications and the effect of interventions".

Rubinstein & Kroese (2017) describe simulation as being appropriate for situations that are too complex to be solved analytically, and note its wide use in the general Operations Research field. They quote Naylor et al. (1966) as defining simulation as follows: "Simulation is a numerical technique for conducting experiments on a digital computer, which involves certain types of mathematical and logical models that describe the behaviour of business or economic systems (or some component thereof) over extended period of real time ".

The same authors describe the process of simulation as follows: (i) Construct a model (defined as "an abstraction of some real system") of the system (defined as "a collection of related entities") being modelled and (ii) Solve the model, either numerically or analytically and either deterministically or stochastically. According to their thinking, a simulation models a system, consisting of entities. Entities may also be called elements, each in turn possessing attributes. These may be in the form of logical or numerical values.

They list the following benefits of simulation: (i) a teaching aid, (ii) a means to understand more of the subject under study, (iii) a tool to identify important variables and (iv) a digital ("in silico") laboratory to study a certain phenomenon. They proceed to explain the need of a model to balance realism and simplicity when constructing a model and distinguish between analytical models and stochastic models.

Finally, they classify simulation models in the following three ways: (i) Static vs Dynamic; (ii) deterministic vs Stochastic and (iii) Continuous vs Discrete.

Simulation, like AI, can be viewed as a general or umbrella term. The description of Rubinstein & Kroese (2017) align very closely to the research objectives for this study and as a result simulation is deemed specifically appropriate. More specifically, a simulation in the Monte Carlo format is appropriate, because of its inherent ability to quantify the uncertainty of a phenomenon in a statistically defensible manner.

Conclusion

This section has provided an overview of the various stochastic techniques and has concluded that there may be a case for applying some of them on the CoPEIQ problem in future research. However, given the current, relatively immature state of knowledge and business drivers for improving CoPEIQ, the primary objective is to expose the value to OpCo management. For that objective, MCS is the appropriate approach. That is therefore the subject of the next section.

4.5.2 Monte Carlo Simulation - an Overview

In the preceding section, it was determined that none of the numerical analysis techniques were appropriate for this study, thereby confirming the original contention that MCS is the appropriate stochastic approach. In this section the subject MCS is reviewed to a sufficient level of detail, using the knowledge gained from the literature. The section ends with an interpretation of the insights specifically gained for this study.

4.5.2.1 Introducing Monte Carlo Simulation

The principles underlying MCS are traced back to the 1700s, when George Louis LeClerc did a study to determine the value of π in the "Buffon's needle" method. (Harrison 2010.) Of the several accounts of the origins of MCS in the literature, that given by Rubinstein & Kroese (2017) is the most informative. According to that account, Von Neumann & Ulam coined "Monte Carlo" as a code word for the secret work they were doing to solve problems around random neutron diffusion, which was needed to develop the atom bomb in the USA during World War II.

There appears to no consensus on how MCS should be formally defined. Ripley (1987) and Sawilowsky (2003) offer differing definitions. For example, Sawilowsky contends that an MCS "uses repeated sampling to determine the properties of some phenomenon". Kalos & Whitelock (2008) feel, however, that these differences are not always applicable. Harrison (2010) suggests that it is difficult to define MCS: "any attempt to define one will inevitably leave out valid examples".

Thomopoulos (2013) distinguished between terminating and non-terminating systems. The former has a defined start and end events, whereas the latter include starting and/or ending transients, steady-state "equilibrium" and cyclical stages. The simulation models for each of these system types should be configured differently.

Recent advancements in this area include using Monte Carlo resampling as Bayesian statistical inference. Gordon et al. (1993) published on this subject in what is now deemed a seminal work.

4.5.2.2 Designing a Monte Carlo Simulation

In this section, some general points regarding MCS design are listed. The section starts with the need to understand the domain, proceeds to some general design points, lists various views on the process to construct a MCS and ends with some pitfalls to avoid.

Mooney (1997) makes the obvious point that an analyst should have a good understanding of the domain of study and the relevant statistical theory. Thomopoulos (2013) supports the notion that input random variables should align as closely as possible with the actual system being studied. Indeed, the analyst is "obliged to seek actual data". Ha (2009) explores the possibility of replacing actual data with assumed distributions, the problems with truncation of assumed or actual distortions and non-normality of distributions. He concludes, for his example of optimizing stock levels, that truncation produces a "disproportionately larger" error during MCS than predicted by deterministic means and shows an interaction between his input variables. He concludes that, in general, "it is prudent to proceed cautiously" when modelling arbitrarily shaped distributions. He also concludes specifically that an *a priori* knowledge of the variable distribution is preferred and, if this is not possible, a sensitivity analysis of the results is recommended.

Harrison (2010) suggests a general pattern for MCS might be:

- develop a model of a system by means of a series of Probability Density Functions (PDF)

- take repeated samples from the PDFs and calculate a result of the model
- analyse the repeated results of the samples to derive a statistical result.

A section on which variables to save in an MCS is dedicated by Mooney (1997). He makes the point that the simulation design should consider saving not only variables about the desired model output, but also include variables that may enable diagnosis of the actual model, thereby assisting in debugging the model during the development phase. The same author discusses the number of MCS trails needed and concludes that, given the ease of modern computing, "lots "of trails should be done. This will reduce the variability and improve the power of the model.

Sawilowsky (2003) contends that a high-quality MCS should meet the following requirements:

- The random number generator is pseudo-random and has a "long" period between repeats
- The numbers generated clear a random test pass
- The sample volume is adequate for accurate simulation
- Sampling is done appropriately
- The algorithm is valid for the problem being studied
- The simulation represents the phenomenon under study adequately.

Mooney (1997) summarizes the procedure to execute an MCS as follows:

- develop a computer algorithm to specify the pseudo-population of the system under study
- collect a pseudo-sample from the pseudo-population in a manner that is comparable to the pseudo-population
- calculate the estimator of the social characteristic under investigation
- repeat for a certain number of trails.
- derive the frequency distribution of the results derived from the trails. This results in the Monte Carlo estimate of the social characteristic under investigation.

Kroese et al. (2014) state that MCS generally take the form:

- define a range of inputs ("input domain")
- draw inputs from the probability distribution at random
- calculate the result deterministically for each input
- aggregate the results of each calculation in the form of a distribution.

Thomopoulos (2013) distinguishes between variable type data and proportional type data. He provides examples of each type:

- Variable type: time to complete an activity, structural steel strength, labour cost for an activity.

- Proportional type: portion of defects in a sample, portion of customer using credit cards, portion of call wait times above a certain time.

He then distinguishes between the calculation of the following variables for each data type variable or proportional:

- Sample mean
- Sample variance
- Confidence interval for normal and not-normal
- Seeking more accuracy
- Comparing two options.

Mooney (1997) reflects on the limitation of specificity during the interpretation of MCS; the results only reflect the situation during the time of the experiment. If the sample size, independence, error distribution or other variables were different, to what extent would the results be different? He suggests that both the potential sources of variation and the interdependency of these sources need to be considered. A response to specificity may be the construction of a matrix of experiments vs factors that impact the experiment. The problem is termed "functional indeterminacy" and Mooney (1997) proposes the use of Response Surface Analysis (after Hendry, 1984). This approach uses polynomial equations of increasing complexity until the data fits satisfactorily.

As early as 1970, Hasting (1970) wrote about errors in Monte Carlo methods and strategies to address them. He quotes Fox & Mayers (1968), who showed how "even the simplest of numerical methods" may yield "spurious results" in the presence of "insufficient care". He lists some common sources of error for MCS methods:

- A low-quality source of uniform random numbers
- Nonnormality of the estimate distribution
- Computational errors arising from either the sample or the actual calculation
- Too small a sample size.

Mooney (1997) makes the following suggestions to reduce the likelihood of error and confusion when designing a n MCS:

- ensure a good understanding of the social processes under study

- ensure a good understanding of the relevant statistical theory

- plan the analysis and its objectives in detail before commencing coding

- develop an explicit model of the characteristics under study

- develop code in a modular fashion, ensuring every element of the code is robust before integrating the elements.

In addition to the general points above, a few specific subjects were identified that are discussed subsequently.

Random Number Generation

This section discusses the subject of Random Number Generation (RNG). After describing the challenge and some general points about it, four methods of RNG are presented.

RNG is universally agreed by the sources quoted to be a difficult subject. Rubinstein & Kroese (2017) explain that early simulation models assumed true randomness was only achieved by mechanical or electronic means. This paradigm, however, is slow and expensive and is not necessarily independent or without bias, and lacks repeatability. Indeed, generating truly random numbers by mechanical or manual means is subject to error, manipulation, exhausting options, or is simply too costly. (Mooney (1997)). As a result, algebraic methods have been developed to generate "pseudo-random" numbers. Mooney (1997) notes that this is a valid approach provided independence and uniformity are "rigorously" confirmed. He also notes the tendency of pseudo-random numbers to repeat at some point.

Thomopoulos (2013) suggests that a sequence of uniform variates can be tested for randomness by means of the cycle length, mean, variance, chi-square and autocorrelation.

Several methods exist to generate pseudo-random numbers. These are listed and discussed next.

Rubinstein & Kroese (2017) mention "linear congruential generators (LCG's), of which the "multiple congruential method" mentioned by Mooney (1997) is a special case. Mooney (1997) contends that "multiple congruential methods" are the most common algorithms. Thomopoulos (2013) contends that LCG is the most common method. He adds that LCG was introduced in 1951 by Lehmer.

Thomopoulos (2013) explains the use of the mathematical function "modular arithmetic" where "for a variable w, the *modulo* of w with *modulus* m returns the remainder of w when divided by m that is an integer". He provides the example for $m = 5$ and $w = 1, w.modulo(m) = 1, modulo(5) = 1$. Ross (1990) summarizes the approach as follows: (i) $X_n = a.x_{n-1}modulo(m)$ m with x_{seed} as the first x_{n-1}. (ii) Repeat until enough x has been generated. The word "*modulo*" means that the product $a.x_{n-1}$ is divided by m, the remainder is retained as x_n. Other authors explain this modulus–based approach using different language, but the principle remains the same. Rubinstein & Kroese (2017) mention the model by Lewis, Goodman & Miller (1969) where $m = 2^{31} - 1$, $a = 7^4$and $c = 0$. Thomopoulos (2013) explains that $2^{31} - 1$ is the largest number recognized by a 32-bit machine.

Mooney (1997) summarize the Inverse Transformation (IT) methods by stating that an inverse distribution function G(a) produces Pr (X, x) = a if X is distributed as F(X). This method is also described by Rubinstein & Kroese (2017). They add the limitation that this method only works where the inverse transform of a distribution can be found.

The Acceptance/Rejection (AR) method uses a random number p that is uniformly distributed. An algorithm samples x from a pseudo-sample; if the density of a PDF using x is less than p, it is accepted; if not, it is rejected. (Mooney (1997)). Rubinstein & Kroese (2017) also mention this method, explain it in a more graphic format and add that its use is indicated when the other methods are inefficient or simply don't work.

Thomopoulos (2013) mentions only two methods, IT and AR, but adds the perspective that IT is preferred if the distribution function can transform. Conversely, if the mathematics for transformation are complex, AR is used, although it requires more steps.

The Composition Method is used when an inverse distribution function G(a) is not able to be derived. This method is based on the combining or transforming of distributions derived from IT or other methods, but doing this composition ensuring that the resultant distribution is of the desired type. Rubinstein & Kroese (2017) explains this method by stating that a cumulative distribution function (CDF) can be expressed as a mixture of individual CDFs.

In addition to the above, Rubinstein & Kroese (2017) makes mention of the so-called Alias method. Since it is applicable to discrete distributions only, it is not discussed further.

The IT method depends on the existence or achievability of the inverse of a PDF. It follows that the RNG method and the PDF are related. As a result, Table 4.5 relates the ideal RNG method per PDF, to the extent that the literature expounds it.

Determining the representative Probability Density Function for the sample data

In this section, the process of assigning a PDF to an experimental data set is explored. As will be seen in Section 4.5.3, this is an important element of this study. After some general comments, relevant PDFs are listed and discussed, together with appropriate RNG methods.

Mooney (1997) starts the discussion of this subject by quoting Johnson (1987) that "little practical or even theoretical guidance" is available for this difficult problem. He suggests that the following three aspects need to be considered, each with respect to the actual problem being investigated:

- the data range
- the shape of the distribution
- the ease by which the distribution function will allow the variation for which the researcher needs to test during the experiment.

For the second of these points, several goodness-of-fit tests exist.

Mooney (1997), Thomopoulos (2013) and Rubinstein & Kroese (2017) all provide a list of CDFs frequently encountered or used in both the discrete and continuous domains. These are listed in Table 4.5. Since it has been determined in Section 4.5.1.1 that CoPEIQ is a continuous problem, only continuous distributions are discussed.

Table 4.5 - Continuous Probability Distributions and Associated RNG Methods

Continuous Distribution	Mooney (1997)		Rubinstein & Kroese (2017)		Thomopoulos (2013)	
	Comment	RNG method	Comment	RNG method	Comment	RNG method
Uniform	U (0,1) is basis of all MCS	IT	Listed	IT	Listed	Routine provided
Pareto		IT				
Exponential	Frequently used in industrial engineering	IT	Listed	IT	Often in queuing systems	IT
Erlang					Listed	Routine provided
Normal	Most frequently used	Composition	Listed	Box & Muller or Accept/Reject	Most widely used	Hastings approximation; Convolution method; Sine/ Cosine method
Lognormal	Close to Pareto to the right of the mode – which portion is important	Composition			Listed	Routine provided
Chi-square	Share changes per DF; useful to assess if the normality assumption is valid	Composition			Usually used to test the variability of the variance of a variable	Approximation
Student's T	Useful to vary leptokurticy	Composition			Used to test the significance of the mean value of a variable	
Beta	Very flexible	AR			Many shapes =f(pars)	Routine provided = f (k1, k2)
Gamma			Listed	Not IT	Listed	Not easy. Method =f(k)
Beta			Listed	IT		
Weibull					Listed	IT
Fischer's F					> 2 variances from normal	Routine provided

Thomopoulos (2013) dedicates an entire chapter on selecting an appropriate PDF from a data set. He proposes the following procedure for fitting a PDF to data:

- Confirm data independence
- Calculate statistics
- Select candidate PD's
- Estimate parameters for each PD
- Test adequacy of the fit.

He suggests that an easy test is to calculate the coefficient of variance (CoV) for the sample data set and comparing the result to expected values of CoV for specific PDs. Table 4.6 therefore summarizes the expected CoV's. Note that the covariance for the PDs Beta and Weibull do not calculate to specific values and are therefore not included. In passing Thomopoulos (2013) notes the expediency of transforming variables to normalized form (0,1) or X≥0. He also explains that estimation may be done by means of either maximum-likelihood and/or method-of-moment techniques.

Table 4.6 - Expected Covariances for Selected Continuous PDF's (after Thomopoulos (2013)

Distribution	Variable range	Expected CoV
Continuous Uniform	(0,1)	$CoV = 0.577$
Normal	$X \geq 0$	$CoV \leq 0.33$
Exponential	$X \geq 0$	$CoV = 1.00$
Lognormal	$X' = \ln(x)$	$CoV' = \sigma'/\mu'$
Gamma	$X \geq 0$	$CoV < 1; K > 1$
		$CoV \geq 1; K \leq 1$

Thomopoulos (2013) demonstrates, by means of several examples, the use of the so-called Q-Q plot to estimate goodness-of-fit between the quartiles of sample data and that of a specified PD.

Rubinstein & Kroese (2017) describe the Limit Theorems, of which the Central Limit Theorem (CLT) is the most useful for this study. Indeed, Thomopoulos (2013) quotes the CLT frequently in the routines for RNG that he provides.

Inference

Comments from the literature regarding the inference of information from sample data are summarized in this section.

Thomopoulos (2013) contends that usual statistical techniques are appropriate to analyse the outputs of simulation runs, but only if the runs were generated using different variates, meaning the individual runs are independent.

Mooney (1997) suggests that the "dominant inferential paradigm today" is to use standard inferential parametric. This paradigm, however, requires proof of the validity of the distributions used and adequate statistical theory. There are, however, two situations for which this paradigm may be invalid:

- A given situation may not be equal to the conditions under with the theory was developed.
- Inadequate statistical theory exists for the appropriate statistic.

He then proceeds to explain how Monte Carlo simulation may be used to address these limitations. Further, MCS is "the only general way" to address the problem of needing accurate population parameters for credible inference. In that way, the quality of an inference method may be determined. He relates this to HT and explains how Type I errors are far worse to commit than Type II errors and that Type I errors are judged against an "absolute standard", as opposed to the relativity of Type II errors.

4.5.2.3 Interpreting Monte Carlo Simulation Design Aspects for This Study

Having explored the specifics of MCS design from the literature in the preceding paragraphs, this section concludes with an interpretation as it pertains to this study specifically. Eleven design aspects will be addressed in the order in which they have been described previously in this section.

- As noted in Section 4.5.1, data is assumed to continuous and normally distributed in this study.

- The issue of domain knowledge is, in this study, deemed addressed adequately. The specifics of the domain have been considered in some detail during the survey design and subsequent verification structured interviews. An idea of the order of magnitude of both inputs and outputs was gained during the verification structured interviews and already used in the deterministic model, Section 4.4.

- The variables to save from the output of the Monte Carlo analysis will be explored in detail in Section 4.5.3.

- Clear guidance on the number of trials is not forthcoming from the literature. The initial design is therefore arbitrarily set at 10,000, after which a confidence test will be calculated.

- Four views on a typical MCS process have been presented. A comparison of these, taken together with the precedent of the deterministic model, yields the process steps for this study as follows:

    - Calculate CoPEIQ using the deterministic model for each response and capture the results as a table.

    - Characterize sample data by calculating the statistics of each question.

    - Generate a series of pseudo-random trials using the variables calculated.

    - Calculate the variables and the confidence interval.

    - Report the results according to the guidance developed in Section 4.2.

More details are provided in Section 4.5.3.

- Based on the definition of Thomopoulos (2013), the data type of this study is clearly of the variable type (as opposed to proportional.)

- There are always changes, organisational or technical, ongoing in an OpCo. For this study, however, the organisational "system" under review is assumed to be a non-terminating one, in equilibrium and steady-state. It may very well be that one or more dimensions of the processes under review may be in a transient state at the point of the survey being undertaken. Since the primary objective is to expose CoPEIQ and not a detailed analysis of the effect of one or more internal or external changes on the organisation and its performance, it must for all practical purposes be assumed that the organisation is in a steady state.

- The same rationale can be applied to the problem of specificity – the intent is not to test the impact of variables on CoPEIQ, but merely to determine CoPEIQ at the point in time when the survey is undertaken.

- Since Microsoft Excel will be used for this study, its default RNG method has been investigated to ensure it meets the requirements discovered in the literature and listed in this section. It was found that the modern version of Excel uses an algorithm called AS 183, written by Wichmann & Hill (1982). (support.microsoft.com, accessed March 30, 2018). It is constructed of three LGCs of which the prime moduli are different and the result of which is added, modulo, 1. In broad terms it can be said that it uses modular arithmetic. It passes the so-called "Diehard" battery of tests, developed by Professor Marsaglia at Florida State University and available at http:/i.cs.hku.h/~diehard. The algorithm is guaranteed to have a life cycle of 10^{13} numbers. It can therefore be safely assured that Excel's RNG functionality will meet the requirements.

The conditions for which it is deemed valid to use standard statistical theory to infer results, as listed by Mooney (1997) and Thomopoulos (2013), have been met by discussion in this section.

A detailed review of MCS has been reported in this section and interpreted for the purposes of this study. The foundation has been created and the discourse of this thesis is ready for the design, creation and testing of the MCS. This follows in Section 4.5.3.

4.5.3 Component Parts of the Monte Carlo Simulation

The MCS will be designed, built and tested in this section. The preceding sections have yielded some design decisions, which will be implemented here. The approach is in four parts, summarized as follows:

- Decide the general model architecture
- Build the model
- Demonstrate the model
- Test the model using artificially generated "laboratory" data.

In this way, the model is constructed and validated, ready for actual field use.

4.5.3.1 Decide the Model Architecture

In this first section, a fundamental architecture for the MCS is selected from three potential options, simply called Options 1, 2 and 3 respectively. These options are described below, after which they are shown schematically in Figure 4.7.

Option 1

Option 1 views all survey questions responses as one population. This means that all responses are aggregated and viewed as the population of one PDF. The logic flow would be as follows:

- To normalize the unit of measure to USD, calculate C_{ijk} for each response

- Aggregate all responses into one sample
- Calculate the parameters of the normal distribution
- Generate 10,000 runs of pseudo-random numbers using the calculated parameters
- Add C_{VAR} to every generated result of C_{ijk}
- Calculate the parameters of this population
- Report results.

This approach is the simplest, therefore requiring the least number of calculations, but raises the risk of inaccuracy and does not enable stochastic disaggregation of the result. Individual elements of C_{ijk} will also vary considerably in terms of order of magnitude. Consider, for example, a hypothetical scenario where one respondent estimates that 1 hour per month is spent on a certain IE and another respondent considers a 5% likelihood of a PSI that costs the OpCo 20 million dollars. The variation in order of magnitude will result in a very flat distribution and statistical results of suspicious validity.

Option 2

Option 2 considers each group g_i as a separate statistical entity. The process described for Option 1 is therefore repeated 4 times, once for each group g_i. The resultant figures are added to yield a distribution of C_{ijk}. After adding C_{VAR}, the resultant CoPEIQ distributions are fit to four PDFs. These PDFs are pseudo-generated individually and individual PDFs fit to each set. The results are then reported separately and added for a final CoPEIQ result and associated confidence interval.

Option 2 is a compromise between the extremes of a detailed, very accurate answer and a simple calculation of adequate accuracy. This approach will allow stochastic disaggregation of the group elements time, production, cost and risk. This stochastic disaggregation may be of value for remediation prioritisation, but the problem of order of magnitude may be further exacerbated by the smaller samples.

Option 3

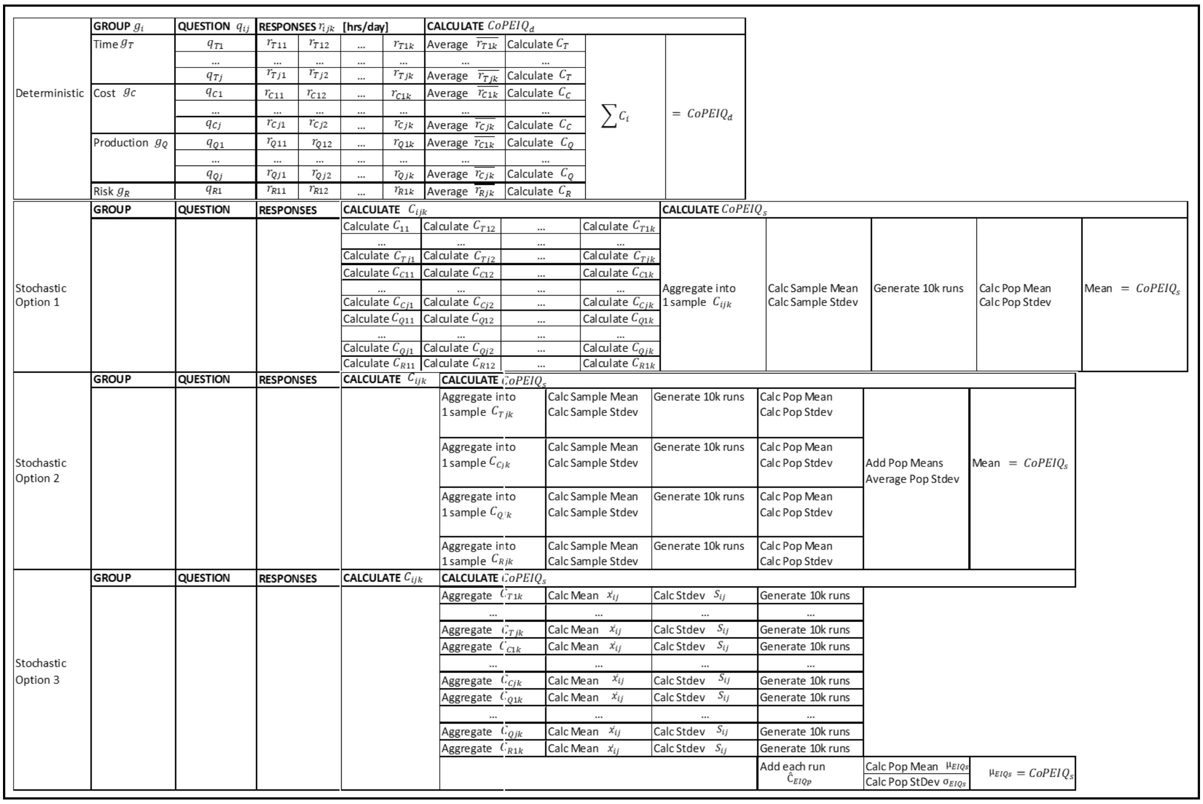
Option 3 considers each set of responses r_{ijk} as a separate distribution. The process is therefore repeated 42 times to yield 42 populations. For each question, parameters are calculated and used to generate 10,000 runs of pseudorandom data. The parameters of this new population constitute the results of stochastic CoPEIQ.

As explained in the next paragraph, this final option is the most accurate in the following sense:

The fundamental approach for this research is to test the opinion of a sample of individuals regarding several IE's of the central problem, namely CoPEIQ. The results effectively added to represent the aggregated opinion of one individual with respect to the total effect of poor EIQ, expressed as CoPEIQ. For the deterministic model, the results were averaged per question and then added; however, the result is the same due to the distributive property. Using Option 3, individual results are quantified statistically and a pseudo-random population of results is generated for each response. The individual results are added per run and therefore, in effect, per pseudo-respondent. The resulting population is deemed a new PDF, which constitutes the result of the MCS. This approach therefore, in effect, simulates 10,000 responses, based on the characteristics of the original sample responses per question. This approach therefore simulates the reality in the closest practical way and is therefore a true MCS of the sample.

Option 3 implies a large amount of calculation; however, this is deemed to fall well within the abilities of Excel and is therefore a negligible constraint.

The preceding discussion has been summarized in graphical form in Figure 4.7. A similar summary of the deterministic model is included for comparative purposes.



