TITLE

A Framework for Valuing the Quality of Customer Information

Gregory Hill

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

October 2009

Department of Information Systems
Faculty of Science
The University of Melbourne

ABSTRACT

This thesis addresses a widespread, significant and persistent problem in Information Systems practice: under-investment in the quality of customer information. Many organisations require clear financial models in order to undertake investments in their information systems and related processes. However, there are no widely accepted approaches to rigorously articulating the costs and benefits of potential quality improvements to customer information. This can result in poor quality customer information which impacts on wider organisational goals.

To address this problem, I develop and evaluate a framework for producing financial models of the costs and benefits of customer information quality interventions. These models can be used to select and prioritise from multiple candidate interventions across various customer processes and information resources, and to build a business case for the organisation to make the investment.

The research process involved:

- The adoption of Design Science as a suitable research approach, underpinned by a Critical Realist philosophy.
- A review of scholarly research in the Information Systems sub-discipline of Information Quality focusing on measurement and valuation, along with topics from relevant reference disciplines in economics and applied mathematics.
- A series of semi-structured context interviews with practitioners (including analysts, managers and executives) in a number of industries, examining specifically information quality measurement, valuation and investment.
- A conceptual study using the knowledge from the reference disciplines to design a framework incorporating models, measures and methods to address these practitioner requirements.
- A simulation study to evaluate and refine the framework by applying synthetic information quality deficiencies to real-world customer data sets and decision process in a controlled fashion.
- An evaluation of the framework based on a number of published criteria recommended by scholars to establish that the framework is a purposeful, innovative and generic solution to the problem at hand.

DECLARATION

This is to certify that:

i. the thesis comprises only my original work towards the PhD,

ii. due acknowledgement has been made in the text to all other material used,

iii. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Gregory Hill

ACKNOWLEDGEMENTS

I wish to acknowledge:

- The **Australian Research Council** for funding the research project,
- **Bill Nankervis** at **Telstra Corp.** for additional funding, guidance and industry access,
- my supervisor, **Professor Graeme Shanks,** for his powers of perseverance and persistence,
- my mother, **Elaine Hill,** for her long-standing encouragement and support,
- and finally, my partner, **Marie Barnard,** for her patience with me throughout this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

INTRODUCTION

1.1 OVERVIEW

Practitioners have long recognised the economic and organisational impacts of poor quality information (Redman 1995). However, the costs of addressing the underlying causes can be significant. For organisations struggling with Information Quality (IQ), articulating the expected costs and benefits of improvements to IQ can be a necessary first step to reaching wider organisational goals.

Information Systems (IS) scholars have been tackling this problem since the 1980s (Ballou and Pazer 1985; Ballou and Tayi 1989). Indeed, information economists and management scientists have been studying this problem since even earlier (Marschak 1971; Stigler 1961). Despite the proliferation of IQ frameworks and models during the 1990s from IS researchers (Strong et al. 1997; Wang 1995) and authors (English 1999), the IQ investment problem has seen relatively scant attention within the discipline.

This research project seeks to develop and evaluate a comprehensive *framework* to help analysts quantify the costs and benefits of improvements to IQ. The framework should cover the necessary definitions, calculations and steps required to produce a business case upon which decision-makers can base a significant investment decision.

The level of abstraction should be high enough that the framework is generic and can apply to a wide range of situations and organisations. It should also be low enough that it can produce useful results to help guide decision-makers in their particular circumstances.

1.2 BACKGROUND AND MOTIVATION

The research project partnered with Australia's leading telecommunications company, Telstra Corp. The industry sponsor was responsible for the quality of information in large-scale customer information systems supporting activities as part of a wider Customer Relationship Management (CRM) strategy. As such, the quality of information about *customers* was the focus for this project. This grounded the research in a specific context (organisational data, processes, systems and objectives) but one that was shared across industries and organisational types. Most organisations, after all, have customers of one sort or another and they are very likely to capture information about them in a database.

A second agreed focus area was the use of automated decision-making at the customer level to support business functions such as marketing campaigns, fraud detection, credit scoring and customer service. These kinds of uses were "pain points" for the sponsor and so were identified as likely areas for improvements in the underlying customer data to be realised. Again, these functions are sufficiently generic across larger organisations that the framework would not become too specialised.