	GROUP g_i	QUESTION q_{ij}	RESPONSES r_{ijk} [hrs/day]				CALCULATE $CoPEIQ_d$			
Deterministic	Time g_T	q_{T1}	r_{T11}	r_{T12}	...	r_{T1k}	Average $\overline{r_{T1k}}$	Calculate C_T	$\sum C_i$	$= CoPEIQ_d$
		...	...	...	...	...	...	...		
		q_{Tj}	r_{Tj1}	r_{Tj2}	...	r_{Tjk}	Average $\overline{r_{Tjk}}$	Calculate C_T		
	Cost g_C	q_{C1}	r_{C11}	r_{C12}	...	r_{C1k}	Average $\overline{r_{C1k}}$	Calculate C_C		
		...	...	...	...	...	...	...		
		q_{Cj}	r_{Cj1}	r_{Cj2}	...	r_{Cjk}	Average $\overline{r_{Cjk}}$	Calculate C_C		
	Production g_Q	q_{Q1}	r_{Q11}	r_{Q12}	...	r_{Q1k}	Average $\overline{r_{C1k}}$	Calculate C_Q		
		...	...	...	...	...	...	...		
		q_{Qj}	r_{Qj1}	r_{Qj2}	...	r_{Qjk}	Average $\overline{r_{Cjk}}$	Calculate C_Q		
	Risk g_R	q_{R1}	r_{R11}	r_{R12}	...	r_{R1k}	Average $\overline{r_{Rjk}}$	Calculate C_R		

	GROUP	QUESTION	RESPONSES	CALCULATE C_{ijk}				CALCULATE $CoPEIQ_s$				
Stochastic Option 1				Calculate C_{11}	Calculate C_{T12}	...	Calculate C_{T1k}	Aggregate into 1 sample C_{ijk}	Calc Sample Mean Calc Sample Stdev	Generate 10k runs	Calc Pop Mean Calc Pop Stdev	Mean $= CoPEIQ_s$
				...	...	...	...					
				Calculate C_{Tj1}	Calculate C_{Tj2}	...	Calculate C_{Tjk}					
				Calculate C_{C11}	Calculate C_{C12}	...	Calculate C_{C1k}					
				...	...	...	...					
				Calculate C_{Cj1}	Calculate C_{Cj2}	...	Calculate C_{Cjk}					
				Calculate C_{Q11}	Calculate C_{Q12}	...	Calculate C_{Q1k}					
				...	...	...	...					
				Calculate C_{Qj1}	Calculate C_{Qj2}	...	Calculate C_{Qjk}					
				Calculate C_{R11}	Calculate C_{R12}	...	Calculate C_{R1k}					

	GROUP	QUESTION	RESPONSES	CALCULATE C_{ijk}	CALCULATE $CoPEIQ_s$					
Stochastic Option 2					Aggregate into 1 sample C_{Tjk}	Calc Sample Mean Calc Sample Stdev	Generate 10k runs	Calc Pop Mean Calc Pop Stdev	Add Pop Means Average Pop Stdev	Mean $= CoPEIQ_s$
					Aggregate into 1 sample C_{Cjk}	Calc Sample Mean Calc Sample Stdev	Generate 10k runs	Calc Pop Mean Calc Pop Stdev		
					Aggregate into 1 sample C_{Qjk}	Calc Sample Mean Calc Sample Stdev	Generate 10k runs	Calc Pop Mean Calc Pop Stdev		
					Aggregate into 1 sample C_{Rjk}	Calc Sample Mean Calc Sample Stdev	Generate 10k runs	Calc Pop Mean Calc Pop Stdev		

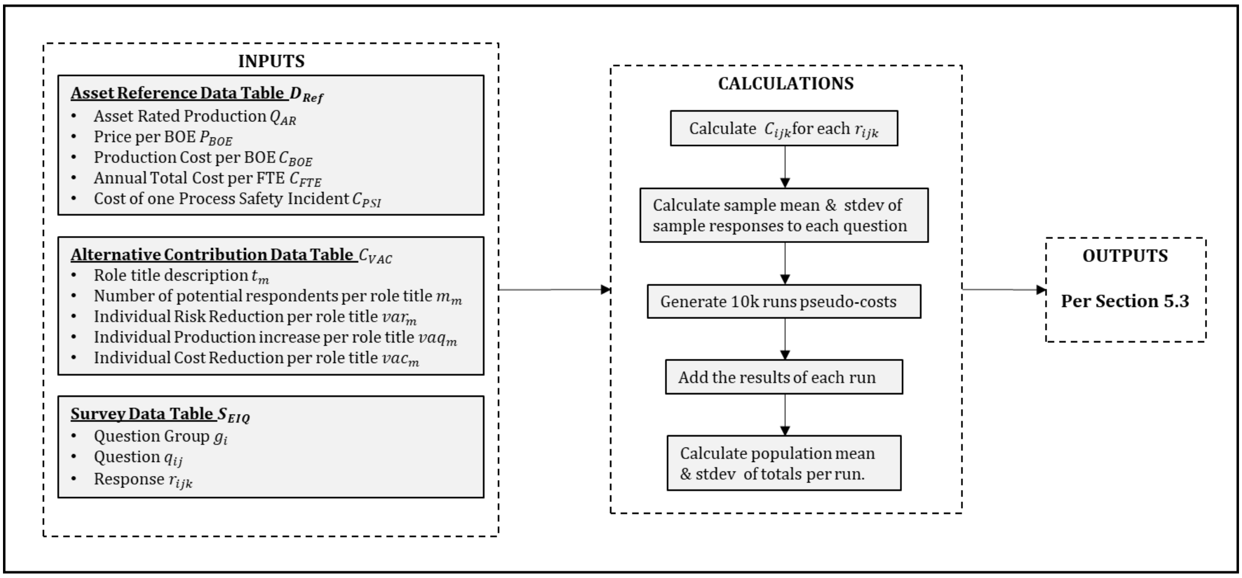
	GROUP	QUESTION	RESPONSES	CALCULATE C_{ijk}	CALCULATE $CoPEIQ_s$					
Stochastic Option 3					Aggregate C_{T1k}	Calc Mean $\bar{x}_{ij}$	Calc Stdev S_{ij}	Generate 10k runs		
					...	...	...	...		
					Aggregate C_{Tjk}	Calc Mean $\bar{x}_{ij}$	Calc Stdev S_{ij}	Generate 10k runs		
					Aggregate C_{C1k}	Calc Mean $\bar{x}_{ij}$	Calc Stdev S_{ij}	Generate 10k runs		
					...	...	...	...		
					Aggregate C_{Cjk}	Calc Mean $\bar{x}_{ij}$	Calc Stdev S_{ij}	Generate 10k runs		
					Aggregate C_{Q1k}	Calc Mean $\bar{x}_{ij}$	Calc Stdev S_{ij}	Generate 10k runs		
					...	...	...	...		
					Aggregate C_{Qjk}	Calc Mean $\bar{x}_{ij}$	Calc Stdev S_{ij}	Generate 10k runs		
					Aggregate C_{R1k}	Calc Mean $\bar{x}_{ij}$	Calc Stdev S_{ij}	Generate 10k runs		
								Add each run $\hat{C}_{EIQp}$	Calc Pop Mean μ_{EIQs} Calc Pop StDev σ_{EIQs}	$\mu_{EIQs} = CoPEIQ_s$

Figure 4.7 – Monte Carlo Simulation Architecture Options

In Sections 4.1 and 4.2 it was concluded that the target audience prefers information in a format that is relatively simple and that the audience is comfortable with stochastic results if the confidence of the results is quantified. The calculation of this data was intended to show an order of magnitude for CoPEIQ, which will enable decision-makers to weigh the benefits of EIQ against other opportunities, to select the best deployment of available improvement funds. There was also a need for disaggregated data in a graphical form, in a way that will permit prioritization of EIQ remediation efforts. The presentation of disaggregated results is only significant once a decision-maker has been convinced that an investment into EIQ is warranted. It may therefore be argued that the stochastic CoPEIQ result prompts a primary decision, and that a review of disaggregated results is secondary.

Based on this argument, Option 3 has been selected for this study. A single stochastic figure for CoPEIQ, with an associated confidence interval, is deemed adequate for the primary decision. Once the decision for investment has been made, the disaggregation required for the prioritization of EIQ remediation efforts may be done deterministically, rendering more detailed stochastic calculation superfluous.

The selected approach may also be shown graphically. Using the same conventions as in Figure 4.5, the selected approach is shown in Figure 4.8 below.



INPUTS
Asset Reference Data Table D_{Ref}
• Asset Rated Production Q_{AR}
• Price per BOE P_{BOE}
• Production Cost per BOE C_{BOE}
• Annual Total Cost per FTE C_{FTE}
• Cost of one Process Safety Incident C_{PSI}
Alternative Contribution Data Table C_{VAC}
• Role title description t_m
• Number of potential respondents per role title m_m
• Individual Risk Reduction per role title var_m
• Individual Production increase per role title vaq_m
• Individual Cost Reduction per role title vac_m
Survey Data Table S_{EIQ}
• Question Group g_i
• Question q_{ij}
• Response r_{ijk}
CALCULATIONS
Calculate C_{ijk} for each r_{ijk}
Calculate sample mean & stdev of sample responses to each question
Generate 10k runs pseudo-costs
Add the results of each run
Calculate population mean & stdev of totals per run.
OUTPUTS
Per Section 5.3

Figure 4.8 – Selected Monte Carlo Simulation Architecture

A primary architecture for the MCS having been selected, the model build can proceed.

4.5.3.2 Inputs of the MCS

Consistent with the advice from Mooney (1997) to build the MCS in incremental procedures first and then combine them, the constituent procedures for the MCS are shown in Figure 4.8, which is an extension of Figure 4.7. After discussing each procedure in turn, this section will conclude with the assembly of the various procedures.

In Section 4.4, the notation C_{XXX} was used consistently for "Cost", in the UoM [USD], with "XXX" representing various cost elements, such as "PSI" for "Process Safety Incident". In this section, the notation $\hat{C}_{XXX}$ is introduced to distinguish between costs that have been calculated deterministically and those generated for stochastic purposes. The notation "$\hat{C}$"may therefore be referred to as "pseudo-cost".

Accordingly, the following additional variables are introduced for the stochastic model:

- C_{ijk}= Individual cost calculation per survey data point r_{ijk} in the UoM [USD/year]
- $\bar{x}_{ij}$ = The sample mean of the responses for a specific question, in [USD/year]
- S_{ij} = The standard deviation of the responses for a specific question, in [USD/year]

- $\hat{C}_{ijp}$ = Pseudo-costs generated for a specific question, in [USD/year]
- p = number of RNG runs, $p \in \{1,2,3 \dots 10{,}000\}$ (in contrast to r_{ijk} = individual responses r_{ijk} ; $k \in \mathbb{R}$)
- $\hat{C}_{EIQp}$= Population of the sum of p individual pseudo-responses, in the UoM [USD/year]
- μ_{EIQs} = Mean of the population $\hat{C}_{EIQp}$
- σ_{EIQs} = Standard deviation of the population $\hat{C}_{EIQp}$
- $\hat{C}_{EIQs}$= Cost of Poor EIQ-stochastic; the equivalent of C_{EIQd}, in the UoM [USD/year]
- $CoPEIQ_s$ = Total Cost of Poor EIQ-stochastic, in the UoM [USD/year]

These additional variables were used in the stochastic calculations as is discussed in the next section.

4.5.3.3 Calculations of the MCS

In Section 4.5.3.1, the selected architecture and the sequence of calculation were described. The variables defined in Section 4.5.3.2 are applied in this section to describe the calculations in detail, using the detail sequence described under “Survey Calculations” in Figure 4.6 as a guide.

- To normalize the unit of measure to USD, calculate C_{ijk} for each response. For this step, four different calculations for the four elements of the group g_i; $i \in \{C, T, R, Q\}$ are described:
  - $C_{Qjk} = \frac{r_{Qjk}}{100} . Q_{AR}.\ A_{Plt}.R_{BOE}.365$
  - $C_{Tjk} = \frac{t_{Tjp}.N.C_{FTE}}{10}$
  - $C_{Cjk} = r_{Cjp}$
  - $C_{Rjk} = r_{Cjp}.\ C_{PSI}$
- Calculate the mean $\bar{x}_{ij}$ and standard deviation S_{ij} for each response:
  - $\bar{x}_{ij} = (\sum_{r=1}^{k} \sum_{j=1}^{n} r_{ijk}) / k$
  - $S_{ij} = \sqrt{\frac{[x-\bar{x}]^2}{n-1}}$
- After generating 10,000 runs of pseudo-random numbers $\hat{C}_{ijp}$ using the NORM.INV command of Excel, add these results per pseudo-respondent using the calculation

  - $\hat{C}_{EIQp} = \sum_{p=1}^{ij} \hat{C}_{ijp}$

- Calculate the mean of the population of pseudo-sums $\hat{C}_{EIQp}$ using the calculation
  - $\mu_{EIQs} = (\sum_{p=1}^{p} \hat{C}_{EIQp})/p$
- Calculate the standard deviation of the population of pseudo-sums $\hat{C}_{EIQp}$ using the calculation
  - $\sigma_{EIQs} = \sqrt{\frac{[x-\bar{x}]^2}{p}}$
- Calculate the 5% confidence interval of the population of pseudo-sums $\hat{C}_{EIQp}$ using the calculation s
  - $\mu_{EIQs} + 1.96\sigma_{EIQs}$ and $\mu_{EIQs} - 1.96\sigma_{EIQs}$
- The calculated mean of the population of pseudo-sums μ_{EIQs} is equal to the Cost of Poor EIQ- stochastic $\hat{C}_{EIQs}$, or:
  - $\hat{C}_{EIQs}= \mu_{EIQs}$
- The Total Cost of Poor EIQ – stochastic $CoPEIQ_s$ is calculated as follows:
  - $CoPEIQ_s = \hat{C}_{EIQs} + C_{VAR}$

These calculations form the basis of the stochastic part of the CoPEIQ model.

4.5.3.4 Outputs of the MCS

As concluded in Section 4.2, the list of outputs of the stochastic model are:

- $\hat{C}_{EIQs}$
- $CoPEIQ_s$
- σ_{EIQs}
- Confidence intervals

As for the outputs of the deterministic model described in Section 4.4.2, additional outputs in graphical form are discussed in Chapter 5.

4.5.3.5 Demonstrating the Stochastic Model

As has been the case in Section 4.4 and consistent with Figures 4.5 and 4.6, this section will demonstrate the model by means of a few screen shots from the Excel file. The reader is reminded of the worksheet and cell colour coding described in Section 4.4.3, which will become more evident in this section.

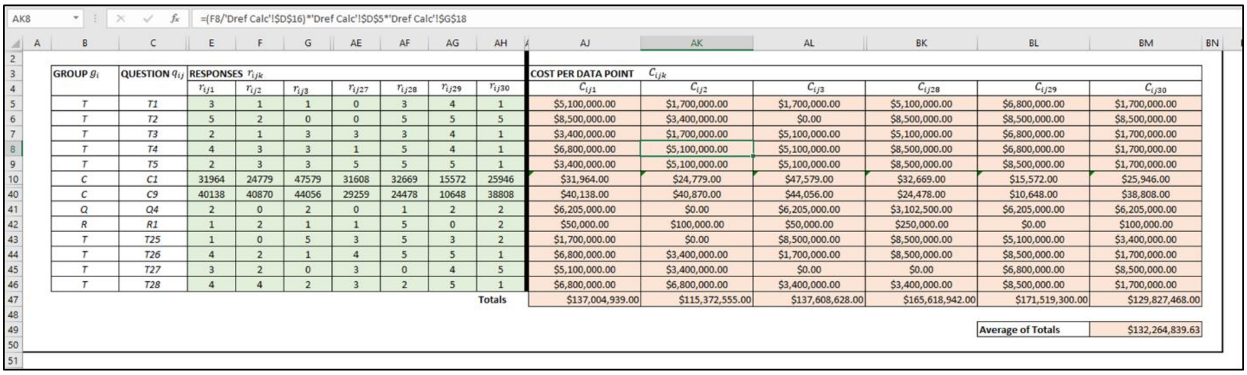
Since the calculations around Asset Reference Data and C_T have already been shown in Figures 4.4 and 4.6 respectively, these foundational elements will not be repeated. Rather, the logic will proceed from that onto stochastic elements.

To calculate the stochastic model, a few additional worksheets have been added, as follows:

- The worksheet “rijk_to_USD” calculates the individual costs per survey question, as explained in Section 4.5.3.3.
- The worksheet “RNG” calculates pseudo-costs $\hat{C}_{EIQp}$
- The worksheet called “CoPEIQ_Stoch” is the presentation of the stochastic calculation $CoPEIQ_s$

Each of these is discussed below and illustrated in Figures 4.9, 4.10 and 4.11.

Figure 4.9 shows part of the worksheet “rijk_to_USD”. Note that column D, Question Text, has been hidden in the interest of graphical expediency, as have been columns H to AD (survey data), columns AM to BJ (C_{ijk} data) and rows 11 to 39. The cursor is on Cell AK8 to demonstrate the calculation of C_{T42}, or the cost calculation of the second response to the fourth question.



AK8 =(F8/'Dref Calc'!$D$16)*'Dref Calc'!$D$5*'Dref Calc'!$G$18

	A	B	C	E	F	G	AE	AF	AG	AH	AJ	AK	AL	BK	BL	BM	BN
2																	
3		GROUP g_i	QUESTION q_{ij}	RESPONSES r_{ijk}							COST PER DATA POINT	C_{ijk}					
4				r_{ij1}	r_{ij2}	r_{ij3}	r_{ij27}	r_{ij28}	r_{ij29}	r_{ij30}	C_{ij1}	C_{ij2}	C_{ij3}	C_{ij28}	C_{ij29}	C_{ij30}	
5		T	T1	3	1	1	0	3	4	1	$5,100,000.00	$1,700,000.00	$1,700,000.00	$5,100,000.00	$6,800,000.00	$1,700,000.00	
6		T	T2	5	2	0	0	5	5	5	$8,500,000.00	$3,400,000.00	$0.00	$8,500,000.00	$8,500,000.00	$8,500,000.00	
7		T	T3	2	1	3	3	3	4	1	$3,400,000.00	$1,700,000.00	$5,100,000.00	$5,100,000.00	$6,800,000.00	$1,700,000.00	
8		T	T4	4	3	3	1	5	4	1	$6,800,000.00	$5,100,000.00	$5,100,000.00	$8,500,000.00	$6,800,000.00	$1,700,000.00	
9		T	T5	2	3	3	5	5	5	1	$3,400,000.00	$5,100,000.00	$5,100,000.00	$8,500,000.00	$8,500,000.00	$1,700,000.00	
10		C	C1	31964	24779	47579	31608	32669	15572	25946	$31,964.00	$24,779.00	$47,579.00	$32,669.00	$15,572.00	$25,946.00	
40		C	C9	40138	40870	44056	29259	24478	10648	38808	$40,138.00	$40,870.00	$44,056.00	$24,478.00	$10,648.00	$38,808.00	
41		Q	Q4	2	0	2	0	1	2	2	$6,205,000.00	$0.00	$6,205,000.00	$3,102,500.00	$6,205,000.00	$6,205,000.00	
42		R	R1	1	2	1	1	5	0	2	$50,000.00	$100,000.00	$50,000.00	$250,000.00	$0.00	$100,000.00	
43		T	T25	1	0	5	3	5	3	2	$1,700,000.00	$0.00	$8,500,000.00	$8,500,000.00	$5,100,000.00	$3,400,000.00	
44		T	T26	4	2	1	4	5	5	1	$6,800,000.00	$3,400,000.00	$1,700,000.00	$8,500,000.00	$8,500,000.00	$1,700,000.00	
45		T	T27	3	2	0	3	0	4	5	$5,100,000.00	$3,400,000.00	$0.00	$0.00	$6,800,000.00	$8,500,000.00	
46		T	T28	4	4	2	3	2	5	1	$6,800,000.00	$6,800,000.00	$3,400,000.00	$3,400,000.00	$8,500,000.00	$1,700,000.00	
47										Totals	$137,004,939.00	$115,372,555.00	$137,608,628.00	$165,618,942.00	$171,519,300.00	$129,827,468.00	
48																	
49															Average of Totals	$132,264,839.63	
50																	
51																	

Figure 4.9 – Calculating the Specific Cost C_{T42}

Figure 4.10 shows part of the worksheet “RNG”, where pseudo-cost data is generated, based on the statistics $\bar{x}_{ij}$ and S_{ij} for each question. Note that columns BHD to NWD (the pseudo-costs for pseudo-respondents 4 to 9998) have been hidden. The cursor is on Cell BR5 to demonstrate the calculation of $\hat{C}_{111}$, or the cost calculation of the first pseudo-respondent to the first question.

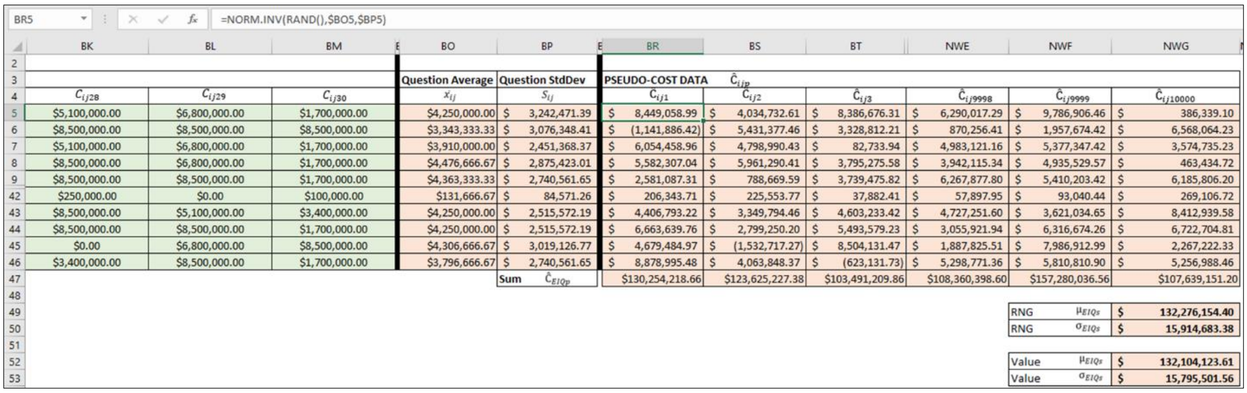
Columns BO and BP calculate the sample mean and standard deviation respectively of the sample of survey data in row 5. Since the survey data is a sample, the Excel function STD.S is used in this case.

The cursor is on Cell BR5 to demonstrate the calculation of $\hat{C}_{111}$, or the cost calculation of the first pseudo-respondent to the first. As is shown, this is again done using the NORM.INV function of Excel, but using the statistics each question, in this example $\bar{x}_{T1}$ and S_{T1}.

Consistent with the cell colour convention, the outputs (red) of the calculation C_{ijk} shown in Figure 4.9, have now become inputs (green), as shown in column BM.

Rows 10 to 41 have also been hidden, to show the first and last 5 questions of the questionnaire. This is done to show the calculations in rows below the matrix. Cell BR47 contains the sum of the cells BR5 to BR46, which has been defined as $\hat{C}_{EIQ1}$. Proceeding along the logic in Section 4.5.3.3, cell NWG49 contains the mean of the population of sums $\hat{C}_{EIQp}$ BR47 to NWG47 and cell MWG50 the population standard deviation of the same population. Since $\hat{C}_{EIQp}$ is a population, the EXCEL function STDEV.P is used in this case.

In cells NWG52 and 53, the values for μ_{EIQs} and σ_{EIQs} have been copied to use in subsequent analysis. This has been done after the initial analysis to compensate for the Excel characteristic to re-calculate randomly generated data every time a command is executed. The values in Figure 4.6, it may be seen, are in the same order of magnitude.

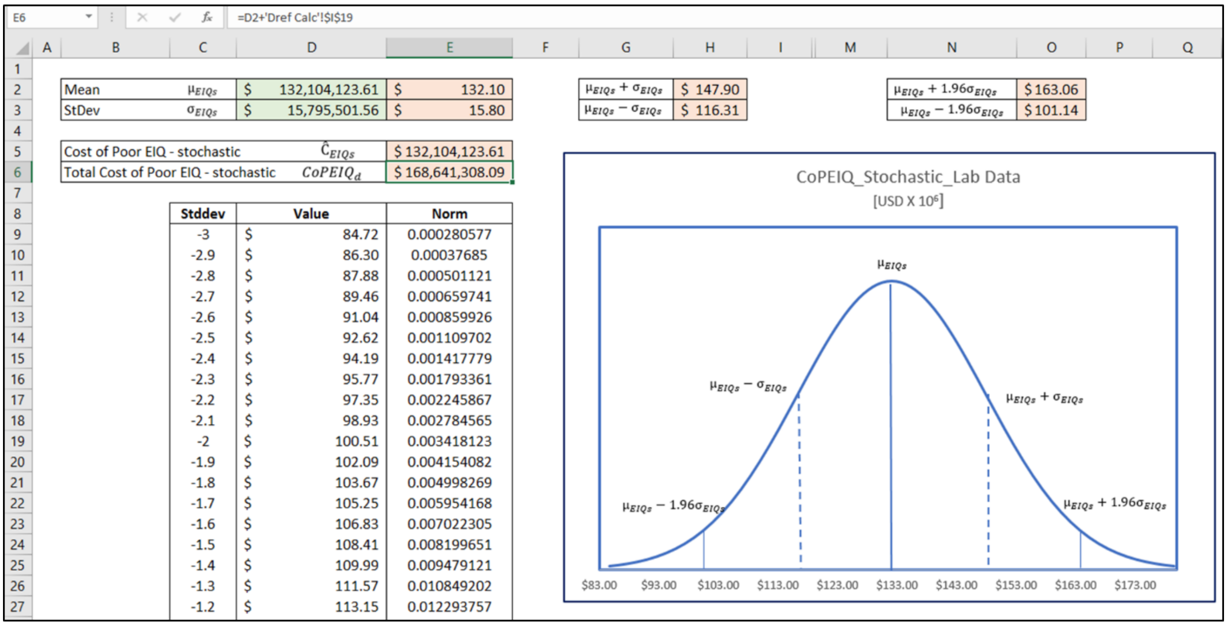


BR5 =NORM.INV(RAND(),$BO5,$BP5)

	BK	BL	BM	BO	BP	BR	BS	BT	NWE	NWF	NWG
2											
3				Question Average	Question StdDev	PSEUDO-COST DATA	$\hat{C}_{ijp}$				
4	C_{ij28}	C_{ij29}	C_{ij30}	x_{ij}	S_{ij}	$\hat{C}_{ij1}$	$\hat{C}_{ij2}$	$\hat{C}_{ij3}$	$\hat{C}_{ij9998}$	$\hat{C}_{ij9999}$	$\hat{C}_{ij10000}$
5	$5,100,000.00	$6,800,000.00	$1,700,000.00	$4,250,000.00	$ 3,242,471.39	$ 8,449,058.99	$ 4,034,732.61	$ 8,386,676.31	$ 6,290,017.29	$ 9,786,906.46	$ 386,339.10
6	$8,500,000.00	$8,500,000.00	$8,500,000.00	$3,343,333.33	$ 3,076,348.41	$ (1,141,886.42)	$ 5,431,377.46	$ 3,328,812.21	$ 870,256.41	$ 1,957,674.42	$ 6,568,064.23
7	$5,100,000.00	$6,800,000.00	$1,700,000.00	$3,910,000.00	$ 2,451,368.37	$ 6,054,458.96	$ 4,798,990.43	$ 82,733.94	$ 4,983,121.16	$ 5,377,347.42	$ 3,574,735.23
8	$8,500,000.00	$6,800,000.00	$1,700,000.00	$4,476,666.67	$ 2,875,423.01	$ 5,582,307.04	$ 5,961,290.41	$ 3,795,275.58	$ 3,942,115.34	$ 4,935,529.57	$ 463,434.72
9	$8,500,000.00	$8,500,000.00	$1,700,000.00	$4,363,333.33	$ 2,740,561.65	$ 2,581,087.31	$ 788,669.59	$ 3,739,475.82	$ 6,267,877.80	$ 5,410,203.42	$ 6,185,806.20
42	$250,000.00	$0.00	$100,000.00	$131,666.67	$ 84,571.26	$ 206,343.71	$ 225,553.77	$ 37,882.41	$ 57,897.95	$ 93,040.44	$ 269,106.72
43	$8,500,000.00	$5,100,000.00	$3,400,000.00	$4,250,000.00	$ 2,515,572.19	$ 4,406,793.22	$ 3,349,794.46	$ 4,603,233.42	$ 4,727,251.60	$ 3,621,034.65	$ 8,412,939.58
44	$8,500,000.00	$8,500,000.00	$1,700,000.00	$4,250,000.00	$ 2,515,572.19	$ 6,663,639.76	$ 2,799,250.20	$ 5,493,579.23	$ 3,055,921.94	$ 6,316,674.26	$ 6,722,704.81
45	$0.00	$6,800,000.00	$8,500,000.00	$4,306,666.67	$ 3,019,126.77	$ 4,679,484.97	$ (1,532,717.27)	$ 8,504,131.47	$ 1,887,825.51	$ 7,986,912.99	$ 2,267,222.33
46	$3,400,000.00	$8,500,000.00	$1,700,000.00	$3,796,666.67	$ 2,740,561.65	$ 8,878,995.48	$ 4,063,848.37	$ (623,131.73)	$ 5,298,771.36	$ 5,810,810.90	$ 5,256,988.46
47					Sum $\hat{C}_{EIQp}$	$130,254,218.66	$123,625,227.38	$103,491,209.86	$108,360,398.60	$157,280,036.56	$107,639,151.20
48											
49										RNG μ_{EIQs}	$ 132,276,154.40
50										RNG σ_{EIQs}	$ 15,914,683.38
51											
52										Value μ_{EIQs}	$ 132,104,123.61
53										Value σ_{EIQs}	$ 15,795,501.56

Figure 4.10 – Calculating Stochastic CoPEIQ $CoPEIQ_s$

Figure 4.11 shows part of the worksheet "CoPEIQ_Stoch", where the additional parameters for $CoPEIQ_s$ and the graphical representation are developed.



E6 =D2+'Dref Calc'!$I$19

	B	C	D	E	G	H	N	O
2	Mean	μ_{EIQs}	$ 132,104,123.61	$ 132.10	$\mu_{EIQs}+\sigma_{EIQs}$	$ 147.90	$\mu_{EIQs}+1.96\sigma_{EIQs}$	$ 163.06
3	StDev	σ_{EIQs}	$ 15,795,501.56	$ 15.80	$\mu_{EIQs}-\sigma_{EIQs}$	$ 116.31	$\mu_{EIQs}-1.96\sigma_{EIQs}$	$ 101.14
4								
5	Cost of Poor EIQ - stochastic		$\hat{C}_{EIQs}$	$ 132,104,123.61				
6	Total Cost of Poor EIQ - stochastic		$CoPEIQ_d$	$ 168,641,308.09				

	Stddev	Value	Norm
8	Stddev	Value	Norm
9	-3	$ 84.72	0.000280577
10	-2.9	$ 86.30	0.00037685
11	-2.8	$ 87.88	0.000501121
12	-2.7	$ 89.46	0.000659741
13	-2.6	$ 91.04	0.000859926
14	-2.5	$ 92.62	0.001109702
15	-2.4	$ 94.19	0.001417779
16	-2.3	$ 95.77	0.001793361
17	-2.2	$ 97.35	0.002245867
18	-2.1	$ 98.93	0.002784565
19	-2	$ 100.51	0.003418123
20	-1.9	$ 102.09	0.004154082
21	-1.8	$ 103.67	0.004998269
22	-1.7	$ 105.25	0.005954168
23	-1.6	$ 106.83	0.007022305
24	-1.5	$ 108.41	0.008199651
25	-1.4	$ 109.99	0.009479121
26	-1.3	$ 111.57	0.010849202
27	-1.2	$ 113.15	0.012293757

CoPEIQ_Stochastic_Lab Data
[USD X 10⁶]
μ_{EIQs}
$\mu_{EIQs}-\sigma_{EIQs}$
$\mu_{EIQs}+\sigma_{EIQs}$
$\mu_{EIQs}-1.96\sigma_{EIQs}$
$\mu_{EIQs}+1.96\sigma_{EIQs}$
$83.00
$93.00
$103.00
$113.00
$123.00
$133.00
$143.00
$153.00
$163.00
$173.00

Figure 4.11 – Stochastic CoPEIQ Parameters

In Figure 4.11, the following comments are relevant:

- Cells D2 and D3 have been copied from cells NWG52 and 53 in Figure 4.10.
- Cells E2 and E3 are simply μ_{EIQs} and σ_{EIQs} divided by 1,000,000 to simplify the subsequent calculations from a visualisation point of view.
- The Cost of poor EIQ-stochastic is shown in cell E5 and the addition of C_{VAC} yields the Total Cost of Poor EIQ-Stochastic, $CoPEIQ_S$in Cell E6.
- Cells H2 and H3 calculate the values of plus or minus one standard deviation, and cells O2 and O3, the 5% confidence intervals.
- The table shown in columns C, D and E is the first part of the preparation for the graphical representation shown, based on a standard deviation range of -3 to +3 in increments of 0.1.

This concludes the description of the stochastic model. The next section will briefly discuss the testing of the model.

4.5.3.6 Testing the Stochastic Model

In this final section of the chapter describing the CoPEIQ models, a brief discussion is given about testing the model. It is not intended to evaluate the reliability or validity of the research design in this section, rather the intent is to demonstrate the accuracy of the Excel calculations in the models.

The testing of the model was done using three simple mechanisms:

- The absence of error messages in Excel
- Continuous testing for expected results
- Populating the “Survey data” table with field data results, as described in Chapter 6.

These mechanisms will be briefly described in order.

Absence of Error Messages

As is evident in the preceding figures of this chapter, no error messages are displayed in the results. There is therefore no fundamental mathematical or data format error in the model.

Expected Results

It is remarked at the outset that a mean result of \$132M, or \$168M if C_{VAC} is included, is roughly in the expected range for the hypothetical test data used in this model development. Further, the median was tested throughout the development of the model for order of magnitude. For example, in the worksheet “rijk to USD”, the average of the totals of each column was calculated and found to be within 0.1% of the eventual result of $CoPEIQ_S$.

Survey Data

Finally, the model was populated with field survey data, and calculated a credible result for both $CoPEIQ_d$ and $CoPEIQ_s$ without the need for intervention, as will be described in Chapter 6.

Chapter 5

Formatting Results Presentation

The preceding chapters have brought this study to the point where a survey is developed and survey data is analysed using an appropriate statistical method. This section will consider how to present the data in a summarized format and graphical presentation suitable for the intended audience. Given that the intended audience comprises senior-level managers in an OpCo context, there is little value in presenting every detail of the survey result. The challenge is to summarize and present the data in a format and language that is palatable to this audience. On this basis a taxonomy is developed in this section and, like the survey, tested with a sample audience before being finalized. Finally, some work is done on the specifics of graphical presentation.

The activities described in this chapter are, to some extent, interwoven with those described in Chapters 3 and 4. For example, the review of the context and length of the survey will necessarily affect the structure and length of the taxonomy, and the understanding gained in section 4.2 will drive the final graphical design in this section.

An overview of the proceedings of this chapter is given in Figure 5.1. A taxonomy is first developed, tested and finalized, after which the graphical design is formalized.

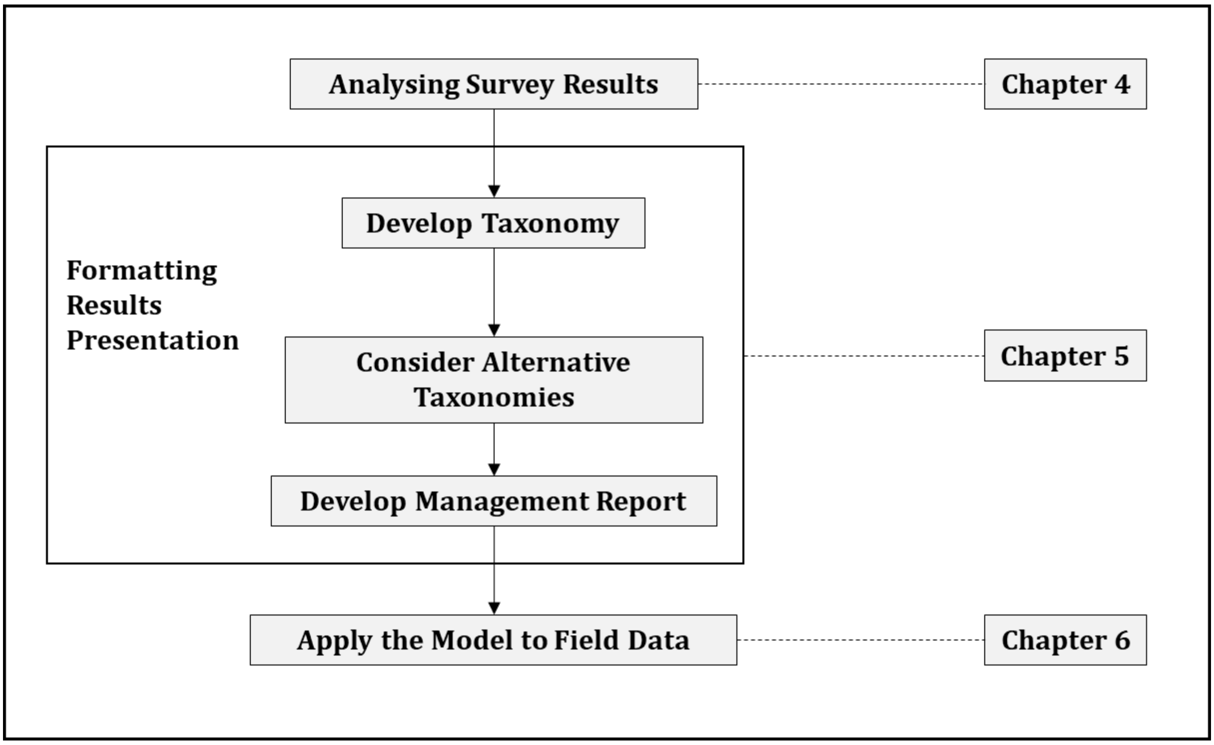

Analysing Survey Results
Chapter 4
Develop Taxonomy
Formatting Results Presentation
Consider Alternative Taxonomies
Chapter 5
Develop Management Report
Apply the Model to Field Data
Chapter 6

Figure 5.1 - Formatting Results Presentation Details

5.1 The Taxonomy

Before delving into the development of the taxonomy, it is useful to carry out a quick review of the literature, which yields interesting insights.

Nickerson et al. (2013) provide a useful methodology for the development of taxonomy. After discussing the merits of using the word "taxonomy", they list the qualitative attributes that a useful taxonomy needs to have. These are listed below, together with a short commentary about the relevance to this study in each bullet point.

- It is concise. If not, it may exceed the cognitive load of the researcher and the research subjects. The length of the Survey had already been reduced after the review cycle; this necessarily shortened the taxonomy to a very focused list.

- It is robust, meaning that it should "clearly differentiate between the objects of interest". This requirement was, likewise, an important consideration during the "mutually exclusive" review of the Survey. The challenge for the taxonomy was to ensure that the clear differentiation remained through the levels of the taxonomy.

- It is comprehensive. All known objects within the domain should be included. This requirement pointed the research to a very crisp definition of the scope of the research. The very approach of extracting a divergent list of IEs from a wide literature sample had the intent of not missing any relevant IE, thereby laying the foundation for this requirement in the taxonomy.

- It is extendible, meaning that it "should allow for inclusion of additional dimensions". The structure of the survey has already excluded Projects from the scope and many attributes have been defined and classified, enabling considerable expansion in several dimensions in future iterations of the taxonomy.

- It is explanatory, or "provides useful explanations of the nature of the objects". Considerable effort and review was undertaken to derive explanatory research questions from the identified IE's. The challenge of the differences in transatlantic English was particularly interesting.

In the same paper, Nickerson et al. (2013) propose an initial process of framing questions during a taxonomy design. These framing questions provide the context for the taxonomy and lead the researcher to fundamental dimensions of the taxonomy. These are repeated in Figure 5.2.

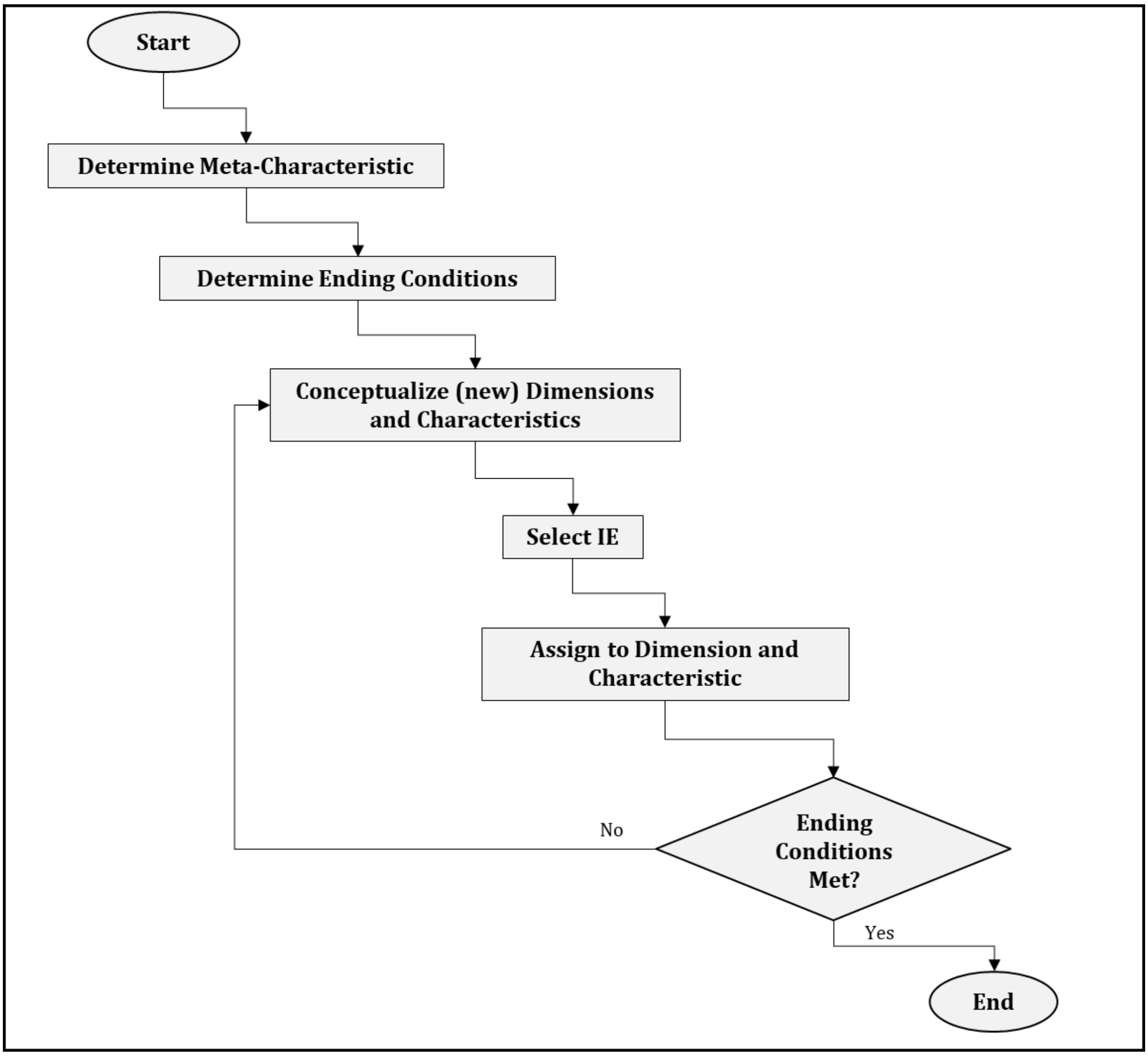

Start
Determine Meta-Characteristic
Determine Ending Conditions
Conceptualize (new) Dimensions and Characteristics
Select IE
Assign to Dimension and Characteristic
Ending Conditions Met?
No
Yes
End

Figure 5.2 - Taxonomy Development Process – adopted from Nickerson et al. (2013)

Gackowski (2009) studied the effectiveness of surveys about information quality from a teleological perspective. He concluded, amongst other things, the need for structured questions. This conclusion has been addressed in Chapter 3; the taxonomy is an extension of that structure.

Eppler & Helfert (2004) had in fact developed a taxonomy for data quality cost, based on a literature review. This taxonomy is shown in Table 5.1.

Table 5.1 - A Data Quality Cost Taxonomy (adopted from Eppler & Helfert (2004))

Data Quality Costs	Costs caused by low data quality	Direct costs	Verification costs
			Re-entry costs
			Compensation costs
		Indirect costs	Cost based on lower reputation
			Costs based on wrong decisions or actions
			Sunk investment costs
	Costs of improving or assuring data quality	Prevention costs	Training costs
			Monitoring costs
			Standard development and deployment costs
		Detection costs	Analysis costs
			Reporting costs
		Repair costs	Repair planning costs
			Repair implementation costs

These three sources of literature provided a handy toolkit to proceed with the construction of the taxonomy.

5.1.1 Initial Taxonomy

Nickerson et al. (2013) quote Batley (1984) as saying that a common approach is to use three-level taxonomy: conceptual, empirical and indicator levels. This approach is consistent with most causal taxonomies in use in the oil and gas industry, and was consequently adopted for this taxonomy, not least because of the instinctive acceptance of the structure by the intended target audience. The initial design decision taken for the taxonomy was therefore that it should consist of three levels that might be called "conceptual", "empirical" and "indicator".

Given the pre-existing taxonomy by Eppler & Helfert (2004), a logical question might be "why it was not adopted?" The answer will be given in this section; however, this question will be revisited in Section 5.2.

From this premise, the design of the taxonomy commenced from the framework shown in Figure 5.2. Table 5.2 summarizes the rationale.

Table 5.2 - Taxonomy Framing Questions

Framing Question	Response for this Taxonomy		
Identify users	Middle and senior level leaders in OGI		
Determine expected use	Investment decisions at various scales and levels regarding the creation, maintenance or recreation of EI.		
Define the purpose of the taxonomy	The taxonomy summarizes the results of the CoPEIQ survey in a structure that is logical for the intended audience. This means: (i) summarizing to decreasing levels of detail into a logical framework for the audience (ii) providing enough granularity to prioritize remediation according to a relatable framework.		
Determine the meta-characteristic*	Option 1	Standard processes	Selected as a broad indicator of the secondary meta-characteristic. Note however the comments in the text.

Framing Question	Response for this Taxonomy		
	Option 2	Generic EIM activities i.e. find/verify/create/store/ compare	Discounted because of the difficulty of extracting this data
	Option 3	Strategic alignment	Selected as primary meta-characteristic since this is the language of the target audience
	Option 4	Asset Life Cycle	Already discounted due to reasons provided in Section 1.2.
	Option 5	IT System or Architecture	Discounted because there is no 1:1 map between systems and the high variability in system deployment.
	Option 6	Department	Selected as denominator. See discussion in text.
Select the basis for the taxonomy	Alternative 1	Empirical to conceptual (inductive)	Not selected.
	Alternative 2	Conceptual to Empirical (deductive)	Selected due to: - low data volume - good understanding of the domain

*"Meta-characteristic" is defined as the "most comprehensive characteristic that will serve as the basis for the choice of characteristics in the taxonomy". For this study it is deemed to be the "relatable framework".

Based on this fundamental design, the iterations proposed in Figure 5.2 proceeded until a provisional taxonomy was prepared.

The top level of provisional taxonomy is shown in Figure 5.3. The five top elements of the Taxonomy have been chosen on the basis that these elements are deemed foremost in the minds of senior OpCo management, regardless of the business cycle, the specific strategy for the year or the specific terminology in force at that OpCo. It is also aligned with Option 3 in Table 5.2, as being the primary meta-characteristic.

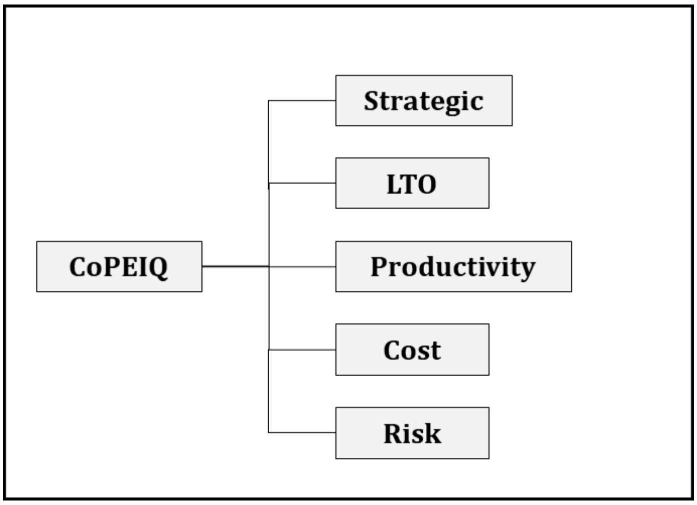


Figure 5.3 - The Initial Taxonomy – Conceptual Level

The intermediate, or "Empirical", level of the taxonomy was derived broadly based on Option 1 – Standard processes. This was initially done based on domain knowledge. More clarity on this level emerged during the validation of the Taxonomy, which is discussed in Section 5.1.2.

From this discussion and further scrutiny of Table 5.2, it is contended here that there can be more than one meta-characteristic. (Specifically, reference is made to a "primary" and "secondary" meta-characteristic. This contention is contrary to the original methodology proposed by Nickerson et al. (2013). The reason for this contention is explained in Section 5.1.2. In addition, the use of the word "Denominator" will also become clear in Section 5.1.2.

Finally, the third "Indicator" Level is the list of Survey questions developed in Chapter 3.

For illustration, Figure 5.4 shows an extract of the Initial Taxonomy.

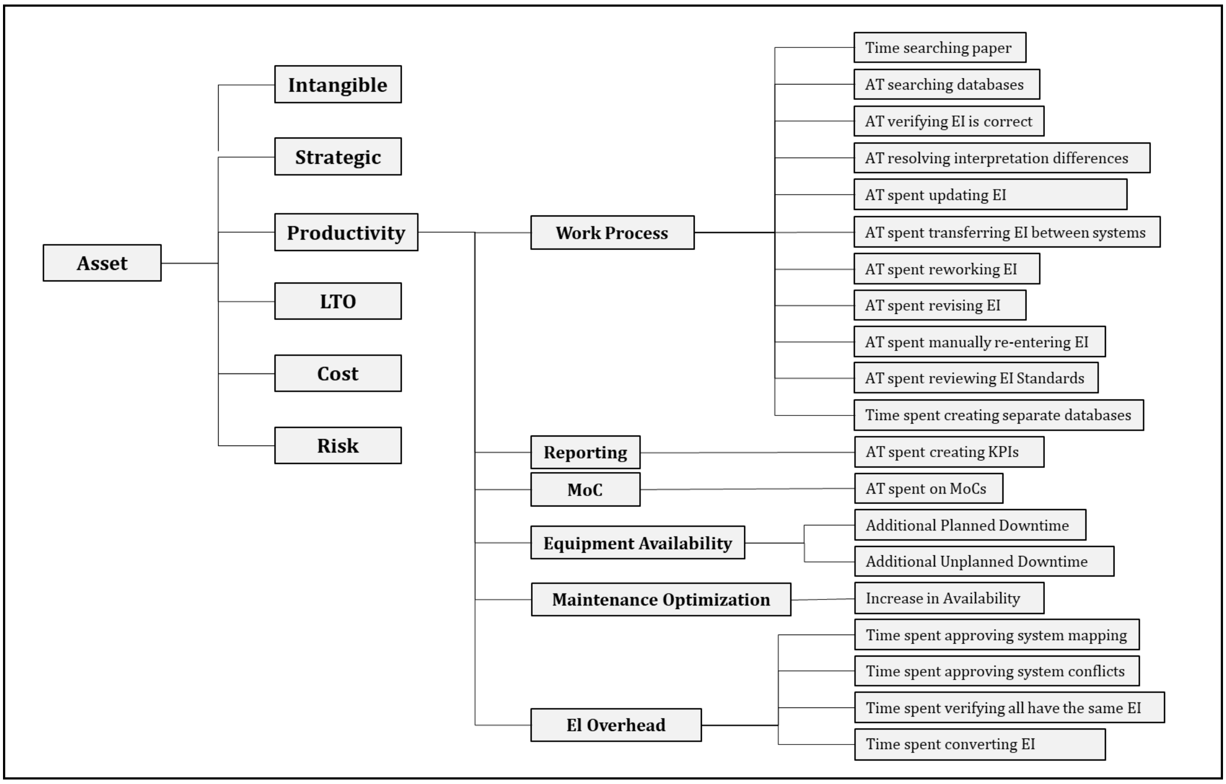

Asset
Intangible
Strategic
Productivity
LTO
Cost
Risk
Work Process
Reporting
MoC
Equipment Availability
Maintenance Optimization
EI Overhead
Time searching paper
AT searching databases
AT verifying EI is correct
AT resolving interpretation differences
AT spent updating EI
AT spent transferring EI between systems
AT spent reworking EI
AT spent revising EI
AT spent manually re-entering EI
AT spent reviewing EI Standards
Time spent creating separate databases
AT spent creating KPIs
AT spent on MoCs
Additional Planned Downtime
Additional Unplanned Downtime
Increase in Availability
Time spent approving system mapping
Time spent approving system conflicts
Time spent verifying all have the same EI
Time spent converting EI

Figure 5.4 - Illustrative Detail of the Initial Taxonomy

The abbreviation "AT" in Figure 5.4 is used to denote "Additional Time". This is done for graphical expediency.