The third principle agreed with the industry partner was that telecommunications would not be the sole industry examined. While arrangements were in place for access to staff in the sponsoring

organisation, it was felt important that approaches, experiences and practices from the wider community would benefit the project.

Lastly, the research project would not address the underlying causes of IQ deficiencies (eg. data entry errors, poor interface design or undocumented data standards) nor their specific remedies (eg. data cleansing, record linking or data model re-design). Instead, the focus would be on a framework for building the case for investing in improvements, independent of the systems or processes under examination. The industry partner was particularly interested in the benefit (or cost avoidance) side of the equation as the view was the costs associated with IQ projects were reasonably well understood and managed within traditional IS systems development frameworks.

Focusing the research on customer information used in customer processes struck the right balance between providing a meaningful context and ensuring the framework could produce useful results.

1.3 Outline of the Thesis

As the research project sought to produce and assess an *artefact* rather than answer a question, Design Science was selected as the most appropriate research approach. With Design Science, *utility* of a designed artefact is explicitly set as the goal rather than the truth of a theory (Hevner et al. 2004). So rather than following a process of formulating and answering a series of research questions, Design Science proceeds by building and evaluating an artefact. In this case, the framework is construed as an *abstract artefact,* incorporating models, measures and a method.

Before tackling the research project, some preliminary work must be completed. Firstly, further understanding of Design Science is required, especially how to distinguish between design as a human activity and Design Science as scholarly research. Further, a method for evaluating the artefact plus criteria for assessing the research itself must be identified. The philosophical position underpinning the research (including the ontological and epistemological stances) must be articulated, along with the implications for gathering and interpreting data. These issues are addressed in Chapter 2, Research Method and Design.

The third chapter (Literature Review) examines critically the current state of IQ research in regards to frameworks, measurement and valuation. The organisational context (CRM, in this case) and related measurement and valuation approaches (from information economics and others) are also examined.

In order to develop a useful artefact, it is necessary to understand what task the artefact is intended to perform and how the task is performed presently. This requires field work with practitioners who deal with questions of value and prioritisation around customer information. A series of semi-structured interviews was selected as the appropriate method here, yielding rich insights into the current "state of the art" including the limitations, difficulties and challenges arising from the existing practices (Chapter 4 – Context Interviews). Further, guidance about what form a solution to this problem could take was sought and this was used as the basis for practical requirements for the framework.

The theoretical knowledge from the Literature Review and the lessons from the Context Interviews were synthesised in Chapter 5 – Conceptual Study. This chapter is where the requirements of the framework are carefully spelled out and the core models and measures are proposed, defined and developed. An outline of the method is also provided.

To move from the development phases to the evaluation phase, Chapter 6 employs simulations and more detailed mathematical modelling to test empirically the emerging framework. This is done

using a realistic evaluation approach, exploring the effect of synthetic IQ deficiencies on real-world data sets and decision-processes. This results in a number of refinements to the framework, the development of a supporting tool and illustration of the method.

Finally, Chapter 7 – Research Evaluation encapsulates the framework (Avison and Fitzgerald 2002) and evaluates it against a set of criteria (Hevner et al. 2004). This is where the argument is made that the framework qualifies as Design Science research.

1.4 CONTRIBUTIONS OF THE RESEARCH

The research is an example of an applied, inter-disciplinary research employing qualitative and quantitative data collection and analysis. It is applied, in the sense that it identifies and addresses a real-world problem of interest to practitioners. It is inter-disciplinary as it draws upon "kernel theories" from reference disciplines in economics, machine learning and applied mathematics and incorporates them into knowledge from the Information Systems discipline. The collection and analysis of both qualitative data (from practitioner interviews) and quantitative data (from simulations) is integrated under a single post-positivist philosophy, Critical Realism.

The key contribution is the development, specification and evaluation of an abstract artefact (a framework comprising of models, measures and a method). This framework is grounded in an existing IQ framework, the *Semiotic Framework for Information Quality* (Price and Shanks 2005a) and extends the *Ontological Model for Information Quality* (Wand and Wang 1996) from the semantic level to the pragmatic. This model is operationalised and rigorously quantified from first principles using Information Theory (Shannon and Weaver 1949). The resulting novel IQ measures are used to identify and prioritise high-value candidate IQ interventions rapidly and efficiently.