The initial taxonomy is validated in the next section.

5.1.2 Validation of the Initial Taxonomy

Along with the validation of the Survey discussed in Section 3.3, the basis for the Taxonomy and its details were also reviewed during the same time period, but using a different process than the structured interviews used to validate the survey questions. This process is described in this section.

The first question was whether the basis of the taxonomy was in fact valid. This was done to test the efficacy of the using a pre-existing taxonomy, rather than develop a specific taxonomy.

As noted in the previous section, Eppler & Helfert (2004) had already developed a taxonomy. In addition, the American Productivity and Quality Center (APQC) publishes a list of standard processes, amongst which is a list of processes for an OpCo. Considering these alternatives, it was necessary to confirm the basis (or primary meta-characteristic) of the taxonomy.

Four senior executives in three different corporations in the OGI were informally interviewed. They were shown the three options and asked which taxonomy would be most suitable for them to enable an informed and reasonable decision for investing in EIQ. They unanimously selected the organigram taxonomy and provided reasons like "this is stated in the language of business", "APQC is too high level for an operating asset".

The taxonomy provided by Eppler & Helfert (2004) was considered a good generic cost categorization model from the point of view of intrinsic cost of data quality. It was not, however, considered specific enough to enable the prioritization of remediation effort for this thesis. It would be difficult, for instance, for an operations manager to prioritise "re-entry costs" versus "analysis costs", whereas a prioritisation between "engineering process efficiency" and "ethical compliance".

As a result, the original taxonomy was confirmed as the basis and was subsequently refined.

During the reviews, the original terminology for the levels of the taxonomy, namely conceptual, empirical and indicator, were found to be counter-intuitive and difficult to relate to. The titles of these levels were therefore changed to align with the organisational levels with which they are most likely to align. These were entitled "Executive", "Management" and "Tactical".

During the review of the Survey, the most significant structural insight gained was that senior managers were not only interested in the aggregate, or "total", financial impact of poor EIQ, but wanted to see a breakdown of the total in terms of a denominator that would enable them to determine with instinctive ease where to prioritize their remediation efforts. Put another way, they wanted to know where the financial impact was the greatest in their areas of responsibility. The taxonomy needed to be structured in a way that facilitated this. For this purpose, the notion of a "Denominator" is introduced. It could be viewed as an "alternative filter" or "another dimension from which to view the data".

This raises the question of exactly which denominator to use to provide this additional granularity. In Table 5.2 two options were identified: Business Process and Department. The denominator "Discipline" was selected, for the following reasons:

- With reference to Section 3.1.12, the Department is a demographic attribute which will be populated unambiguously as a nominal variable.

- In most OGI contexts, Discipline will correlate unambiguously to organisational unit, which in turn will usually provide a good indication of where to prioritize remediation efforts. For example, if the Department "Electrical" indicated the highest financial impact of poor EIQ, it would follow with reasonable certainty that the Electrical Department would be the starting point of a remediation effort.

- If a specific OGI manager wished to prioritise data remediation efforts on the grounds of business process, the Department will correlate well for that specific context. For example, if the specific asset under review was organised in such a way that maintenance planning was done in one centralized team and the Discipline "Planning" was reported in the survey as having the highest impact, the manager could infer with high certainty that remediation should start with Maintenance Planning.

Because of the preceding, the construct of the taxonomy was enhanced as shown in Figure 5.5. There are two dimensions: the horizontal level shows the three-level Taxonomy with one illustrative example, while the Denominator (Engineering Discipline) is shown on the vertical.

A detailed look will also reveal that the "Executive Level" has been modified. This was in response to feedback during the reviews.

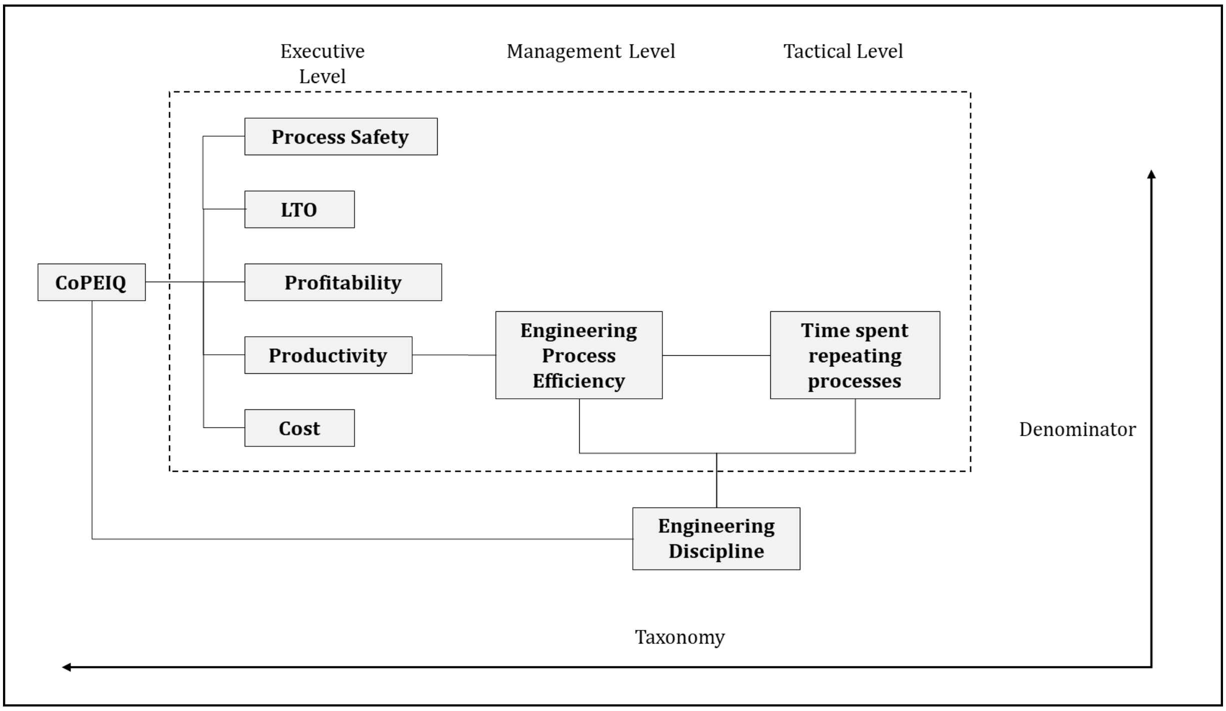


Figure 5.5 - The Final Taxonomy Construct

The decision to use Department as a denominator leads to the question "what exactly is a standardized list of Departments in the OGI?" The answer to this question is remarkably difficult to get, given how fundamental it is to the whole issue of engineering information management and given the wide variety or organisational constructs in the OGI. The concept of "Engineering Discipline" was considered as a useful proxy for Department and investigated. ISO 14224 (2006) refers to Disciplines only in passing: "Maintenance man-hours per discipline (mechanical, electrical, instrument, others)". EPISTLE Part 2 (1999) infers a list of standard engineering disciplines by specifying data handover requirements per Disciplines. This is repeated in Table 5.3.

Table 5.3 - Engineering Disciplines

Engineering Disciplines per EPISTLE Part 2 (1999)
Process Engineering
Instrumentation
Fire and Gas
Telecommunications
Architectural
Electrical
Maintenance
Maintenance Planning
Mechanical
Pipelines[4]
Piping
Corrosion
HVAC
Civil and Structural

EPISTLE Part 2(1999) in fact adds a few additional disciplines into the specification. Since these disciplines are, however, disciplines related specifically to the execution of projects, they have not been included in this analysis. They are Contracts, Planning/Scheduling/Cost Control, Materials Management, Construction, QA/QC, HSE, Loss Control, Fabrication Control.

After further review of the data sets of several OpCo's, a simplified list of Departments was selected. This list is shown below.

- Operations
- Electrical
- Mechanical (Machinery)
- Mechanical (Integrity)
- Instrumentation/Control/Automation
- Maintenance
- Maintenance Planning
- Turnaround
- Process Engineering
- Process Safety
- Corporate Planning/Analysis/Strategy
- Projects
- Engineering Information/Technical Documentation
- Other (please specify)

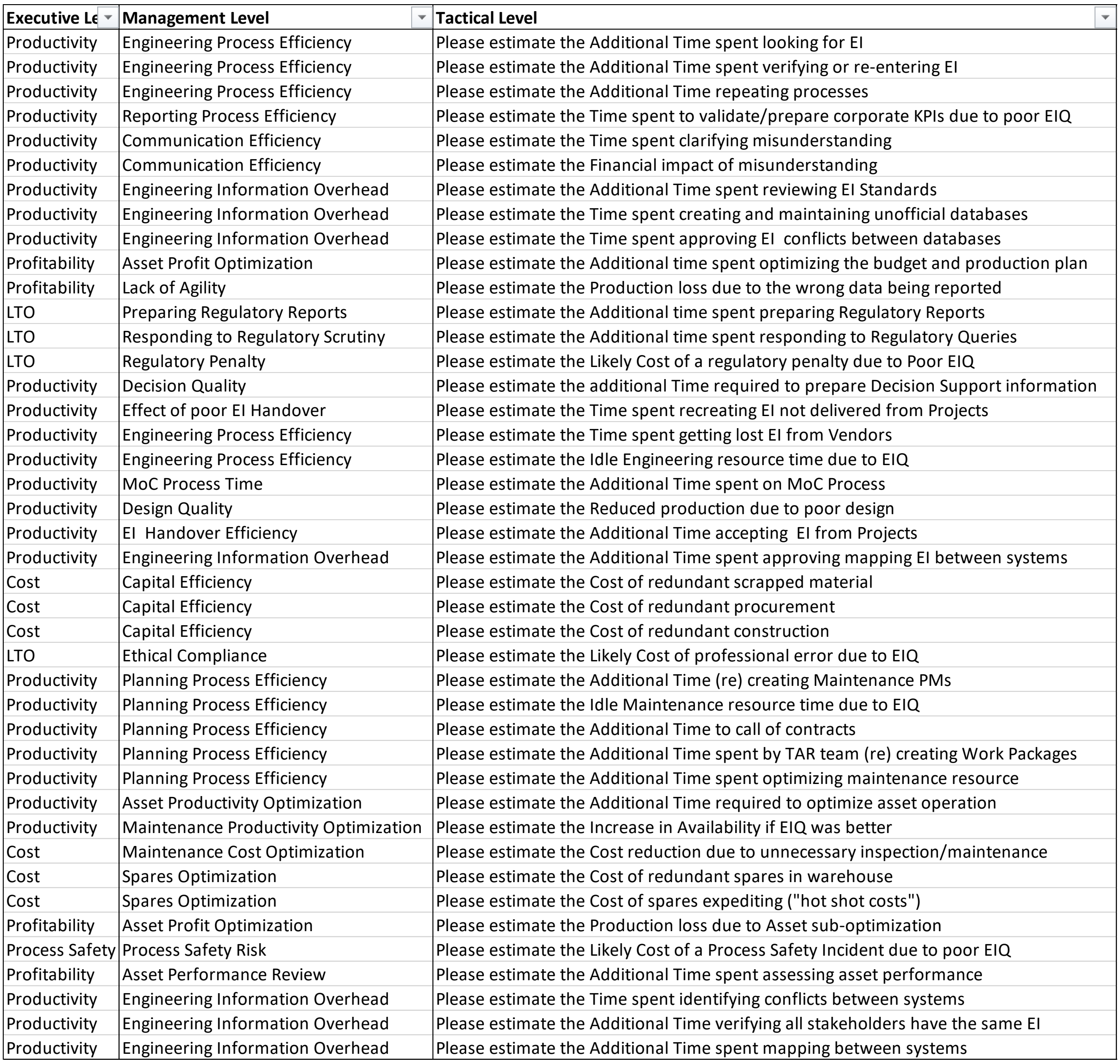
The enhancements found during the review cycle were incorporated into the Final Taxonomy, which follows in the next section.

[4] "Pipelines" are conventionally meant to mean long-distance overland pipes, while "Piping" refers to pipework within OGI plants battery limits

5.1.3 The Final Taxonomy

The result of the preceding Initial Taxonomy design and subsequent enhancements, have led to the Final Taxonomy, which is shown in Table 5.4. This has also been renamed to "Primary Taxonomy", for reasons shown in Section 5.2.

Table 5.4 - Final Taxonomy



Executive Le	Management Level	Tactical Level
Productivity	Engineering Process Efficiency	Please estimate the Additional Time spent looking for EI
Productivity	Engineering Process Efficiency	Please estimate the Additional Time spent verifying or re-entering EI
Productivity	Engineering Process Efficiency	Please estimate the Additional Time repeating processes
Productivity	Reporting Process Efficiency	Please estimate the Time spent to validate/prepare corporate KPIs due to poor EIQ
Productivity	Communication Efficiency	Please estimate the Time spent clarifying misunderstanding
Productivity	Communication Efficiency	Please estimate the Financial impact of misunderstanding
Productivity	Engineering Information Overhead	Please estimate the Additional Time spent reviewing EI Standards
Productivity	Engineering Information Overhead	Please estimate the Time spent creating and maintaining unofficial databases
Productivity	Engineering Information Overhead	Please estimate the Time spent approving EI conflicts between databases
Profitability	Asset Profit Optimization	Please estimate the Additional time spent optimizing the budget and production plan
Profitability	Lack of Agility	Please estimate the Production loss due to the wrong data being reported
LTO	Preparing Regulatory Reports	Please estimate the Additional time spent preparing Regulatory Reports
LTO	Responding to Regulatory Scrutiny	Please estimate the Additional time spent responding to Regulatory Queries
LTO	Regulatory Penalty	Please estimate the Likely Cost of a regulatory penalty due to Poor EIQ
Productivity	Decision Quality	Please estimate the additional Time required to prepare Decision Support information
Productivity	Effect of poor EI Handover	Please estimate the Time spent recreating EI not delivered from Projects
Productivity	Engineering Process Efficiency	Please estimate the Time spent getting lost EI from Vendors
Productivity	Engineering Process Efficiency	Please estimate the Idle Engineering resource time due to EIQ
Productivity	MoC Process Time	Please estimate the Additional Time spent on MoC Process
Productivity	Design Quality	Please estimate the Reduced production due to poor design
Productivity	EI Handover Efficiency	Please estimate the Additional Time accepting EI from Projects
Productivity	Engineering Information Overhead	Please estimate the Additional Time spent approving mapping EI between systems
Cost	Capital Efficiency	Please estimate the Cost of redundant scrapped material
Cost	Capital Efficiency	Please estimate the Cost of redundant procurement
Cost	Capital Efficiency	Please estimate the Cost of redundant construction
LTO	Ethical Compliance	Please estimate the Likely Cost of professional error due to EIQ
Productivity	Planning Process Efficiency	Please estimate the Additional Time (re) creating Maintenance PMs
Productivity	Planning Process Efficiency	Please estimate the Idle Maintenance resource time due to EIQ
Productivity	Planning Process Efficiency	Please estimate the Additional Time to call of contracts
Productivity	Planning Process Efficiency	Please estimate the Additional Time spent by TAR team (re) creating Work Packages
Productivity	Planning Process Efficiency	Please estimate the Additional Time spent optimizing maintenance resource
Productivity	Asset Productivity Optimization	Please estimate the Additional Time required to optimize asset operation
Productivity	Maintenance Productivity Optimization	Please estimate the Increase in Availability if EIQ was better
Cost	Maintenance Cost Optimization	Please estimate the Cost reduction due to unnecessary inspection/maintenance
Cost	Spares Optimization	Please estimate the Cost of redundant spares in warehouse
Cost	Spares Optimization	Please estimate the Cost of spares expediting ("hot shot costs")
Profitability	Asset Profit Optimization	Please estimate the Production loss due to Asset sub-optimization
Process Safety	Process Safety Risk	Please estimate the Likely Cost of a Process Safety Incident due to poor EIQ
Profitability	Asset Performance Review	Please estimate the Additional Time spent assessing asset performance
Productivity	Engineering Information Overhead	Please estimate the Time spent identifying conflicts between systems
Productivity	Engineering Information Overhead	Please estimate the Additional Time verifying all stakeholders have the same EI
Productivity	Engineering Information Overhead	Please estimate the Additional Time spent mapping between systems

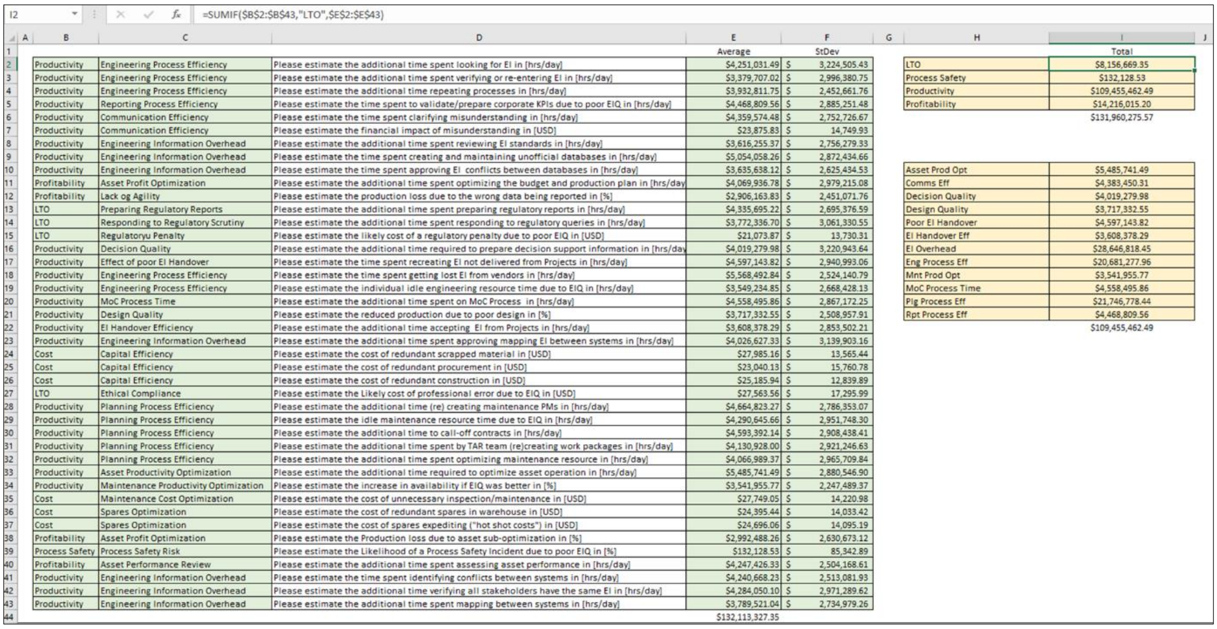
This concludes the development of the taxonomy. Its application to the laboratory data is discussed in the next section.

5.1.3.1 Demonstrating the Final Taxonomy

Having selected the final taxonomy, this section will describe the application of the taxonomy, using same results obtained in the laboratory model demonstrated in Section 4.5.3.

The approach has simply been to apply Excel's SUMIF function to the C_{ijk} data shown in Figure 4.9 and applying Excel's standard "Pareto" to the table so derived. This is done for the Executive Level and at the Management level for the highest subset of data in the Executive category.

To demonstrate this method, an additional worksheet entitled “Prioritisation” has been added to the development worksheet, of which an extract is shown in Figure 5.6. The cursor is on cell I2 to demonstrate the use of the SUMIF function in the formula bar.



I2

=SUMIF($B$2:$B$43,"LTO",$E$2:$E$43)

	A	B	C	D	E	F	G	H	I	J
1					Average	StDev			Total	
2		Productivity	Engineering Process Efficiency	Please estimate the additional time spent looking for EI in [hrs/day]	$4,251,031.49	$ 3,224,505.43		LTO	$8,156,669.35	
3		Productivity	Engineering Process Efficiency	Please estimate the additional time spent verifying or re-entering EI in [hrs/day]	$3,379,707.02	$ 2,996,380.75		Process Safety	$132,128.53	
4		Productivity	Engineering Process Efficiency	Please estimate the additional time repeating processes in [hrs/day]	$3,932,811.75	$ 2,452,661.76		Productivity	$109,455,462.49	
5		Productivity	Reporting Process Efficiency	Please estimate the time spent to validate/prepare corporate KPIs due to poor EIQ in [hrs/day]	$4,468,809.56	$ 2,885,251.48		Profitability	$14,216,015.20	
6		Productivity	Communication Efficiency	Please estimate the time spent clarifying misunderstanding in [hrs/day]	$4,359,574.48	$ 2,752,726.67			$131,960,275.57	
7		Productivity	Communication Efficiency	Please estimate the financial impact of misunderstanding in [USD]	$23,875.83	$ 14,749.93				
8		Productivity	Engineering Information Overhead	Please estimate the additional time spent reviewing EI standards in [hrs/day]	$3,616,255.37	$ 2,756,279.33				
9		Productivity	Engineering Information Overhead	Please estimate the time spent creating and maintaining unofficial databases in [hrs/day]	$5,054,058.26	$ 2,872,434.66				
10		Productivity	Engineering Information Overhead	Please estimate the time spent approving EI conflicts between databases in [hrs/day]	$3,635,638.12	$ 2,625,434.53		Asset Prod Opt	$5,485,741.49	
11		Profitability	Asset Profit Optimization	Please estimate the additional time spent optimizing the budget and production plan in [hrs/day	$4,069,936.78	$ 2,979,215.08		Comms Eff	$4,383,450.31	
12		Profitability	Lack og Agility	Please estimate the production loss due to the wrong data being reported in [%]	$2,906,163.83	$ 2,451,071.76		Decision Quality	$4,019,279.98	
13		LTO	Preparing Regulatory Reports	Please estimate the additional time spent preparing regulatory reports in [hrs/day]	$4,335,695.22	$ 2,695,376.59		Design Quality	$3,717,332.55	
14		LTO	Responding to Regulatory Scrutiny	Please estimate the additional time spent responding to regulatory queries in [hrs/day]	$3,772,336.70	$ 3,061,330.55		Poor EI Handover	$4,597,143.82	
15		LTO	Regulatoryu Penalty	Please estimate the likely cost of a regulatory penalty due to poor EIQ in [USD]	$21,073.87	$ 13,730.31		EI Handover Eff	$3,608,378.29	
16		Productivity	Decision Quality	Please estimate the additional time required to prepare decision support information in [hrs/day	$4,019,279.98	$ 3,220,943.64		EI Overhead	$28,646,818.45	
17		Productivity	Effect of poor EI Handover	Please estimate the time spent recreating EI not delivered from Projects in [hrs/day]	$4,597,143.82	$ 2,940,993.06		Eng Process Eff	$20,681,277.96	
18		Productivity	Engineering Process Efficiency	Please estimate the time spent getting lost EI from vendors in [hrs/day]	$5,568,492.84	$ 2,524,140.79		Mnt Prod Opt	$3,541,955.77	
19		Productivity	Engineering Process Efficiency	Please estimate the individual idle engineering resource time due to EIQ in [hrs/day]	$3,549,234.85	$ 2,668,428.13		MoC Process Time	$4,558,495.86	
20		Productivity	MoC Process Time	Please estimate the additional time spent on MoC Process in [hrs/day]	$4,558,495.86	$ 2,867,172.25		Plg Process Eff	$21,746,778.44	
21		Productivity	Design Quality	Please estimate the reduced production due to poor design in [%]	$3,717,332.55	$ 2,508,957.91		Rpt Process Eff	$4,468,809.56	
22		Productivity	EI Handover Efficiency	Please estimate the additional time accepting EI from Projects in [hrs/day]	$3,608,378.29	$ 2,853,502.21			$109,455,462.49	
23		Productivity	Engineering Information Overhead	Please estimate the additional time spent approving mapping EI between systems in [hrs/day]	$4,026,627.33	$ 3,139,903.16				
24		Cost	Capital Efficiency	Please estimate the cost of redundant scrapped material in [USD]	$27,985.16	$ 13,565.44				
25		Cost	Capital Efficiency	Please estimate the cost of redundant procurement in [USD]	$23,040.13	$ 15,760.78				
26		Cost	Capital Efficiency	Please estimate the cost of redundant construction in [USD]	$25,185.94	$ 12,839.89				
27		LTO	Ethical Compliance	Please estimate the Likely cost of professional error due to EIQ in [USD]	$27,563.56	$ 17,295.99				
28		Productivity	Planning Process Efficiency	Please estimate the additional time (re) creating maintenance PMs in [hrs/day]	$4,664,823.27	$ 2,786,353.07				
29		Productivity	Planning Process Efficiency	Please estimate the idle maintenance resource time due to EIQ in [hrs/day]	$4,290,645.66	$ 2,951,748.30				
30		Productivity	Planning Process Efficiency	Please estimate the additional time to call-off contracts in [hrs/day]	$4,593,392.14	$ 2,908,438.41				
31		Productivity	Planning Process Efficiency	Please estimate the additional time spent by TAR team (re)creating work packages in [hrs/day]	$4,130,928.00	$ 2,921,246.63				
32		Productivity	Planning Process Efficiency	Please estimate the additional time spent optimizing maintenance resource in [hrs/day]	$4,066,989.37	$ 2,965,709.84				
33		Productivity	Asset Productivity Optimization	Please estimate the additional time required to optimize asset operation in [hrs/day]	$5,485,741.49	$ 2,880,546.90				
34		Productivity	Maintenance Productivity Optimization	Please estimate the increase in availability if EIQ was better in [%]	$3,541,955.77	$ 2,247,489.37				
35		Cost	Maintenance Cost Optimization	Please estimate the cost of unnecessary inspection/maintenance in [USD]	$27,749.05	$ 14,220.98				
36		Cost	Spares Optimization	Please estimate the cost of redundant spares in warehouse in [USD]	$24,395.44	$ 14,033.42				
37		Cost	Spares Optimization	Please estimate the cost of spares expediting ("hot shot costs") in [USD]	$24,696.06	$ 14,095.19				
38		Profitability	Asset Profit Optimization	Please estimate the Production loss due to asset sub-optimization in [%]	$2,992,488.26	$ 2,630,673.12				
39		Process Safety	Process Safety Risk	Please estimate the Likelihood of a Process Safety Incident due to poor EIQ in [%]	$132,128.53	$ 85,342.89				
40		Profitability	Asset Performance Review	Please estimate the additional time spent assessing asset performance in [hrs/day]	$4,247,426.33	$ 2,504,168.61				
41		Productivity	Engineering Information Overhead	Please estimate the time spent identifying conflicts between systems in [hrs/day]	$4,240,668.23	$ 2,513,081.93				
42		Productivity	Engineering Information Overhead	Please estimate the additional time verifying all stakeholders have the same EI in [hrs/day]	$4,284,050.10	$ 2,971,289.62				
43		Productivity	Engineering Information Overhead	Please estimate the additional time spent mapping between systems in [hrs/day]	$3,789,521.04	$ 2,734,979.26				
44					$132,113,327.35					

Figure 5.6 - Applying the Taxonomy to Laboratory Data

The results shown in Figure 5.6 are presented in Pareto charts for the Executive and Management levels, as shown in Figures 5.7 and 5.8 respectively. These figures jointly constitute an indication of the highest potential improvement opportunities for the asset, based on the calculated data. For the laboratory data used to develop the model, these opportunities are in Productivity (Figure 5.7) and specifically in the areas EI Overhead, Planning Process Efficiency and Engineering Process Efficiency (Figure 5.8).

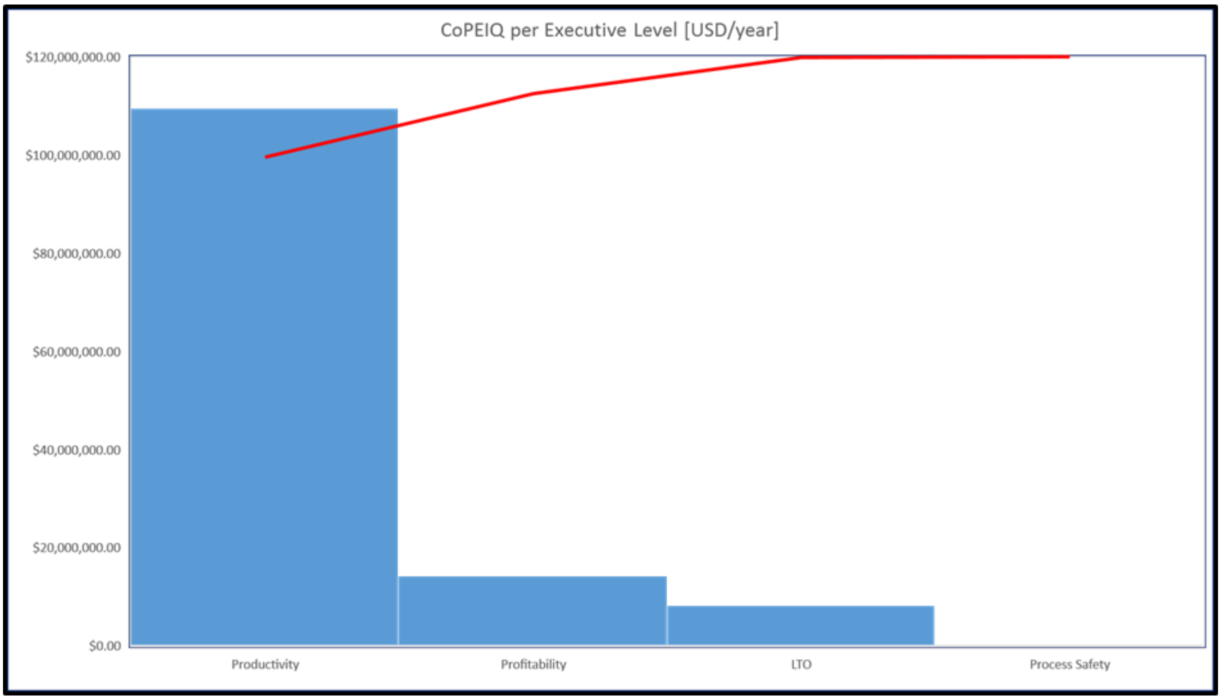


Figure 5.7 - Executive Level Pareto Chart for Laboratory Data

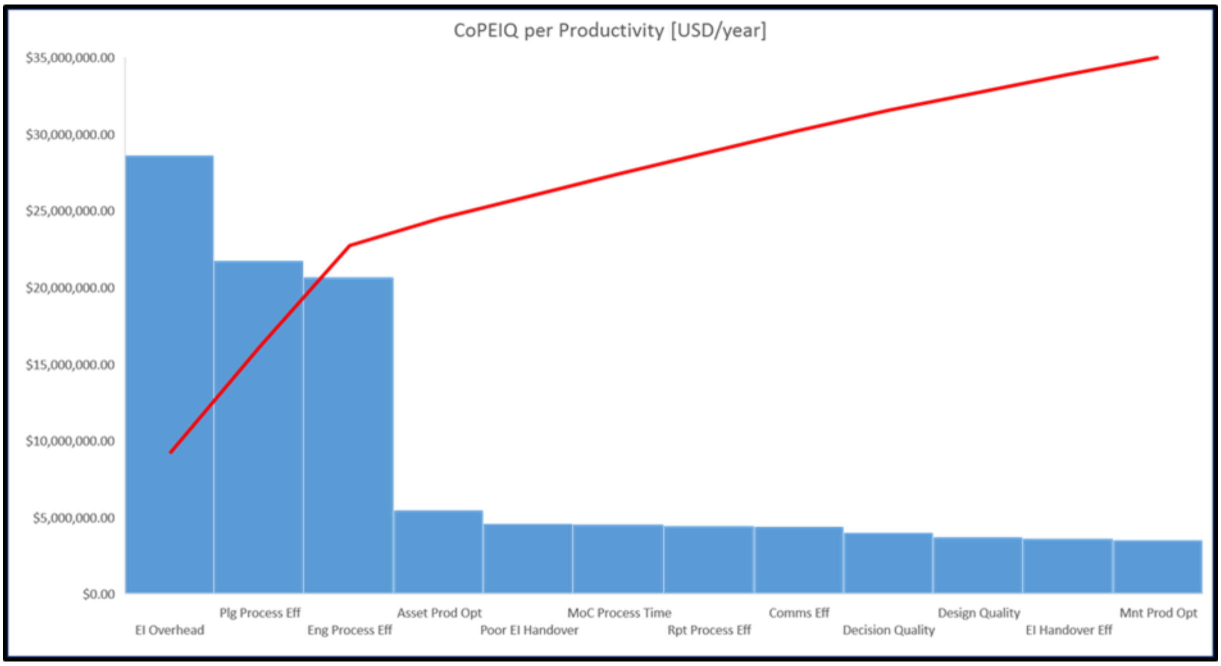


Figure 5.8 - Management Level Pareto Chart for Laboratory Data

To demonstrate how these Pareto charts may be useful in practice, the sponsoring manager may direct initial remediation efforts towards providing validated data to planners and engineers (to improve productivity), and review the EI management processes to understand the high overhead requirement. These Pareto charts are therefore intended to assist the sponsoring manager in prioritising EIQ remediation efforts.

A few alternative taxonomies have also been identified and evaluated. This will be discussed in the next section.

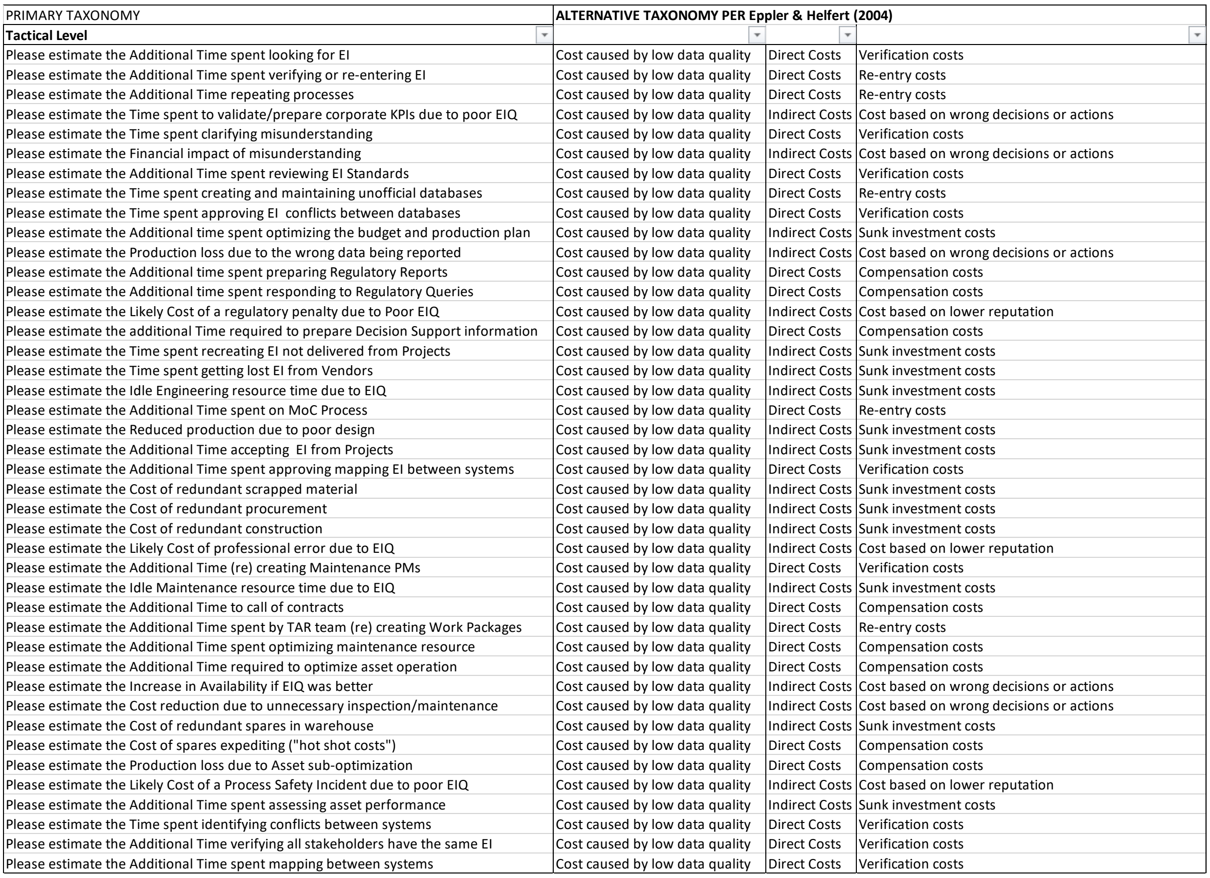
5.2 Alternative Taxonomies

As has become clear in the preceding sections of this chapter, two alternative taxonomies have been considered. This was done to test the efficacy of using existing taxonomies, rather than develop a specific taxonomy. These are discussed in turn below.

5.2.1 The Eppler & Helfert Taxonomy

The cost-based taxonomy developed by Eppler & Helfert (2004) was included as Table 5.1. Although it had been discounted as a primary taxonomy, it was considered a valuable alternative insight into the CoPEIQ problem from a financial point of view. The reasoning was roughly as follows: in the event where the CoPEIQ survey result was significant enough to warrant investment into EIQ remediation, the funding for such remediation would follow the normal Capex/Feasex/Opex processes in force at the OpCo. In that case the classification of cost might be useful. As a result, Eppler & Helfert was mapped against the primary taxonomy, as is shown in Table 5.5. For simplicity, only the mapping to IE (Tactical) Level is shown.

Table 5.5 - Alternative Taxonomy of Eppler & Helfert (2004)



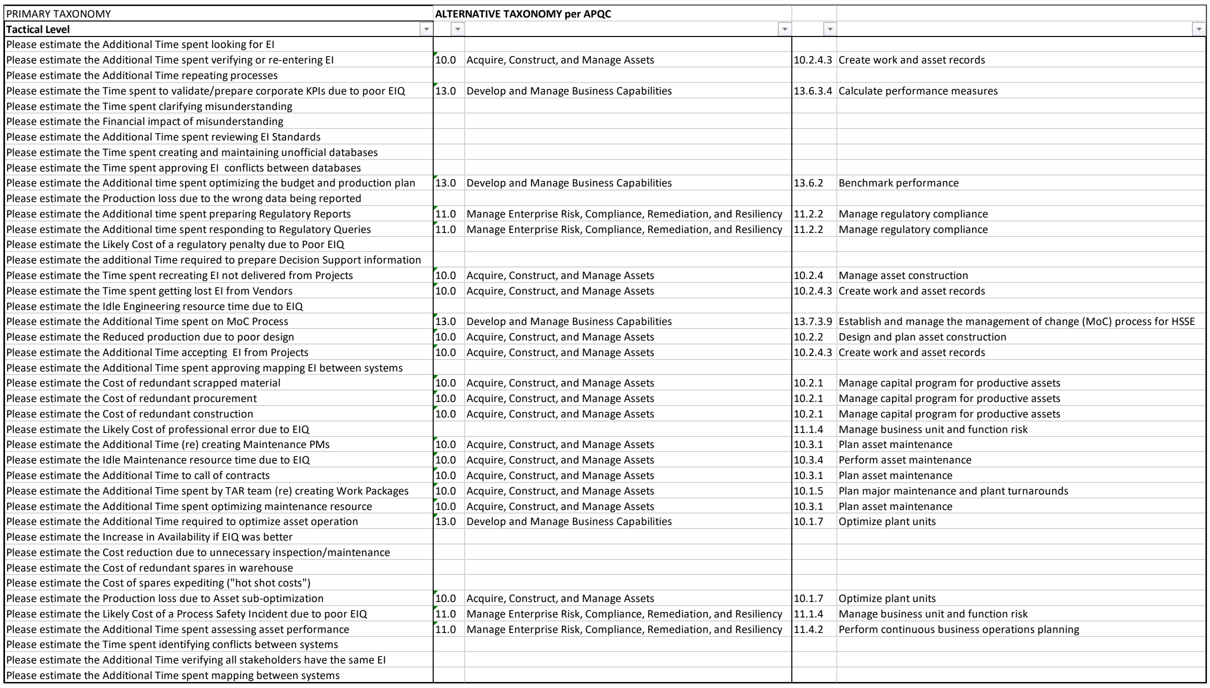
PRIMARY TAXONOMY	ALTERNATIVE TAXONOMY PER Eppler & Helfert (2004)		
Tactical Level			
Please estimate the Additional Time spent looking for EI	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Additional Time spent verifying or re-entering EI	Cost caused by low data quality	Direct Costs	Re-entry costs
Please estimate the Additional Time repeating processes	Cost caused by low data quality	Direct Costs	Re-entry costs
Please estimate the Time spent to validate/prepare corporate KPIs due to poor EIQ	Cost caused by low data quality	Indirect Costs	Cost based on wrong decisions or actions
Please estimate the Time spent clarifying misunderstanding	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Financial impact of misunderstanding	Cost caused by low data quality	Indirect Costs	Cost based on wrong decisions or actions
Please estimate the Additional Time spent reviewing EI Standards	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Time spent creating and maintaining unofficial databases	Cost caused by low data quality	Direct Costs	Re-entry costs
Please estimate the Time spent approving EI conflicts between databases	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Additional time spent optimizing the budget and production plan	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Production loss due to the wrong data being reported	Cost caused by low data quality	Indirect Costs	Cost based on wrong decisions or actions
Please estimate the Additional time spent preparing Regulatory Reports	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Additional time spent responding to Regulatory Queries	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Likely Cost of a regulatory penalty due to Poor EIQ	Cost caused by low data quality	Indirect Costs	Cost based on lower reputation
Please estimate the additional Time required to prepare Decision Support information	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Time spent recreating EI not delivered from Projects	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Time spent getting lost EI from Vendors	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Idle Engineering resource time due to EIQ	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Additional Time spent on MoC Process	Cost caused by low data quality	Direct Costs	Re-entry costs
Please estimate the Reduced production due to poor design	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Additional Time accepting EI from Projects	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Additional Time spent approving mapping EI between systems	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Cost of redundant scrapped material	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Cost of redundant procurement	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Cost of redundant construction	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Likely Cost of professional error due to EIQ	Cost caused by low data quality	Indirect Costs	Cost based on lower reputation
Please estimate the Additional Time (re) creating Maintenance PMs	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Idle Maintenance resource time due to EIQ	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Additional Time to call of contracts	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Additional Time spent by TAR team (re) creating Work Packages	Cost caused by low data quality	Direct Costs	Re-entry costs
Please estimate the Additional Time spent optimizing maintenance resource	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Additional Time required to optimize asset operation	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Increase in Availability if EIQ was better	Cost caused by low data quality	Indirect Costs	Cost based on wrong decisions or actions
Please estimate the Cost reduction due to unnecessary inspection/maintenance	Cost caused by low data quality	Indirect Costs	Cost based on wrong decisions or actions
Please estimate the Cost of redundant spares in warehouse	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Cost of spares expediting ("hot shot costs")	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Production loss due to Asset sub-optimization	Cost caused by low data quality	Direct Costs	Compensation costs
Please estimate the Likely Cost of a Process Safety Incident due to poor EIQ	Cost caused by low data quality	Indirect Costs	Cost based on lower reputation
Please estimate the Additional Time spent assessing asset performance	Cost caused by low data quality	Indirect Costs	Sunk investment costs
Please estimate the Time spent identifying conflicts between systems	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Additional Time verifying all stakeholders have the same EI	Cost caused by low data quality	Direct Costs	Verification costs
Please estimate the Additional Time spent mapping between systems	Cost caused by low data quality	Direct Costs	Verification costs

One clarification is required with respect to the mapping to the Eppler & Helfert taxonomy. This is best explained by means of an example: It may be argued that the IE "Additional time spent mapping between systems" in the original taxonomy is best mapped to the taxonomy element "Standard Development and Deployment Costs"; however, in the final taxonomy it is mapped to "Verification Costs". The former example is part of the high-level grouping "Costs of improving or assuring data quality", whereas in the latter, actual mapping is part of the high-level grouping "Cost caused by low quality data". The rationale for this decision and others like it was as follows: this thesis develops a method to measure the cost of poor EIQ, which include only the grouping "Costs caused by low quality data". This implies that the various elements of the cost category "Cost of improving or assuring data quality" is not implemented in an asset and therefore cannot be measured.

5.2.2 The APQC List of Standard Processes

An attempt was made to map the primary taxonomy against the APQC Process List, but this has not been successful, as is evident by the gaps shown in Table 5.6. This is likely due to the APQC processes being at a very high level, whereas CoPEIQ is interested in detailed processes within an Asset.

Table 5.6 - Attempted Mapping Against APQC



PRIMARY TAXONOMY	ALTERNATIVE TAXONOMY per APQC			
Tactical Level				
Please estimate the Additional Time spent looking for EI				
Please estimate the Additional Time spent verifying or re-entering EI	10.0	Acquire, Construct, and Manage Assets	10.2.4.3	Create work and asset records
Please estimate the Additional Time repeating processes				
Please estimate the Time spent to validate/prepare corporate KPIs due to poor EIQ	13.0	Develop and Manage Business Capabilities	13.6.3.4	Calculate performance measures
Please estimate the Time spent clarifying misunderstanding				
Please estimate the Financial impact of misunderstanding				
Please estimate the Additional Time spent reviewing EI Standards				
Please estimate the Time spent creating and maintaining unofficial databases				
Please estimate the Time spent approving EI conflicts between databases				
Please estimate the Additional time spent optimizing the budget and production plan	13.0	Develop and Manage Business Capabilities	13.6.2	Benchmark performance
Please estimate the Production loss due to the wrong data being reported				
Please estimate the Additional time spent preparing Regulatory Reports	11.0	Manage Enterprise Risk, Compliance, Remediation, and Resiliency	11.2.2	Manage regulatory compliance
Please estimate the Additional time spent responding to Regulatory Queries	11.0	Manage Enterprise Risk, Compliance, Remediation, and Resiliency	11.2.2	Manage regulatory compliance
Please estimate the Likely Cost of a regulatory penalty due to Poor EIQ				
Please estimate the additional Time required to prepare Decision Support information				
Please estimate the Time spent recreating EI not delivered from Projects	10.0	Acquire, Construct, and Manage Assets	10.2.4	Manage asset construction
Please estimate the Time spent getting lost EI from Vendors	10.0	Acquire, Construct, and Manage Assets	10.2.4.3	Create work and asset records
Please estimate the Idle Engineering resource time due to EIQ				
Please estimate the Additional Time spent on MoC Process	13.0	Develop and Manage Business Capabilities	13.7.3.9	Establish and manage the management of change (MoC) process for HSSE
Please estimate the Reduced production due to poor design	10.0	Acquire, Construct, and Manage Assets	10.2.2	Design and plan asset construction
Please estimate the Additional Time accepting EI from Projects	10.0	Acquire, Construct, and Manage Assets	10.2.4.3	Create work and asset records
Please estimate the Additional Time spent approving mapping EI between systems				
Please estimate the Cost of redundant scrapped material	10.0	Acquire, Construct, and Manage Assets	10.2.1	Manage capital program for productive assets
Please estimate the Cost of redundant procurement	10.0	Acquire, Construct, and Manage Assets	10.2.1	Manage capital program for productive assets
Please estimate the Cost of redundant construction	10.0	Acquire, Construct, and Manage Assets	10.2.1	Manage capital program for productive assets
Please estimate the Likely Cost of professional error due to EIQ			11.1.4	Manage business unit and function risk
Please estimate the Additional Time (re) creating Maintenance PMs	10.0	Acquire, Construct, and Manage Assets	10.3.1	Plan asset maintenance
Please estimate the Idle Maintenance resource time due to EIQ	10.0	Acquire, Construct, and Manage Assets	10.3.4	Perform asset maintenance
Please estimate the Additional Time to call of contracts	10.0	Acquire, Construct, and Manage Assets	10.3.1	Plan asset maintenance
Please estimate the Additional Time spent by TAR team (re) creating Work Packages	10.0	Acquire, Construct, and Manage Assets	10.1.5	Plan major maintenance and plant turnarounds
Please estimate the Additional Time spent optimizing maintenance resource	10.0	Acquire, Construct, and Manage Assets	10.3.1	Plan asset maintenance
Please estimate the Additional Time required to optimize asset operation	13.0	Develop and Manage Business Capabilities	10.1.7	Optimize plant units
Please estimate the Increase in Availability if EIQ was better				
Please estimate the Cost reduction due to unnecessary inspection/maintenance				
Please estimate the Cost of redundant spares in warehouse				
Please estimate the Cost of spares expediting ("hot shot costs")				
Please estimate the Production loss due to Asset sub-optimization	10.0	Acquire, Construct, and Manage Assets	10.1.7	Optimize plant units
Please estimate the Likely Cost of a Process Safety Incident due to poor EIQ	11.0	Manage Enterprise Risk, Compliance, Remediation, and Resiliency	11.1.4	Manage business unit and function risk
Please estimate the Additional Time spent assessing asset performance	11.0	Manage Enterprise Risk, Compliance, Remediation, and Resiliency	11.4.2	Perform continuous business operations planning
Please estimate the Time spent identifying conflicts between systems				
Please estimate the Additional Time verifying all stakeholders have the same EI				
Please estimate the Additional Time spent mapping between systems				

As a result, APQC was not included in the Data Model.

5.2.3 Other Alternatives Taxonomies

In addition to the discussion in Section 5.1.2 with respect to Department serving as a proxy for Denominator, manipulation of the data model with respect to other Demographic variables could enable further granularity of results. A full list of these variables in shown in Table 3.11.

This section has explained the Taxonomy by which results are reported to OpCo management. A final step in the definition of how to present results is to specify the graphics. This is discussed the next section.

5.3 Develop Management Report

This final section of the chapter will summarize the various analyses done in previous sections in the form of a draft report of results to OpCo management. The graphical formats were taken from Table 4.1. Entries specific to the asset survey are in <Brackets> and graphical representations in *italics.* The management report is shown in Table 5.7.

Table 5.7 - Management Report

1. EXECUTIVE SUMMARY

This report summarizes the results of a survey done at <Asset> during the weeks <Date>to <Date>.

The intent was to determine the financial impact of poor Engineering Information Quality (EIQ) at <Asset>.
The methodology and limitations of the Survey are detailed below.
EIQ is defined as "Engineering Information in the form of data, documents, drawings and models, that is complete and accurate to the specification required by <Asset> and can be found by its defined user population in the correct repository".

The results of the simulation show that the Cost of poor EIQ (CoPEIQ) is shown in Figure A.

Figure A - COPEIQ for <Asset>, showing mean, standard deviation and 5% confidence limits.

The response rate was <Response Rate>.

Additional views of the results are shown below.
The areas of largest opportunity for <Asset> are:
<List of priorities per Pareto>

It is hoped that this study will enable informed decisions regarding remediation and preventative actions of EIQ at your asset.

2. INTRODUCTION

It is recognized that poor EIQ costs the energy industry a significant amount in lost production, additional cost, inefficiency and increased risk. A standardized method to calculate the financial impact, based on a survey methodology and a Monte Carlo simulation, has been developed to quantify this opportunity. With your permission granted on <Date>, the survey was deployed at your asset during the weeks <Date> to <Date>. The target respondent population was:

- Engineers, Technicians and Specialists supporting <Asset>.
- Maintenance and Inspection Planners
- Maintenance Leadership at all levels
- Corporate Planners
- Those involved in Process Safety

This report details the results of the study for your review and consideration.

3. METHODOLOGY

The survey questions were developed using an initial literature review and subsequent proof of concept at an operating asset.
Data collection is by voluntary (and therefore random) sampling of the target respondent audience by means of an email survey over two weeks.
Data analysis is by standardized Monte Carlo simulation and selected Pareto analyses to assist you in prioritizing interventions.

4. LIMITATIONS

The study is limited by the fact that it is measuring perception at <Asset> at the time of the Survey. The response is stochastic and therefore not absolute and only correct within the confidence limits stated in Figure A. The measurement excludes the perceived impact on the following groups:

- Capital Projects
- Information Technology
- 3rd parties
- Other functions in the organizations like Finance, HR and Supply Chain Management

The result is calculated stochastically and this therefore only accurate within the stated confidence limits.

5. RESULTS

5. 1. Overall Result

The summary of results of the CoPEIQ analysis is shown in Figure A.

Figure A .

It may be seen that the CoPEIQ result is between <Lower Limits> and <Upper Limit> with a confidence of 95%, a mean of <Mean> and a standard deviation of <StdDev>.

5. 2. Areas of Priority

The prioritized areas, based on a Pareto Analysis on a standardized taxonomy, were:

Figure B – Bar Graph of P50 results at Executive Level

The secondary prioritized areas, were:

Figure C – Second level bar graph for two top results in Figure B.

5. 3. Response Rate

The response rate for this Survey was <Response Rate>.

From these results it is suggested that:
CoPEIQ is costing <Asset> between and <Lower Limit> and <Upper Limit> USD per year, with the most likely figure being <Mean> USD.
The <First and second bars in Figure C> departments or teams will benefit most from a remediation effort.
The most likely improvement will be seen in <Top bar in Figure B>.
Further analyses can be made if required.

6. CONCLUSION

<Asset> has a potential opportunity of <Mean> USD per year if an EIQ remediation effort is initiated.

This section has concluded the specification of results presentation. With the survey, MCS and results presentation complete, the study is ready for implementation. This is discussed in Chapter 6.

Chapter 6

Applying the CoPEIQ Model to Field Data

The preceding chapter has developed the theme of presenting the results of the model developed in Chapter 4, which in turn was based on the development work in the preceding chapters. This body of work has prepared the CoPEIQ model to be applied in practice. That is the subject of this chapter.

Figure 6.1 provides an overview of the chapter. A short description is given of the preparation for the survey and for the survey itself, and a few comments are provided regarding the survey data received. The survey data is then loaded into the Excel model that had been developed in Chapter 4 and the results are displayed. The chapter concludes with a short discussion of results and their implications to the OpCo.

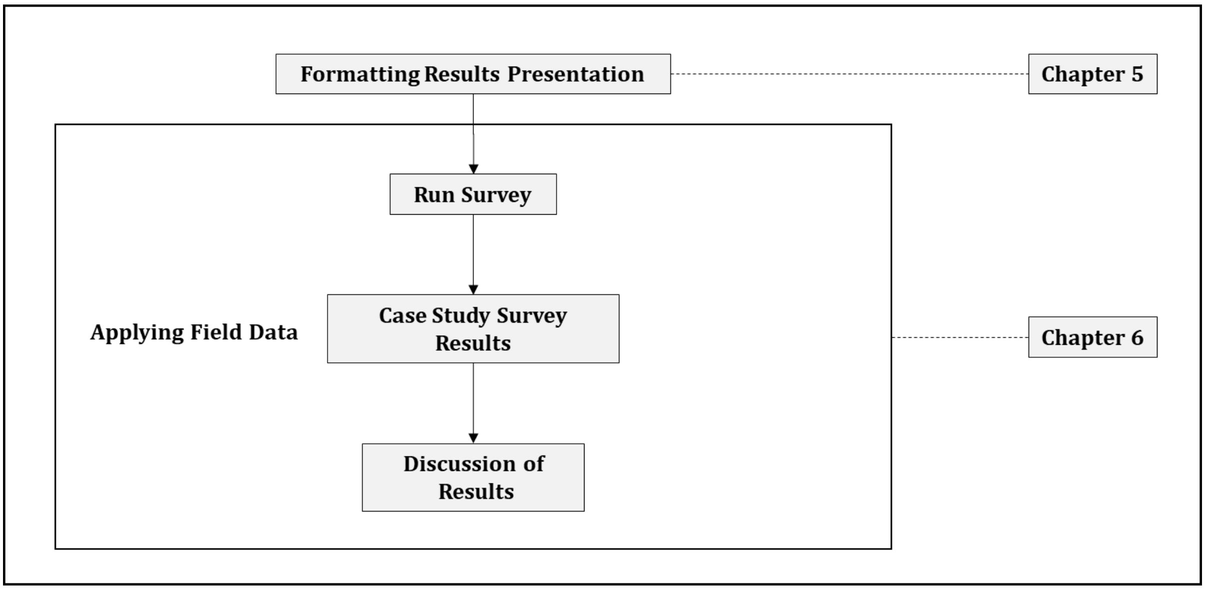


Figure 6.1 - Model and Validation Details

6.1 The Survey

The survey design has been explained in Chapter 3, and the preparatory mechanics to run a survey, in Section 3.4.3. This section will briefly describe the acquisition of the field survey data used for this thesis.

The field data was gathered during the development of a business case for an actual EIQ drive in an eminent oil multinational. The asset under review is a conventional, mature asset with a stable and mature workforce. This context is fortunate for this study since it negates the potential effects of unique technology or transient phenomena in the organisation.

Depending on the context of the OpCo or asset under review, considerable efforts may be necessary to raise awareness and convince the asset sponsoring manager of the efficacy of conducting a CoPEIQ survey. The specifics of this process during this case study are out of the scope of this study and confidential to the asset; suffice it to say that some work is required to enable the survey to be conducted.

Upon receiving permission, the steps laid out in Section 3.4.3 were followed largely unchanged. For reasons specific to the asset, SurveyMonkey was replaced with an internal tool and the acquired survey data, together with D_{Ref} data, was received anonymously from the OpCo for this study.

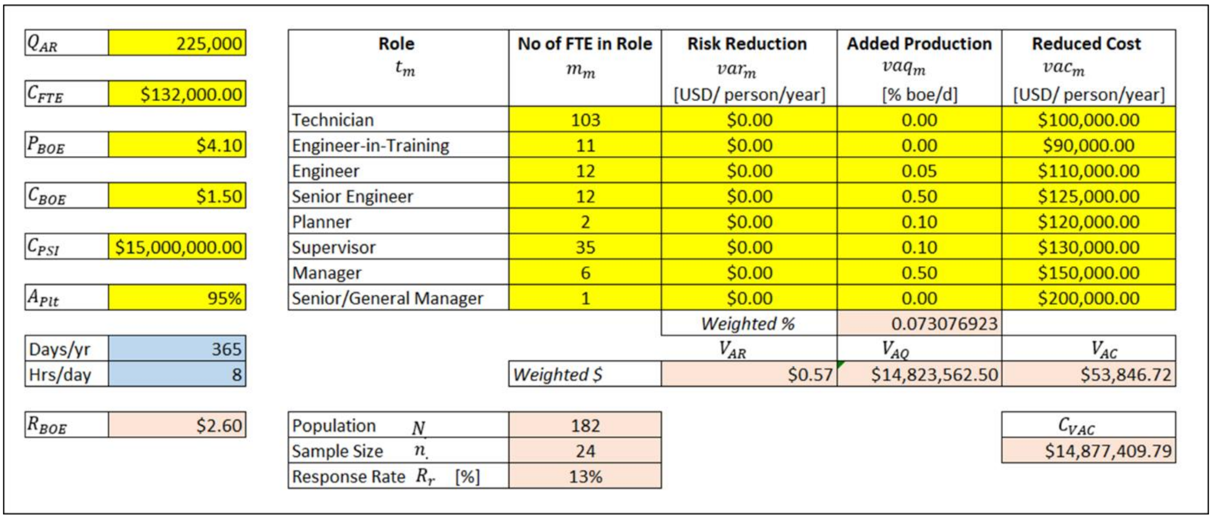
Ironically for this thesis, a brief comment on data quality is also required. The survey data received for this test required not insignificant efforts to ensure consistency in data format and UoM. This resulted in several errors when the data was first loaded into the CoPEIQ model. Excel has a helpful "Trace Error" function which enabled isolation of the sources and confirmed that there were no programming errors in the CoPEIQ model. This is hardly surprising: Klein (2000) concluded that between 0 and 10% serious data errors and between 10 and 40 trivial data errors for municipal bond data could be expected.

Upon ensuring that the data UoM and format are in a suitable format for the model, the data was entered into the CoPEIQ model and the results achieved as expected. This is discussed in Section 6.2.

6.2 Survey Data Analysis

The results of the CoPEIQ calculations of field survey data are discussed in this section. Since the model design has been discussed in detail in Chapter 4, the format is essentially in a list of screen shots of the actual data.

The display starts with a view of the input data D_{Ref} provided by the sponsoring manager. It provides an overview of the order of magnitudes of production, cost and organisational numbers.



Q_{AR}	225,000
C_{FTE}	$132,000.00
P_{BOE}	$4.10
C_{BOE}	$1.50
C_{PSI}	$15,000,000.00
A_{Plt}	95%
Days/yr	365
Hrs/day	8
R_{BOE}	$2.60

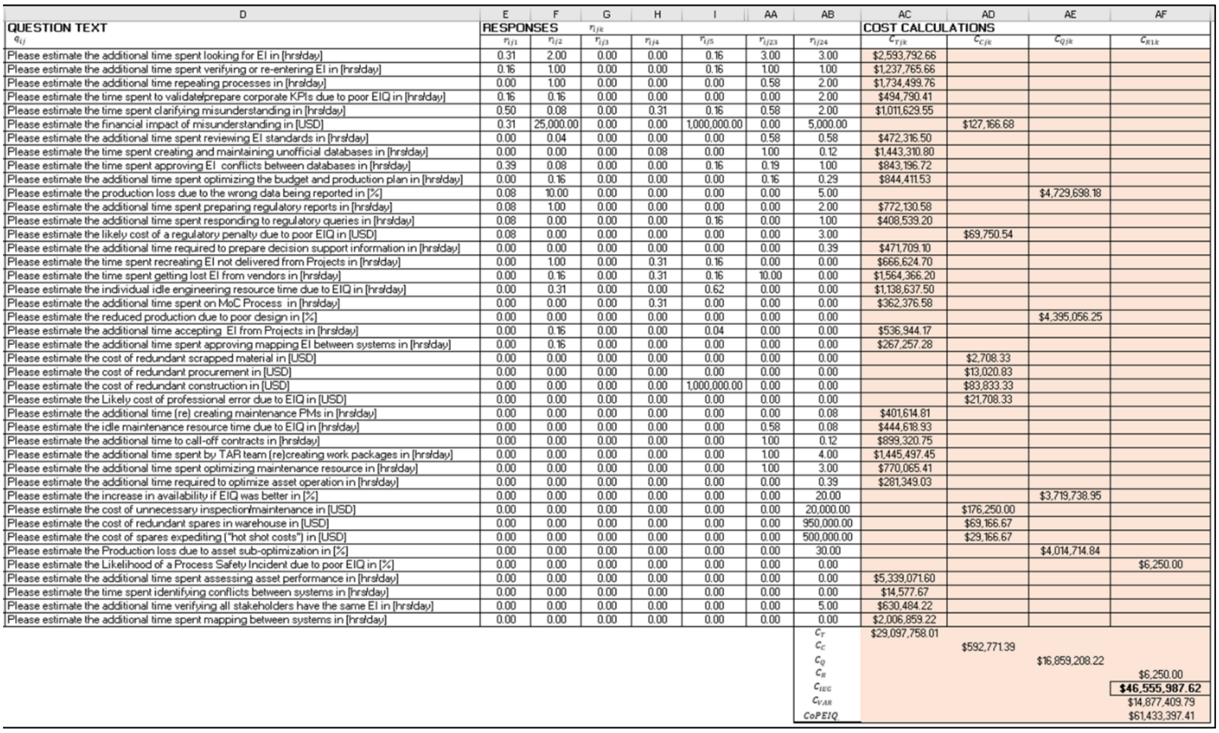
Role t_m	No of FTE in Role m_m	Risk Reduction var_m [USD/ person/year]	Added Production vaq_m [% boe/d]	Reduced Cost vac_m [USD/ person/year]
Technician	103	$0.00	0.00	$100,000.00
Engineer-in-Training	11	$0.00	0.00	$90,000.00
Engineer	12	$0.00	0.05	$110,000.00
Senior Engineer	12	$0.00	0.50	$125,000.00
Planner	2	$0.00	0.10	$120,000.00
Supervisor	35	$0.00	0.10	$130,000.00
Manager	6	$0.00	0.50	$150,000.00
Senior/General Manager	1	$0.00	0.00	$200,000.00
		Weighted %	0.073076923	
		V_{AR}	V_{AQ}	V_{AC}
	Weighted $	$0.57	$14,823,562.50	$53,846.72

Population	N	182
Sample Size	n	24
Response Rate	R_r [%]	13%

C_{VAC}
$14,877,409.79

Figure 6.2 – Reference Data D_{Ref} for the Case Study

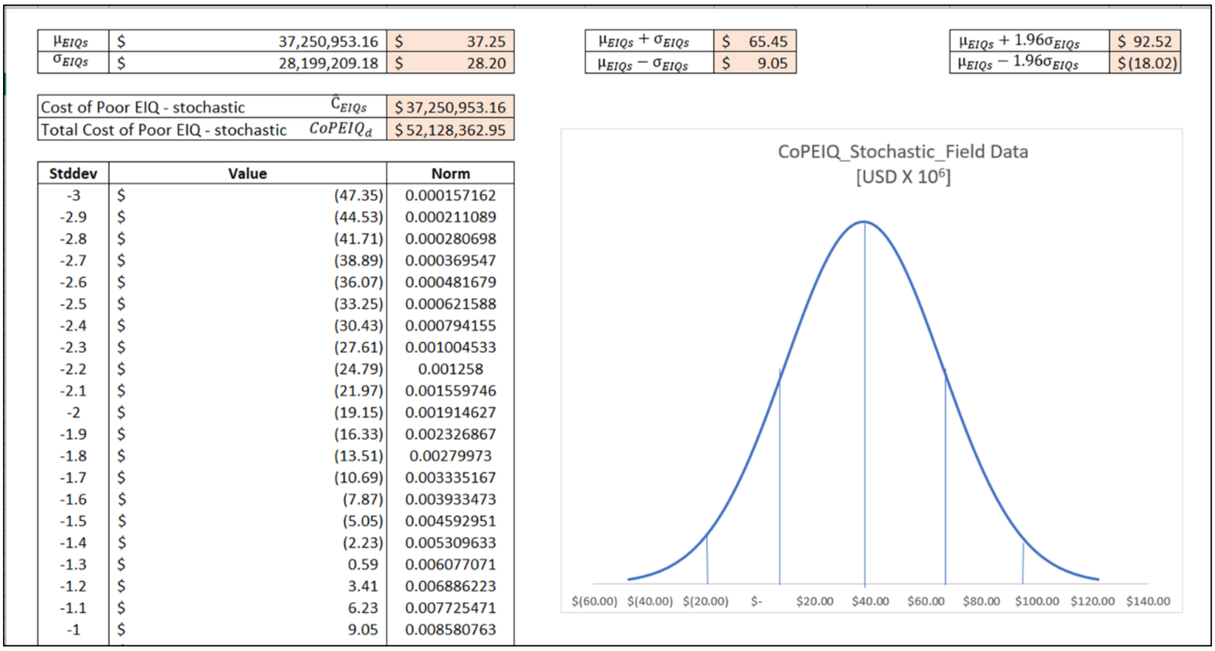
D_{Ref} data is followed by a view of the $CoPEIQ_d$ calculation in Figure 6.3. Note that rows 11-39 and columns G to AE have been hidden for graphical expediency.



D	E	F	G	H	I	AA	AB	AC	AD	AE	AF
QUESTION TEXT	RESPONSES	r_{ijk}						COST CALCULATIONS			
q_{ij}	r_{ij1}	r_{ij2}	r_{ij3}	r_{ij4}	r_{ij5}	r_{ij23}	r_{ij24}	C_{Tjk}	C_{Cjk}	C_{Qjk}	C_{R1k}
Please estimate the additional time spent looking for EI in [hrs/day]	0.31	2.00	0.00	0.00	0.16	3.00	3.00	$2,593,792.66			
Please estimate the additional time spent verifying or re-entering EI in [hrs/day]	0.16	1.00	0.00	0.00	0.16	1.00	1.00	$1,237,765.66			
Please estimate the additional time repeating processes in [hrs/day]	0.00	1.00	0.00	0.00	0.00	0.58	2.00	$1,734,499.76			
Please estimate the time spent to validate/prepare corporate KPIs due to poor EIQ in [hrs/day]	0.16	0.16	0.00	0.00	0.00	0.00	2.00	$494,790.41			
Please estimate the time spent clarifying misunderstanding in [hrs/day]	0.50	0.08	0.00	0.31	0.16	0.58	2.00	$1,011,629.55			
Please estimate the financial impact of misunderstanding in [USD]	0.31	25,000.00	0.00	0.00	1,000,000.00	0.00	5,000.00		$127,166.68		
Please estimate the additional time spent reviewing EI standards in [hrs/day]	0.00	0.04	0.00	0.00	0.00	0.58	0.58	$472,316.50			
Please estimate the time spent creating and maintaining unofficial databases in [hrs/day]	0.00	0.00	0.00	0.08	0.00	1.00	0.12	$1,443,310.80			
Please estimate the time spent approving EI conflicts between databases in [hrs/day]	0.39	0.08	0.00	0.00	0.16	0.19	1.00	$843,196.72			
Please estimate the additional time spent optimizing the budget and production plan in [hrs/day]	0.00	0.16	0.00	0.00	0.00	0.16	0.29	$844,411.53			
Please estimate the production loss due to the wrong data being reported in [%]	0.08	10.00	0.00	0.00	0.00	0.00	5.00			$4,729,698.18	
Please estimate the additional time spent preparing regulatory reports in [hrs/day]	0.08	1.00	0.00	0.00	0.00	0.00	2.00	$772,130.58			
Please estimate the additional time spent responding to regulatory queries in [hrs/day]	0.08	0.00	0.00	0.00	0.16	0.00	1.00	$408,539.20			
Please estimate the likely cost of a regulatory penalty due to poor EIQ in [USD]	0.08	0.00	0.00	0.00	0.00	0.00	3.00		$69,750.54		
Please estimate the additional time required to prepare decision support information in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.00	0.39	$471,709.10			
Please estimate the time spent recreating EI not delivered from Projects in [hrs/day]	0.00	1.00	0.00	0.31	0.16	0.00	0.00	$666,624.70			
Please estimate the time spent getting lost EI from vendors in [hrs/day]	0.00	0.16	0.00	0.31	0.16	10.00	0.00	$1,564,366.20			
Please estimate the individual idle engineering resource time due to EIQ in [hrs/day]	0.00	0.31	0.00	0.00	0.62	0.00	0.00	$1,138,637.50			
Please estimate the additional time spent on MoC Process in [hrs/day]	0.00	0.00	0.00	0.31	0.00	0.00	0.00	$362,376.58			
Please estimate the reduced production due to poor design in [%]	0.00	0.00	0.00	0.00	0.00	0.00	0.00			$4,395,056.25	
Please estimate the additional time accepting EI from Projects in [hrs/day]	0.00	0.16	0.00	0.00	0.04	0.00	0.00	$536,944.17			
Please estimate the additional time spent approving mapping EI between systems in [hrs/day]	0.00	0.16	0.00	0.00	0.00	0.00	0.00	$267,257.28			
Please estimate the cost of redundant scrapped material in [USD]	0.00	0.00	0.00	0.00	0.00	0.00	0.00		$2,708.33		
Please estimate the cost of redundant procurement in [USD]	0.00	0.00	0.00	0.00	0.00	0.00	0.00		$13,020.83		
Please estimate the cost of redundant construction in [USD]	0.00	0.00	0.00	0.00	1,000,000.00	0.00	0.00		$83,833.33		
Please estimate the Likely cost of professional error due to EIQ in [USD]	0.00	0.00	0.00	0.00	0.00	0.00	0.00		$21,708.33		
Please estimate the additional time (re) creating maintenance PMs in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.00	0.08	$401,614.81			
Please estimate the idle maintenance resource time due to EIQ in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.58	0.08	$444,618.93			
Please estimate the additional time to call-off contracts in [hrs/day]	0.00	0.00	0.00	0.00	0.00	1.00	0.12	$899,320.75			
Please estimate the additional time spent by TAR team (re)creating work packages in [hrs/day]	0.00	0.00	0.00	0.00	0.00	1.00	4.00	$1,445,497.45			
Please estimate the additional time spent optimizing maintenance resource in [hrs/day]	0.00	0.00	0.00	0.00	0.00	1.00	3.00	$770,065.41			
Please estimate the additional time required to optimize asset operation in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.00	0.39	$281,349.03			
Please estimate the increase in availability if EIQ was better in [%]	0.00	0.00	0.00	0.00	0.00	0.00	20.00			$3,719,738.95	
Please estimate the cost of unnecessary inspection/maintenance in [USD]	0.00	0.00	0.00	0.00	0.00	0.00	20,000.00		$176,250.00		
Please estimate the cost of redundant spares in warehouse in [USD]	0.00	0.00	0.00	0.00	0.00	0.00	950,000.00		$69,166.67		
Please estimate the cost of spares expediting ("hot shot costs") in [USD]	0.00	0.00	0.00	0.00	0.00	0.00	500,000.00		$29,166.67		
Please estimate the Production loss due to asset sub-optimization in [%]	0.00	0.00	0.00	0.00	0.00	0.00	30.00			$4,014,714.84	
Please estimate the Likelihood of a Process Safety Incident due to poor EIQ in [%]	0.00	0.00	0.00	0.00	0.00	0.00	0.00				$6,250.00
Please estimate the additional time spent assessing asset performance in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	$5,339,071.60			
Please estimate the time spent identifying conflicts between systems in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	$14,577.67			
Please estimate the additional time verifying all stakeholders have the same EI in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.00	5.00	$630,484.22			
Please estimate the additional time spent mapping between systems in [hrs/day]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	$2,006,859.22			
							C_T	$29,097,758.01			
							C_C		$592,771.39		
							C_Q			$16,859,208.22	
							C_R				$6,250.00
							C_{IEC}				**$46,555,987.62**
							C_{VAR}				$14,877,409.79
							$CoPEIQ$				$61,433,397.41

Figure 6.3 -Deterministic CoPEIQ for Case Study Data

Equivalent to Figure 4.11, the parameters of the stochastic analysis for field data are shown in Figure 6.4.



μ_{EIQs}	$	37,250,953.16	$ 37.25
σ_{EIQs}	$	28,199,209.18	$ 28.20

$\mu_{EIQs} + \sigma_{EIQs}$	$ 65.45
$\mu_{EIQs} - \sigma_{EIQs}$	$ 9.05

$\mu_{EIQs} + 1.96\sigma_{EIQs}$	$ 92.52
$\mu_{EIQs} - 1.96\sigma_{EIQs}$	$(18.02)

Cost of Poor EIQ - stochastic	$\hat{C}_{EIQs}$	$ 37,250,953.16
Total Cost of Poor EIQ - stochastic	$CoPEIQ_d$	$ 52,128,362.95

Stddev	Value		Norm
-3	$	(47.35)	0.000157162
-2.9	$	(44.53)	0.000211089
-2.8	$	(41.71)	0.000280698
-2.7	$	(38.89)	0.000369547
-2.6	$	(36.07)	0.000481679
-2.5	$	(33.25)	0.000621588
-2.4	$	(30.43)	0.000794155
-2.3	$	(27.61)	0.001004533
-2.2	$	(24.79)	0.001258
-2.1	$	(21.97)	0.001559746
-2	$	(19.15)	0.001914627
-1.9	$	(16.33)	0.002326867
-1.8	$	(13.51)	0.00279973
-1.7	$	(10.69)	0.003335167
-1.6	$	(7.87)	0.003933473
-1.5	$	(5.05)	0.004592951
-1.4	$	(2.23)	0.005309633
-1.3	$	0.59	0.006077071
-1.2	$	3.41	0.006886223
-1.1	$	6.23	0.007725471
-1	$	9.05	0.008580763

Figure 6.4 – Stochastic CoPEIQ Parameters for Case Study Data

This review concludes with a view of the taxonomy analysis in Figures 6.5 and 6.6. These figures provide an insight into the calculated CoPEIQ. As has been explained in Section 5.1.3, they are simply Pareto plots of the taxonomy, at the Executive and Management level respectively.

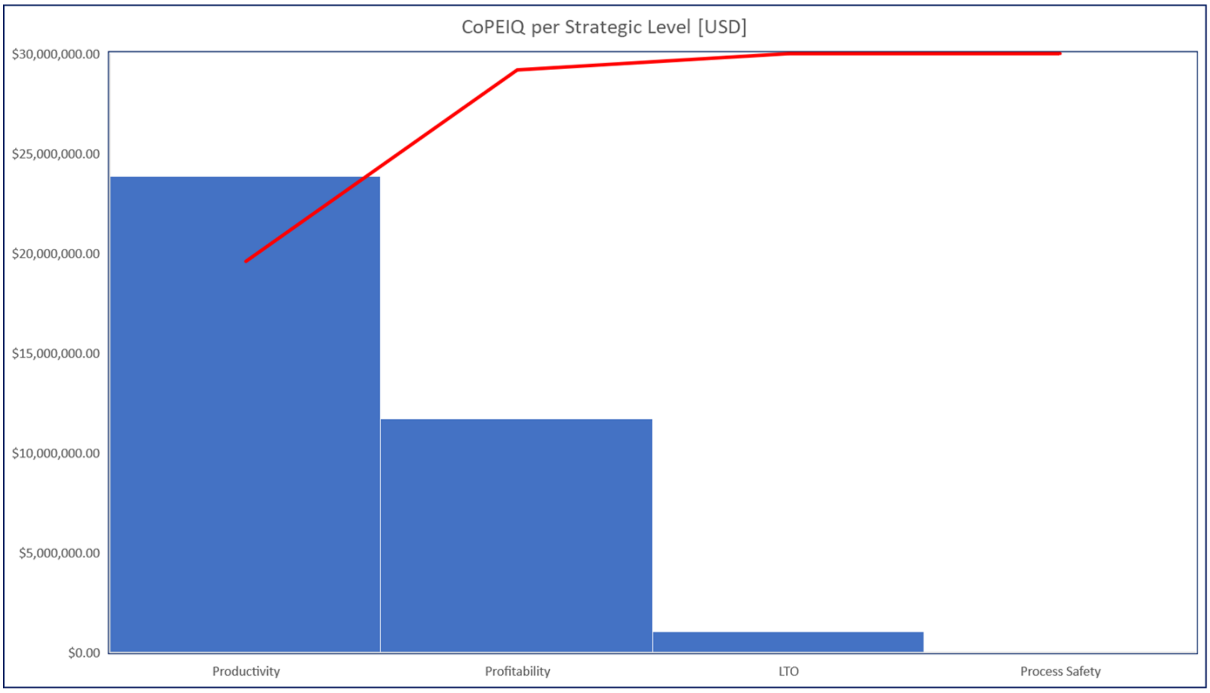


Figure 6.5 - Executive Level Pareto Chart for Case Study Data

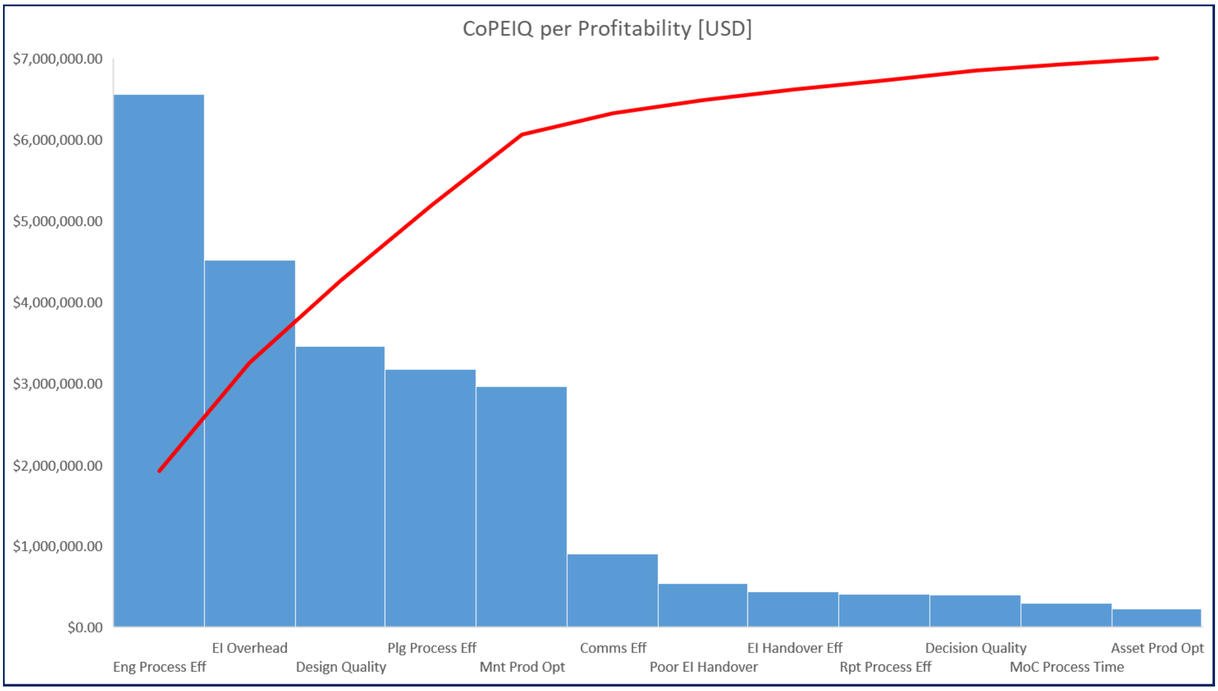


Figure 6.6 - Management Level Pareto Chart of Productivity Result for Case Study Data

Similar to the discussion in Section 5.1.3.1, these results provide useful clues about which data sets are in the worst state, or the parts of the organisation where the biggest challenges exist for the asset under review. The results are discussed in the next section.

6.3 Discussion of Results

This section briefly reviews the results presented in section 6.2.

Already in Section 1.3.6, one limitation of this study was that it is “by nature an approximation”. Indeed, in Table 3.2 the survey respondent was reminded that “the objective is not to be exactly accurate, rather to be a reasonable and realistic estimate”. Section 4.1, where data presentation requirements to management were discussed, concluded that “that the actual magnitude of uncertainty is less important than its attributes”.

These three examples demonstrate the general gist of the CoPEIQ calculation: it is intended to provide the sponsoring manager with an estimate of CoPEIQ that is accurate enough to enable a relative ranking when compared to other candidates for a necessarily finite improvement budget. The intended audience expects uncertainty in the response; a quantification of that uncertainty is deemed adequate.

On the basis of this argument, the response rate of 13% for this field data is considered valid to meet the intent of the CoPEIQ calculation. It does, however, call for introspection about how to improve response rates in practice. This is discussed in Section 7.2. The low response rate may indicate several organisational or cultural characteristics, for which the temptation to enter into speculation is resisted for the purposes of this thesis.

The unit orders of magnitude for the case study are smaller than the structured interview-based results obtained at the start of this study. This difference may be explained by differences in asset size, geographic factors, more complex technology or subjective bias on the part of the interviewer during the early structured interviews.

The standard deviation of the MCS results is around 75% of the mean. This is indicative of a large spread of opinion in the asset on the research subject. This may conceivably be explained by differences in the relative maturity of data or subcultures in the teams; the researcher does not have access to enough information to explore this further.

The large spread in the data also explains the almost 30% difference between the deterministic and stochastic results of CoPEIQ.

The Pareto result of Productivity, the most significant Executive factor, is followed closely by Profitability (Figure 6.5). For Productivity, the biggest opportunity is in Engineering Process Efficiency, followed by a group of five Management Level results (Figure 6.6). It may therefore be said that the Pareto results for the case study are in general more balanced than those presented for the laboratory data presented in Chapter 5. These results may indicate to the sponsoring manager that a more holistic intervention into EIQ is indicated, perhaps after a more immediate drive to improve Engineering Process Efficiency.

In any event, the results are a first indication of what may be expected for a stochastic study of this nature and, as such, are significant for both this study and the OGI in general.

In terms of value to the OpCo, the results may be interpreted as follows:

The characterization of USD 37M (or USD75M if Alternative Contribution is included), but with a standard deviation of nearly three-quarters of that figure, provides management with the information it needs to judge the efficacy of investing in an EIQ remediation initiative. It is significant for this specific case study that the results show that there is a possibility of 1.3 standard deviations that the effort will result in a negative outcome. This is offset by the possibility of saving a mean of $52 million, a not insignificant figure on any scale. An ambiguous result of this nature is typical for the industry and indicates a cautious, phased approach to the CoPEIQ problem. This is also typical for the industry.

This chapter has described the application of the CoPEIQ model in a real-life setting and displayed the results of the analysis. A discussion of these results, together with a much wider introspection, follows in Chapter 7.

Chapter 7

Conclusion

7.1 Conclusions

During the initial phases of this study, the following research questions have been posed in Section 1.3:

- Has a method been developed previously to measure the financial impact of poor EIQ?
- What is a sensible classification of Impact Elements that jointly constitute the financial impact of poor EIQ?
- What standardized list of aggregated dimensions is appropriate, against which to report the results (outputs) of IEs? (For instance, there is no standardized list of risks associated with poor EIQ in the literature).
- What are the appropriate units of measure for the outputs?
- How should data of this nature be presented to management in the OGI?
- What is the appropriate model to use for analysing this data?
- What are the appropriate statistical instruments to analyse this data?

These questions culminating in the following research objective:

Develop a standardized model to quantify the financial impact of Engineering Information Quality in the OGI, for a specific context.

This research objective was deemed to be of value by virtue of its ability to enable the OGI by to assess the value of an EI remediation effort on a comparable basis to other improvement or expansion opportunities in an OpCo.

This section will in the first instance draw conclusions regarding the extent to which these research questions have been answered and the extent to which the research objective has been met. A few general conclusions will be drawn subsequently.

7.1.1 Have the Research Questions Been Answered?

The seven research questions noted above are answered in turn in this section.

1. Has a method been developed previously to measure the financial impact of poor EIQ?

   Despite previous efforts and precedent in related industries, Section 2.2 concludes decisively that the answer to this question, prior to this current study, is “No”.

2. What is a sensible classification of Impact Elements that jointly constitute the financial impact of poor EIQ?

The development of the Taxonomy, described in Chapter 5, answers this question. Chapter 5 also tests the notion of an alternative taxonomy and concludes that such a possibility may be feasible.

3. What standardized list of aggregated dimensions is appropriate, against which to report the results (outputs) of IEs?

Chapter 5 also answers this question decisively.

4. What are the appropriate units of measure for the outputs?

   For the purposes of the intended audience, it was concluded in Section 3.3.2 that [USD/year] is the most appropriate. Four interim units of measure had been derived:

   - Hours/day
   - USD/year
   - % Cost difference [USD/year]
   - % Risk reduction [USD/year]

5. How should data of this nature be presented to management in the OGI?

   Section 4.2 summarizes the conclusions drawn in this regard.

6. What is the appropriate model to use for analysing this data?

   The CoPEIQ model has been described in Chapter 4.

7. What are the appropriate statistical instruments to analyse this data?

At the end of Section 4.5.1.2 it is concluded that MCS is the most appropriate for this study.

It may therefore be said in summary that all the research questions have been answered.

7.1.2 Has the Research Objective been met?

Based on the detailed responses in Section 7.1.1, it can be said that the research objective "To develop a standardized Model to quantify the financial impact of Engineering Information Quality in the OGI, for a specific context" has been met. The same can be said for its rationale, to "develop a standardized Model to quantify the financial impact of Engineering Information Quality in the OGI, for a specific context", has been met, judging by the adoption of the principles by several entities in the OGI.

7.1.3 Other Conclusions

This section reflects briefly on additional conclusions drawn during the literature survey and the many conversations that ensued during the study.

It is concluded in the first instance that the problem of poor EIQ extends well beyond the scope of this study, and that it is instinctively understood by most stakeholders involved.

The appropriate process to address the problem is not well understood by stakeholders and is beyond the scope of this document.

Regardless of the appropriate response to a reality of poor EIQ, a defensible business case is required in every instance except in the rare case where a senior manager is convinced or frustrated enough to fund the remediation without challenge.

The conclusions drawn above are tempered somewhat by the introspection discussed in the next section.

7.2 Introspection

The decisions, approaches and calculations in the preceding chapters are reviewed in this section from a critical viewpoint and with the benefit of hindsight. Several points of introspection are discussed:

- The balance of granularity: the conflict between the need for granularity and the need for a concise survey is a challenge for work of this nature. A concise survey is deemed more palatable to a survey audience and more likely to remain within the researcher's cognitive bandwidth. Empirically, a questionnaire demanding more than 20 minutes' time to complete is likely to be ignored or done poorly. Conversely, the need for rich data and the potential value of secondary research into the data set demands deep and wide detail. The cost and effort of getting permission to conduct a survey and the effort of acquiring and analysing the data, creates a very tempting environment to gather more data for related reasons. In Section 3.3 it is described how this study yielded no more than 20 questions for each class of respondent. This is deemed a reasonable optimum.

- Following from the previous point, the survey questionnaire is considered to be sub-optimal at the time of writing. Further refinement to achieve a higher degree of mutual exclusivity is not only possible, but desirable, not least due to the benefits of a shorter survey. Given that this study is a first exploration of the subject, this is a predictable conclusion.

- An assessment of the Research Objective and Research Questions yields the following list of requirements of the eventually selected Research Model. Each is reviewed in turn:

    - Disaggregation: it should be able to quantify the impact of poor EIQ through a number of scopes, 'ranges', 'maturities' or 'use cases' as defined in EPRI (2014).

      This requirement is necessitated by the fact that every OpCo will have a variety of contexts and scoping decisions which this study is intended to support. This is supported by ISO 14224, which contains a recommendation that a cost-benefit analysis be done before an extensive asset data remediation programme is initiated.

      The disaggregation requirement has been addressed during the development of the taxonomy. The current research is limited by the scope limitation shown in Figure 3.4. The foundation exists, however, for expansion of that scope.

- Simplicity: it should be as simple as the functional requirement permits (EPRI 2014) and CEN fig 5 (CEN 2009).

  The need to simplify the survey questionnaire was confirmed during the validation structured interviews (Section 3.3.1). The initial questionnaire was simplified and shortened, and an explanatory introduction added for survey respondents. This, however, remains an inherent limitation for an uninitiated respondent. One response to this challenge might be to use neutral interviewers and a carefully controlled, structured interview; however, this reduces the ease of execution and may impede permission to gather CoPEIQ data under the auspices of a hesitant sponsoring manager.

- Palatability: Its design should be of palatable complexity for mid-level OpCo managers, who would need to be convinced of the value of the model and surrounding methodology.

  The palatability requirement has been optimised by means of the choice of language, simplification of the questionnaire and extensive efforts to present the data in a palatable form.

- Introspection from an academic point of view is related to survey reliability. The considerable attention in the literature regarding survey reliability is demonstrated by this study. Several technical and methodological caveats exist for this study, which are required to be specified when reporting CoPEIQ results. Criticism of these assumptions may include:

  - Independence. It may be argued that the behaviour of the target population for a survey is determined by the team or site culture in force at the time, which implies that the individual responses are biased towards the culture and that they are therefore not independent. The only response to this challenge is to assume independence until an adequate sample of results exist in the OGI to test for correlation of this assumption.

  - Statistical validity: The work relies on the validity of the Central Limit Theorem for the sample data. Results should therefore be reported with care in the case of a low response rate, such as occurs in the case with the case study in Chapter 6. It may be argued that re-sampling will improve this situation, but it may conversely be argued that re-sampling in short order will invoke the Hawthorne Effect, which is likely to distort the responses even more.

  - Statistical validity is also a function of sample size. The low response rate of 13% for the field data analysis had been deemed adequate for the intent of this study in Section 6.3, but nevertheless calls for some thought on how to improve the response rate for future studies. Possible options might be to raise awareness in the target audience before sending the survey, or for the sponsoring manager to state that responses to the survey is mandatory. This latter option does however introduce a higher possibility of bias in the responses. A final option is to use structured interviews to capture data, albeit this is a considerably more expensive option.

- Introspection at the macro level yielded the following insights:
    - The fundamental problem of poor EIQ, together with the multitude of factors leading to it of this problem, persists apace in the OGI. The method developed here is one contributor to the eventual resolution of the problem.
    - The inherent challenge of response rate was demonstrated by the case study result.
    - A large population of survey data, coupled with demographic data, constitutes exponentially larger value to the OGI than the CoPEIQ calculations imply, particularly with the rapid advancement of AI technologies and their adoption in the OGI.
    - Conversely, gaining permission to run one or more surveys of this nature requires more effort than the potential benefits imply, reflecting perhaps on the irrationality of OGI decision-making discussed in Section 4.1.

These challenges, or similar ones, are inherent in most data presented to senior managers when requesting funding for an improvement initiative. Notwithstanding this introspection, therefore, the final point of introspection is that the CoPEIQ model meets its original intent.

7.3 Future Research

In this penultimate section, a few perspectives are given about possible subsequent research. These considerations are provided from three viewpoints: academic, practical and macro-behavioural.

7.3.1 Academic Viewpoint

Many opportunities exist to further research this subject from an academic point of view. The following opportunities come to mind:

- The variables used in the C_{VAC} calculation are, as stated, in fact potential sources of MCS themselves. Just one example will suffice: Plant availability and the factors driving it are well-researched and understood.

- The benefits to an OpCo and third parties that have been excluded from the scope of this study, will add to the total benefit of improved EIQ by a large but unknown factor. The core methodology developed in this study may be expanded to include more benefits with relatively little effort.

- The preceding chapters emphasize the early, exploratory nature of this research. With an adequate number of studies, the benefits of Bayesian statistics may very well be applied to refine CoPEIQ, as has been explored in Section 4.5.3.

7.3.2 Practical Viewpoint

From a practical perspective, the initial realization reached during this study is that the quantification of CoPEIQ is in fact a very small part of the EIQ problem. Much remains to be understood regarding the internal mechanisms and reasons that cause poor EIQ in the first instance and/or what perpetuates it inside an OpCo. Examples of further research topics may be:

- What are standard time benchmarks for each time-related IE?
- Which of the demographic variables are predominant?
- What is the specific impact of poor EIQ on other organisational entities, notably Supply Chain Management?
- What is the impact of applying sophisticated software systems to the core processes related to EIQ?
- What impact does the accelerated application of AI have on CoPEIQ and EIQ in general?

Contantinou et al. (2016) interpreted a complex and unstructured data set for a complex problem into a structured Bayesian network for medical decision support. This method may arguably be used to derive causal relationships for the problem of poor EIQ and provide OpCo management with a predictive model to make the correct decisions early in an asset life.

Finally, mention was made in Section 4.4.2 of the so-called "Data Island Hypothesis". This concept was first mentioned during the initial structured interviews, where one interviewee mentioned the possibility that an engineer who has been working in a certain OpCo for an extended period will gather for himself/herself a copy of the EI needed to perform their specific duties. With the passage of time, the specified elements of EIQ required are reviewed until the subject is comfortable with the EIQ, after which the efficiency of the subject improves, since the lost time to confirm EIQ and rework certain elements is reduced. A hypothetical "Data Island Effect" is shown in Figure 7.1.

With reference to Table 2.1, one technique for Option A may be to derive survey questions from a Value Driver Tree (VDT) of a number of standardised engineering processes. The VDT so derived could conceivably mapped to the relative contribution of EIQ to the effectiveness of each process. This approach will yield a series of survey questions which respondents may find easier to answer. One example may be "what is the value of predictive maintenance lost due to poor EIQ?" This approach presumes that the sponsoring manager is convinced of the value of this approach to an extent that will warrant the investment of agreeing standard processes, mapping a VDT and populating the node data required. Conversely, considerable additional diagnostic value is achievable from this approach, such as for example a sensitivity analysis of the relative value of processes and the consequent behavioural interventions.

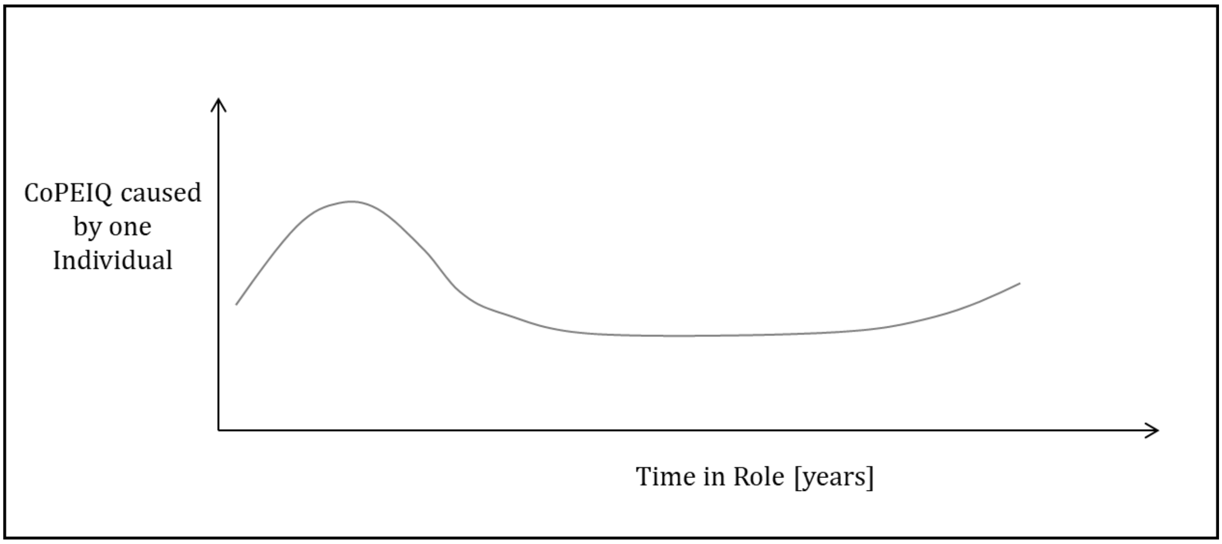


Figure 7.1 - Hypothetical "Data Island Effect"

It would be useful to test the following in future research:

- Does such a "data island" phenomenon exist?
- What is its effect on the asset, either positive or negative?
- How long should a subject be in a role before this phenomenon manifests?
- Does the effect change over time, or does it approach some asymptotic value?

The final section discusses research from a macro-behavioural point of view.

7.3.3 Macro-Behavioural Viewpoint

Quantification of the cost of poor EIQ does nothing more than to provide the basis for the approval of funding to start a remediation effort. Much opportunity for research remains about subjects like how to prevent poor EIQ in the first instance, what is an appropriate remediation strategy or how exactly should EIQ be defined. One approach might be to view poor EIQ predominantly as a behavioural problem and to explore the attributes of the problem further with a view of preventing them. Examples research questions might be:

- Does the EIQ problem arise due to a lack of trust of the data, a lack of time to maintain it, both, or something else?
- Has the advent of technology changed the prevalence of the problem?
- Does the prevalence of poor EIQ correlate with the oil price, possibly due to some intermediate variable such as personnel turnover rate?
- Since one catastrophic "black swan" PSI could possible pay for an EIQ remediation initiative for an entire OpCo, and poor EIQ is a known contributor to PSI's, why does the irrational status persist for OpCo's to allow poor EIQ in their operations?

A final perspective might be to view the behaviour leading to poor EIQ as a symptom of a larger behavioural problem, which also manifests in other problems, such as perhaps a decrease in maintenance or design quality. If so, what might the contributing factors be? This thought may be expanded as follows: Reason (1990 and1998) contends that safety performance and commercial performance are related because the psychological roots of these indices are the same. Could the same be said for poor EIQ?

Section 7.3 proposes several opportunities for further research. These thoughts conclude this study into EIQ. The final section reports briefly on the changes in the context that originally prompted this study and the potential impact of those changes on this study.

7.4 Endnote

The primary motivation for this study has been to develop an innovation that will quickly and in real terms improve asset performance in the OGI. During the conducting of this study, two events in the macro-environment have occurred that emphasize the original contentions stated in Section 1.3:

- Continued dynamics in the socio-political and macro-economic environments have increased the pressure on OpCo's to improve their internal efficiency.
- Accelerated emergence and innovation of so-called "data science" initiatives have increased the visibility of EIQ.

The likelihood of satisfying the original motivation is therefore rapidly increasing.

References

Aletras, N., Tsarapatsais, D., Preotuic-Pietro, D. and Lampos, V. (2016). 'Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective.' *Peer J Computer Science, e93,* October 24, 2016. https://doi.org/10.7717/peerj-cs.93

American Petroleum Institute (API), (2016). *Risk-Based Inspection.* API Recommended Practice 580. 3rd Edition. [Online]. Available: http://www.techstreet.com/api/searches/12764433

American Productivity and Quality Center (APQC), (2016). *APQC Process Classification Framework (PCF) – Downstream Petroleum – Excel Version 7.0.5.* Published September 28, 2016. https://www.apqc.org/knowledge-base/documents/apqc-process-classification-framework-pcf-downstream-petroleum-excel-vers-0 (Accessed: 20th February 2017).

Arlbjørn, J.S., Wong, C.Y. & Seerup, S. (2007). 'Achieving competitiveness through supply chain integration.' *International Journal of Integrated Supply Management*, 3(1), pp. 4-24.

Axxin, W.G., Link, C.F. and Groves, R.M. (2011). 'Responsive survey design, demographic data collection and models of demographic behaviour.' *Demography,* 48, pp. 1127-1149.

Bailey, K.D. (1984). A three-level measurement model. *Quality and Quantity,* 18(3), pp. 225-245.

Ballou, D.P. and Pazer, H. (1985). 'Modelling data and process quality in multi-input multi-output information systems.' *Management Science*, 32(2), pp. 150-162.

Bryman, A., Bell, E., Hirschsohn, P., Dos Santos, A., Du Toit, J., Masenge, A., Van Aardt, I. and Wagner, C. (2014). *Research Methodology*. Oxford: Oxford University Press.

Buchanan, B.G. (2006). 'A (very) brief history of artificial intelligence.' *AI Magazine,* 26(4).

Bureau of Safety and Environmental Enforcement (BSSE), (2011). *BOEMRE releases report of investigation on BP's Atlantis Platform* [Online] Available at: https://www.bsee.gov/site-page/bp-atlantis-report-march-2011 (Accessed: 28th April 2016).

Burgess,T.F., McKee, D. and Kidd, C. (2005). "Configuration management in the aerospace industry: a review of industry practice*." International Journal of Operations & Production Management,* 25(3), pp. 290-301.

Cabantous, L., Gond, J-P. and Johnson-Cramer, M. (2010). 'Decision theory as practice: crafting rationality in organisations.' *Organisation Studies,* 31(11), pp. 1531-1566.

Capital Facilities Information Handover Specification (CFIHOS), (2017). http://uspi-global.org/index.php/projects/frameworks-methodologies/136-cfihos(Accessed: 11th February, 2017).

CEN: The European Committee for Standardization, (2009). *CEN Workshop Agreement, CWA 16180-1. 2009. The CEN ORCHID Roadmap Standardising Information Across the Plant Engineering Supply Chain - Part 1: Direction and Framework*.

Constantinou, A., Fenton, N. and Neil, M. (2016). 'Integrating Expert Knowledge with Data in Bayesian Networks: Preserving Data-Driven Expectations with the Expert Variables Remain Unobserved.' *Expert Systems with Applications*, 56: pp. 197-208.

Constantinou, A., Fenton, N., Marsh, W. and Radlinski, L. (2016). 'From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decisions support.' *Artificial Intelligence in Medicine,* 67, pp. 75-93.

Coopers & Lybrand (1997). POSC/CAESAR for better business – report to the POSC/CEASAR initiative. Oslo.

Cotton, D., Grissom, M., Spalding, D. and Want, R. (2012). *Standardization Barriers in the Petroleum Industry*. Boulder, CO: (M Thesis) University of Colorado.

Eckerson W.W. (2002). *Data Quality and the Bottom Line*. Chatsworth, CA: 101 Communications LLC, The Data Warehousing Institute.

Electric Power Research Institute (EPRI), (2014). *Data-Centric Configuration Management for Efficiency and Cost Reduction: An Economic Basis for Implementation*. Palo Alto, CA. 3002003126.

Eppler, M. and Helfert. M. (2004). 'A classification and analysis of data quality costs.' *MIT International Conference on Information Quality, November 5-6, 2004*, Boston, MA.

European Process Industries STEP Technical Liaison Executive (EPISTLE), (1998a). *Process Industries Data Handover Guide Part 2.* [Online]. Available: https://www.posccaesar.org/raw-attachment/wiki/.../HandoverGuide (06 February 2017)

European Process Industries STEP Technical Liaison Executive(EPISTLE), (1998b). *Process Industries Data Handover Guide Part 1.* [Online]. Available: https://www.posccaesar.org/raw-attachment/.../HandoverGuide (29 April 2016)

Farrington-Darby, T. and Wilson, J.R. (2006). 'The nature of expertise: a review.*' Applied Ergonomics,* 37(1), pp. 17-32. Elsevier.

Fess, E.E. 'Evaluating published research: method-statistical instrument selection.' *Journal of Hand Therapy,* October-December 1991, pp.181-182.

Fogel, R.W. (1979). 'Notes on the social saving controversy.' *The Journal of Economic History,* 39(1), pp. 1-54.

Fouhy, K. (1998). 'Managing plant data.' *Chemical Engineering*, 105.1, pp. 135-138.

Gackowski, Z.J. (2009). 'Information quality survey seen from the perspective of operations.' *International Conference on Computing, Engineering and Information, IEEE 2009*.

Gallaher, M.P., O'Connor, A.C., Dettbarn Jr., J.L. and Gilday, L.T. (2004). *Cost analysis of inadequate interoperability in the U.S. capital facilities industry*. U.S. Department of Commerce Technology Administration. NIST GCR 04-867. Gaithersburg, MD.

Ghosh, D. and Crain, T.L. (1993). 'Structure of uncertainty and decision-making: an experimental investigation.' *Decision Sciences,* 24(4).

Gordon, N.J., Salmond, D.J. and Smith, A.F.M (1993). 'Novel approach to nonlinear/non-Gaussian Bayesian state estimation.' *Radar and Signal Processing, IEE Proceedings F*. 140(2), pp. 107-113.

Grant, M.R. (2013). 'The Development of Knowledge Management in the Oil and Gas Industry.' *Universia Business Review*, pp. 92-125, Cuarto Trimestre

Ha, S. (2009). 'Finding Time to Find data.' *Industrial Engineering*, pp. 42, 12.

Haider, A. (2008). 'Conceptual and operational limitations of evaluating IS for engineering asset management.' *PACIS 2008 Proceedings* Paper 244 2008, AISeL.

Hamburg, M. (1974.) *Basic statistics-a modern approach*. 3rd Edition, Harcourt Brace Jovanovich. ISBN 0-15-505113-X.

Hamby, T. and Taylor, W. (2016). 'Survey satisficing inflates reliability and validity measures: an experimental comparison of college and amazon mechanical turk samples.' *Educational and Psychological Measurement*, 76(6), pp. 919-932.

Haney, M. (2016). 'Teaching the concept of investment risk through spreadsheet Monte Carlo simulations.' *Journal of the Academy of Business Education,* Winter 2016, pp. 236-256.

Harrison, R.L. (2010). *Introduction to Monte Carlo Simulation*. U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2924739/ (Accessed: May 5, 2017).

Haug, A., Arlbjørn, J.S., Zachariassen F. and Schlichter J. (2013). 'Master data quality barriers: an empirical investigation.' *Industrial Management & Data Systems*, 113(2), pp. 234-249.

Haug, A., Zachariassen, F., Van Liempd, D. (2011), 'The costs of poor data quality.' *Journal of Industrial Engineering and Management*, 2011-4(2), pp. 168-193.

Hawking, S. (2015). 'Stephen Hawking warns artificial intelligence could end mankind.' *BBC News*. Retrieved October 30, 2015.

Hayes-Roth, F., Waterman, D., Lenat, D. (1983). *Building Expert Systems*. Addison-Wesley. ISBN 0-2-1-10686-8.

Hendry, D.F. (1984). 'Monte Carlo experimentation in econometrics.' In *Handbook of Econometrics,* 2. (Editors Griliches, Z. and Intriligator, M.D. Elsevier, Amsterdam.

Hines, W.W. and Montgomery. D.C. (1980). *Probability and Statistics in Engineering and Management Science*. 2nd Edition. Wiley and Sons.

Hober,S., Burbank,D., Bradley, C. and Pomraning, M. 2009. *Data Modelling for the Business*. Bradley Beach, NJ: Technics Publications.

Institute of Nuclear Power Operations (INPO), (1985). *A maintenance analysis of safety significant events*. Nuclear Utility Management and Human Resources Committee, Maintenance Working Group. Institute of Nuclear Power Operations. Atlanta, GA.

Internal Revenue Service (2017). Available at: https://www.irs.gov/pub/irs-drop/n-99-18.pdf, (Accessed: 19th April 2017).

International Standards Organisation (ISO). 2003. *ISO 15926-2:2003. Industrial automation systems and integration – integration of life-cycle data for process plans including oil and gas production facilities.* [Online]. www.iso.org. (Accessed 29th April 2016).

International Standards Organisation (ISO), (2006). *ISO 14224:2006. Petroleum, petrochemical and natural gas industries — collection and exchange of reliability and maintenance data for equipment.* 2nd Edition [Online]. www.iso.org. (Accessed: 29th April 2016).

Johnson, M.E. (1987). *Multivariate statistical simulation*. New York, John Wiley.

Jooste, J.L. (2014). A critical success factor model for asset management services. Doctoral Thesis, the University of Stellenbosch.

Kim, W. and Choi, B. (2003). 'Towards quantifying data quality costs.' *Journal of Object Technology*, 2(4), pp. 69-76.

Klein, B.D. (2000). 'The detection of data errors in computer information systems: field interviews with municipal bond analysts.' *Information Resources Management Journal*, Jul-Sep 2000, 13(3), p. 23.

Kohli, R., Johnson, S. 2011. "Digital transformation in latecomer industries: CIO and CEO Leadership from Encana Oil & Gas (USA) Inc.' *MIS Quarterly Executive*, 10(4).

Kroese, D. P., Brereton, T., Taimre, T. and Botev, Z. I. (2014). 'Why the Monte Carlo method is so important today.' *WIREs Comput Stat*. 6 (6), pp. 386–392.

Langley, P. (2011). 'The Changing Science of Machine Learning.' *Machine Learning*, 82, pp. 275-279.

Layer, C., McLeod, M. and Blythe, S. 2013. 'Online Survey design and development: a Janus-faced approach.' *Written Communication,* 30(3), pp. 330-357.

Li, H. and Yamanishi, K. (2001). 'Mining from open answers in questionnaire data.' *KDD '01 Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. San Francisco, California — August 26 - 29, 2001, pp.443-449.

Lorenzo, D.K. and Van Den Heuval, L. (2000). 'Preventing human error.' *American Institute of Chemical Engineers*, Course 579, 2000.

Malhotra, V., Lee, M.D. and Khurana, A. (2007). '*Domain experts influence decision quality: Towards a robust method for their identification.' Journal of Petroleum Science and Engineering*, 57, pp. 181-194. Elsevier.

Mannila, H. (1996). 'Data mining: machine learning, statistics and databases.' *International Conference. Scientific and Statistical Database Management*. IEEE Computer Society.

Marsh, R. (2005). 'Drowning in dirty data? It's time to sink or swim: a four-stage methodology for total data quality management.' *Database Marketing & Customer Strategy Management*, 12 (2), pp. 105-112.

Mooney, C.Z. (1997). *Monte Carlo Simulation*. Thousand Oaks, CA: Sage Publications.

Mukhopadhyay, T., Kekre, S. and Kalathur, S. (1995). Business Value of Information Technology: A Study of Electronic Data Interchange. *MIS Quarterly*, June, pp. 137-156.

Naylor, T.J., Balintfy, J.L., Burdick, D.S. and Chu, K. (1966) *Computer simulation techniques*. New York, 1966. John Wiley & Sons.

Neeley, M.P., Lin, S. and Koronios, A. (2006). 'The deficiencies of current data quality tools in the realm of engineering asset management.' Proceedings of the *Twelfth Americas Conference in Information Systems*.

Nickerson, R.C., Varshney, U. and Muntermann, J. (2013). *'A method for taxonomy development and its application in information systems.' European Journal of Information Systems*, 22, pp. 336-359. Operational Research Society Ltd.

Noller, D., Myren, F., Haaland, O., Brisco, J. and Bryan, E.W. (2012). 'Improved decision-making and operational efficiencies through integrated production operations solutions.' *Offshore Technology Conference*, Houston, TX, 30 April 3 to May 2012.

Novak, V. (2005). 'Are fuzzy sets a reasonable tool for modelling vague phenomena?' *Fuzzy Sets and Systems,* 156, pp. 341-348. Elsevier.

Olson, D.L., Moshkovich, H.M., Schellenberger, R. and Mechitov, A.I. (1995). 'Consistency and accuracy in decision aids: experiments with four multiattribute systems.' *Decision Sciences*, 26(6), p. 723.

Prawel, D. (2003). Quoted in *Strategic Directions*, 20(5), pp. 31–33. Emerald Group Publishing Ltd. (2004).

Peytchev, A., Couper, M.P., McCabe, S.E. and Crawfors, S.D. (2006). 'Web survey design – paging versus scrolling.' *Public Opinion Quarterly*, 70(4), Winter 2006, pp. 596–607.

PriceWaterhouseCoopers (2002). *Global data management survey*. London: Citreon Wolf Communication.

Reason, J. (1990). *Human error*. Cambridge: Cambridge University Press.

Reason, J. (1998). *Managing the risks of organisational accidents*. Ashgate.

Redman, T.C. (1998). 'The impact of poor data quality on the typical enterprise.' *Communications of the ACM*, 41(2), pp. 79-82.

Ring, M.J. (1997). 'POSC and the emerging E&P business perspective.' *15th World Petroleum Conference*, Beijing, October 12-17, 1997.

Rubinstein, R.Y. and Kroese, D.P. (2017). *Simulation and the Monte Carlo method*. 3rd Edition, New Jersey: John Wiley & Sons, Hoboken.

Russel, S.J., Norvig, P. (2003). *Artificial intelligence: a modern approach*. 2nd Edition, ISBN 0-13-790395-2. Upper Saddle River, New Jersey: Prentice Hall.

Russel, S.J., Norvig, P. (2009). *Artificial intelligence: a modern approach*. 3rd Edition, ISBN 0-13-604259-7Upper saddle River, New Jersey: Prentice Hall.

Sawilovsky, S.S. (2003). 'You think you've go trivials?' *Journal of Modern Applied Statistical Methods*, 2(1), pp. 218-225.

Schenk, R.W. (1985). 'Configuration management as applied to engineering projects.' *J. Manage. Eng*., 1(3), pp. 157-165, 1985.

Smith, D.J. (1997). *Reliability, maintainability and risk*. 5th Edition. ISBN 0 7506 3752 8 Butterworth Heinemann.

Society of Petroleum Engineers (SPE), (2017). *Unit conversion factors*. Available at: http://www.spe.org/industry/unit-conversion-factors.php. (Accessed: 19th April 2017).

Spangler, I.S. and Pompper, D. (2011). 'Corporate social responsibility and the oil industry: Theory and perspective fuel a longitudinal view.' *Public Relations Review*, 37, pp. 217-225. Elsevier.

Spear, M. and Leis, M. (1997). 'Artificial neural networks and the accounting method choice in the oil and gas industry.' *Accting, Mgmnt & Info.Tech*., 7(3), pp. 169-181.

Speier, C., Vessey, I. and Valacich, J.S. (2003). 'The effects of interruptions, task complexity and information presentation on computer-supported decision-making performance.' *Decisions Sciences*, 34(4), Fall 2003, p. 771.

Tam, A.S.B. and Price, J.W.H. (2008). 'A generic asset management framework for optimising maintenance investment decision.' *Production Planning & Control*. 19(4), pp. 287–300, June 2008.

Thawesaengskulthai, N., Tannock, J.D.T. (2008). "A decision aid for selecting improvement methodologies." *International Journal of Production Research*, 46(23), pp. 6721-6737.

The Bureau of Ocean Energy Management, Regulation and Enforcement (BOEMRE), (2011). "Report Regarding the Causes of the April 20, 2010 Macondo Well Blowout." https://repository.library.noaa.gov/view/noaa/279. (Accessed: 11th February 2017).

Thomopoulos, N.T. (2013). *Essentials of Monte Carlo simulation, statistical methods for building simulation models,* DOI: 10.1007/978-1-4614-6022-0. New York. Springer Science + Business Media.

University of North Caroline at Chapel Hill (2007). www.enc.edu. *Introduction to Knowledge Management*. From Wikipedia. (Accessed: 18th April 2017).

Van den Honert, A. (2014). *Estimating the continuous risk of accidents occurring in the South African mining industry*. (M.Eng. Thesis) Stellenbosch: University of Stellenbosch.

Vayghan, J.A., Garfinkle, S.M., Walenta, C., Healy, D.C. and Valentin, Z. (2007). "The internal information transformation of IBM." *IBM Systems Journal*, 46(4), pp. 669-684.

Walls, M.R. (ca. 2003). *Managing risks and strategic decisions in petroleum exploration and production.* Petroleum Institute for Continuing Education http://peice.com/coursedetails.aspx?course=1885

Wand, Y., Wang, R.Y. (1996). 'Anchoring data quality dimensions in ontological foundations.' *Communications of the ACM*, 39(11), pp. 86-95.

World Bank. 2015. *Global economic prospects*, Chapter 4, pp. 155-168. https://www.worldbank.org/content/dam/Worldbank/GEP/GEP2015a/pdfs/GEP2015a_chapter4_report_oil.pdf (Accessed: 29th April 2016).

Zastron, C.M. (2016). *Improving information reporting in data-intensive organisations by determining individual data presentation preferences*. (M.Eng. Thesis) Stellenbosch: University of Stellenbosch.

Appendix A – Complete Impact Element List

Impact Element	**Impact Element**
Ability to predict asset remnant lifecycle	IT Standard: shorter time to market
Asking for the same data from vendors	IT: higher degree of mutual understanding
Asset Database is foundational to Strategic Decisions and Optimization	IT: less time to mobilize projects
Asset need redefinition	IT: reduced IT costs
Asset operation profiling	KM Definition: converting tacit into explicit knowledge
Asset redesign/rehabilitation	KM gained at each stage of the asset lifecycle
Being blackmailed by a vendor	KM: Difficult to standardize with multiple simultaneous projects
Better environment	KM: Multiple classifications
Better reliability	KM: Time needed to update Engineering modifications down 75%
Better safety	KM: Time required to solve difficult operational problems cut by 95%
Building new plants faster	Lack of common interpretation
Business value chain integration	Lack of situational awareness
Changing asset management strategies to condition-based	Less commissioning activities
Construction: Avoid: IT support for redundant systems	Less lead time to get PO reversed
Construction: Avoid: Data translation cost	Less variation orders
Construction: Avoid: Inefficient business processes	Limited re-use of knowledge
Construction: Avoid: IO R&D	Loss of agility
Construction: Avoid: Productivity and training loss	Low efficiency transfers between FEED/Construction/CSU
Construction: Avoid: Redundant systems	Lower costs
Construction: Mitigation: Design & Construction Verify	Maintenance and other economic trade-offs
Construction: Mitigation: Manual re-entry	Making information more available through common databases
Construction: Mitigation: reworking design files	Ongoing mapping activities
Continuous improvement of the asset management plan	Ops: Mitigation: Redundant info transfer
Continuous verification that all stakeholders have the same info	Optimized decision support
Continuously reinventing the wheel due to lack of standards	Outage Planning: Outage Schedules
Cost of building a plant could be reduced by 10%	Outage Planning: Work Week Planning
Cost of low IO: Acquisition of redundant systems	Output Dimension: Predictability
Cost of low IO: Cost avoidance	Output Dimension: Unplanned Downtime
Cost of low IO: Translation costs between systems	Planners required more info than necessary - extra workload
Cost of Maintenance reduced by 10%	Planning for support resource
Data-to-information-to-decisions	Planning/Engineering/Design: Avoid: IT support for redundant systems
Delays: late penalties	Planning/Engineering/Design: Avoid: Data translation cost

Impact Element	Impact Element
Demonstrate solid proof of software value	Planning/Engineering/Design: Avoid: Inefficient business processes
Design Changes: Document change only	Planning/Engineering/Design: Avoid: I-O R&D
Design changes: Major Mods	Planning/Engineering/Design: Avoid: Productivity and training loss
Design Changes: Minor Mods	Planning/Engineering/Design: Avoid: Redundant systems
Design Changes: Set point change	Planning/Engineering/Design: Mitigation: Design & Construction Verify
Detailed work processes are hindered by lack of definition	Planning/Engineering/Design: Mitigation: Manual re-entry
Different interpretations of data meaning	Planning/Engineering/Design: Mitigation: reworking design files
Divergences from standard procedures	Predictive modelling
Engineering Evaluation: Operability Determinations	Process Safety Risk
Engineering Evaluation: Operating Experience Evaluations	Procurement Engineering: Commercial Grade Dedication
Engineering Evaluation: Procedures Changes	Procurement Engineering: Equivalency Evaluations
Engineering Evaluation: Work Week PRA Coding	Quality of informed choices
Engineering Evaluations: Condition Report	Recued cost of Quality
Engineering Evaluations: Contractor oversight	Recued cost to develop Final Investment Decision
Engineering Evaluations: Field Change Notices	Recued interface management costs
Engineering Evaluations: Prepare response to urgent request	Recued volume of surplus parts
Engineering Programmes: Environmental quantification	Reduced backlog of updating information
Engineering Programmes: External Events	Reduced call off times from contracts
Engineering Programmes: License Renewal/Ageing Management	Reduced design time
Engineering Programmes: Mitigating System Performance Index	Reduced Engineering budget by 15%
Enhancement in solution design	Reduced execution time
Enhancing competitiveness	Reduced handover costs
Environmental and regulatory concerns	Reduced information volumes: distribution
Estimated savings 2-3% of investment	Reduced information volumes: duplication
Ethical risks	Reduced information volumes: less revisions
FLNG $12m cost avoidance	Reduced number of systems
FTEs building PMs for TAR outside normal system	Reduced time and cost of custom programming
Future suitability	Reduced time in data exchange
Greater predictability when changing sites	Relationships with third parties
Hard = reduction in headcount – efficiency	Repetition of data across sources
Hard = reduction in headcount – productivity	Resources Dimensions: Spares
Hard = reduction in headcount - quality	Resources Dimensions: Facilities and Tools
Hard benefits to financial statements	Resources Dimensions: Headcount
Hard: MH reduce due to enhanced data retrieval 1: assure data is correct	Risk Dimension - Design Risk
Hard: MH reduce due to enhanced data retrieval 1: id other disciplines	Risk Dimension - Maintenance Risk
Hard: MH reduce due to enhanced data retrieval 1: provide data	Risk Dimension - Regulatory Compliance

Impact Element	Impact Element
Hard: MH reduce due to enhanced data retrieval 1: provide data to resolve conflict	Risk Management
Higher quality design due to more time being available in Appraise/Select stages	Saving of Engineering time
Improved analysis	Simplifying documentation
Improved capabilities of applications available in the marketplace	Soft qualitative benefits: behaviour
Improved collaboration and data sharing & integration with partners	Soft qualitative benefits: culture
Improved concurrent Engineering between EPC and main suppliers	Soft qualitative benefits: intellectual capital
Improved MoC time	Soft: Errors
Improved start-up efficiency due to information being available to Operations sooner	Soft: Insurance Premiums
Improved working relationships	Soft: CRM
Increased standardization	Soft: Delays
Integration of all activities - 15% cost, 25% time	Soft: Reduction in Regulatory Risk
IO and IM Standards	Soft: Reduction in Regulatory Scrutiny
IO: Better decision making	Soft: Reduction in Unplanned Downtime
IO: CAD to prototype to tooling rework cost	Soft: Rework
IO: cost of checking data	Soft: Stakeholder satisfaction
IO: cost of delay - Profit loss due to delay of revenues	Soft: work process redesign: electronic mock-ups
IO: cost of delay at handover - most difficult to quantify	Soft: work process redesign: better resource loading
IO: cost of delay between phases - most difficult to quantify	Soft: work process redesign: enhanced collaboration
IO: cost of manual re-entry of data	Standards: Cost of change of software vendors
IO: cost of rework	Systems Engineering: Primary/BOP/El/etc.
IO: delays between Engineering systems	Technology refresh
IO: Inefficiency between Engineering systems	Total potential savings of 4.2% of investment
IO: Mitigation cost - manual updates	Transformation of patterns of business
IO: outsourcing ability	Translation costs of data between software vendors
IO: redundant Construction	Unclear priorities
IO: Training CAD will improve IO	Verifiable automatic transfer between systems
IT Standard benefit: economic competitiveness	Work Planning: WO Preparation for non-Outage periods
IT Standard: decreasing supply chain cost	Work Planning: WO Preparation for Outages
IT Standard: Lower infrastructure vulnerability	Work Planning: WO Preparation for PMs
IT Standard: reducing duplication effort	Work prioritization

Appendix B – Sources of Impact Elements

American Petroleum Institute. 2016. API 580 Risk-Based Inspection, Third Edition. [Online]. Available: http://www.techstreet.com/api/searches/12764433

Arlbjørn, J.S., Wong, C.Y. and Seerup, S. 2007. Achieving competitiveness through supply chain integration. International Journal of Integrated Supply Management, 3(1): 4-24.

Burgess,T.F., McKee, D., Kidd, C. 2005. Configuration management in the aerospace industry: a review of industry practice. International Journal of Operations & Production Management. 25(3): 290-301, 2005.

Bryman, A., Bell, E., Hirschsohn, P., Dos Santos, A., Du Toit, J., Masenge, A., Van Aardt, I. & Wagner, C. 2014. Research Methodology. Oxford, Oxford University Press.

Coopers & Lybrand. 1997. POSC/CAESAR for better business. Oslo.

CEN: The European Committee for Standardization. 2009. CEN Workshop Agreement, CWA 16180-1. 2009. The CEN ORCHID Roadmap Standardising Information Across the Plant Engineering Supply Chain - Part 1: Direction and Framework.

Cotton, D., Grissom, M., Spalding, D.& Want, R. 2012. Standardization Barriers in the Petroleum Industry. Boulder, CO: (M Thesis) University of Colorado.

Electric Power Research Institute (EPRI). 2014. Data-Centric Configuration Management for Efficiency and Cost Reduction; An Economic Basis for Implementation. EPRI, Palo Alto, CA, 2014. 3002003126.

European Process Industries STEP Technical Liaison Executive. 1998. Process Industries Data Handover Guide Part 1. [Online]. Available: https://www.posccaesar.org/raw-attachment/.../HandoverGuide (29 April 2016)

Fogel, Robert W. March 1979. "Notes on the Social Saving Controversy." The Journal of Economic History 39(1):1-54

Fouhy, K. 1998. Managing Plant Data. Chemical Engineering, 105.1: 135-138, Jan 1998

Grant, M.R. 2013. The Development of Knowledge Management in the Oil and Gas Industry. Universia Business Review, 92-125,Cuarto Trimestre

Haider, A. 2008. Conceptual and Operational Limitations of Evaluating IS for Engineering Asset Management. PACIS 2008 Proceedings Paper 244 2008 : AISeL

Haug, A., Arlbjørn, J.S., Zachariassen F., Schlichter J. 2012. Master data quality barriers: an empirical investigation. Industrial Management & Data Systems, 113(2) : 234-249, 2013

INPO. A Maintenance Analysis of Safety Significant Events. Nuclear Utility Management and Human Resources Committee, Maintenance Working Group. Institute of Nuclear Power Operations. Atlanta, GA,1985

International Standards Organization. 2006. ISO 14224:2006. Petroleum, petrochemical and natural gas industries — Collection and exchange of reliability and maintenance data for equipment. Second edition [Online]. Available: www.iso.org (29 April 2016)

International Standards Organization. 2003. ISO 15926. Industrial automation systems and integration – Integration of life-cycle data for process plans including oil and gas production facilities. [Online]. Available: www.iso.org (29 April 2016)

Kohli, R., Johnson, S. 2011. Digital Transformation in Latecomer Industries: CIO and CEO Leadership from Encana Oil & Gas (USA) Inc. MIS Quarterly Executive, Vol. 10, No. 4, Dec. 4.

Marsh, R. 2005. "Drowning in dirty data? It's time to sink or swim: a four-stage methodology for total data quality management". Database Marketing & Customer Strategy Management, 12 (2):105-12, 2005

Mukhopadhyay, T., Kekre, S., Kalathur, S. 1995. Business Value of Information Technology: A Study of Electronic Data Interchange. MIS Quarterly,: 137-156, June 1995

Noller, D., Myren, F., Haaland, O., Brisco, J. & Bryan, E.W. 2012. Improved Decision-making and Operational Efficiencies through Integrated Production Operations Solutions. Offshore Technology Conference, Houston, TX, 30 Apr-3 May, 2012. OnePetro.

Prawel, D. 2003. Interoperability Best Practices. Strategic Directions, Vol. 20. No. 5,pp.31 – 33.2004

Reason, James. 1990. Human Error. Cambridge: Cambridge University Press

Redman, T.C. 1998. “The Impact of Poor Data Quality on the Typical Enterprise”, Communications of the ACM, 41(2) :79-82, 1998.

Ring, M.J. 1997. POSC and the Emerging E&P Business Perspective. 15th World Petroleum Conference, Beijing, 12-17 October 1997.

Schenk, R.W. 1985. Configuration Management as Applied to Engineering Projects. J. Manage. Eng., 1(3): 157-165, 1985.

Tam, A.S.B., Price, J.W.H. 2008. A generic asset management framework for optimising maintenance investment decision. Production Planning & Control. 19(4): 287–300, June 2008

Bureau of Safety and Environmental Enforcement. 2011. BOEMRE Releases Report of Investigation on BP’s Atlantis Platform [Online] Available: http://www.bsee.gov/BSEE-Newsroom/Press-Releases/2011, Accessed 28/04/2016

U.S. Department of Commerce Technology Administration. 2004. Gallaher, M. P., O’Connor, A.C., Dettbarn Jr., J.L. & Gilday, L.T. Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry. Gaithersburg, MD.

Van den Honert, A. 2014. Estimating The Continuous Risk Of Accidents Occurring In The South African Mining Industry. (M.Eng Thesis) Stellenbosch: University of Stellenbosh.

World Bank. 2015. Global Economic Prospects, Chapter 4, p155. https://www.worldbank.org/content/dam/Worldbank/GEP/GEP2015a/pdfs/GEP2015a_chapter4_report_oil.pdf, accessed 29 April 2016