At the core, this contribution stems from re-conceptualising the Information System as a communications channel between the external world of the customer and the organisation's internal representation of the customer. The statistical relationships between external-world customer attributes and those of the internal representation can be modelled using the entropy measures developed by Shannon in his Information Theory. In this way, the research builds on an existing rigorous IS theory and integrates an important "reference discipline" (Information Theory) in a novel way.

The next step is the use of these internal representations of customer attributes to drive organisational decision-making. By employing Utility Theory to quantify the costs and benefits of customer-level decision-making, the costs to the organisation of mistakes can be quantified. By identifying how representational errors cause mistaken actions, the value of improving IQ deficiencies can be calculated. Here, Utility Theory is used as a "reference theory" to develop a novel normative theory for how rational organisations should invest in the IQ aspect of their Information Systems.

Finally, a systematic and efficient framework (comprising models, measures and a method) for identifying and measuring these opportunities is developed and assessed. This is important in practice, as well as theory, as it means that the time and resources likely required to undertake such an analysis are not unfeasibly demanding.

The contributions to Information Systems theory are:

- the application of Utility Theory and Information Theory to address rigorously the value measurement problems in existing Information Quality frameworks,

- the use of Critical Realism in Design Science research as a way to incorporate qualitative data collection (for requirements) and quantitative data collection (for evaluation) within a unified and coherent methodology,

The contributions to Information Systems practice are:

- an understanding of how organisations fail to invest in Information Quality interventions,
- a framework for producing financial models of the expected costs and benefits of Information Quality interventions to help analysts make the case for investment.

Further, the financial models produced by the framework could also be used by researchers as the basis for an instrument in Information Quality research. For instance, they could be used to compare the efficacy of certain interventions, to quantify the impact of various deficiencies or to identify Critical Success Factors for Information Quality projects.

Chapter 2

Research Method and Design

RESEARCH METHOD AND DESIGN

2.1 SUMMARY

This research project employs a research approach known as *Design Science* to address the research problem. While related work predates the use of the term, it is often presented as a relatively new approach within the Information Systems discipline (Hevner et al. 2004) . Hence, this chapter explains the historical development of the approach, its philosophical basis and presents an argument for its appropriateness for this particular project as justification. Subsequent sections deal with the selection and justification of particular data collection (empirical) and analysis phases of the research:

1. Review of Relevant Literature
2. Semi-Structured Interview Series
3. Conceptual Study and Mathematical Modelling
4. Model Simulation Experiments
5. Research Evaluation

This project undertakes both qualitative (textual) and quantitative (numerical) data collection and analysis. A hybrid approach that encompasses both domains is a necessary consequence of building and evaluating a framework that entails the use of measurements by people in a business context.

2.2 INTRODUCTION TO DESIGN SCIENCE

While humans have been undertaking design-related activities for millennia, many authors – for example, Hevner et al. (2004) and March and Storey (2008) – trace the intellectual origins of Design Science to Herbert Simon's ongoing study of the *Sciences of the Artificial* (Simon 1996). Simon argues that, in contrast to the natural sciences of eg. physics and biology, an important source of knowledge can be found in the human-constructed world of the "artificial". The kinds of disciplines that grapple with questions of design include all forms of engineering, medicine, aspects of law, architecture and business (Simon 1996). In contrast to the natural sciences (which are concerned with *truth* and *necessity*), these artificial sciences are focused on *usefulness* and *contingency* (possibility). The common thread throughout these disparate fields is the notion of an *artefact*: the object of design could be an exchange-traded financial contract or a public transport system.

However, Simon argues that since the Second World War the validity of such approaches has succumbed to the primacy of the natural sciences. As a consequence, the artefact has been pushed into the background. Simon's work is in essence a call-to-arms for academics to embrace these artificial sciences and in particular, design as a means for undertaking research.

Since then, Design Science has been examined within Information Systems as a research method (Gregor 2006; Gregor and Jones 2007; Hevner et al. 2004; Jörg et al. 2007; Peffers et al. 2007) as well as used for conducting research on IS topics (Arnott 2006).

2.3 MOTIVATION

Firstly, I provide background and context for the project. The five steps outlined in the methodology from Takeda et al. (1990) form a natural way of presenting the history of the development of the project.

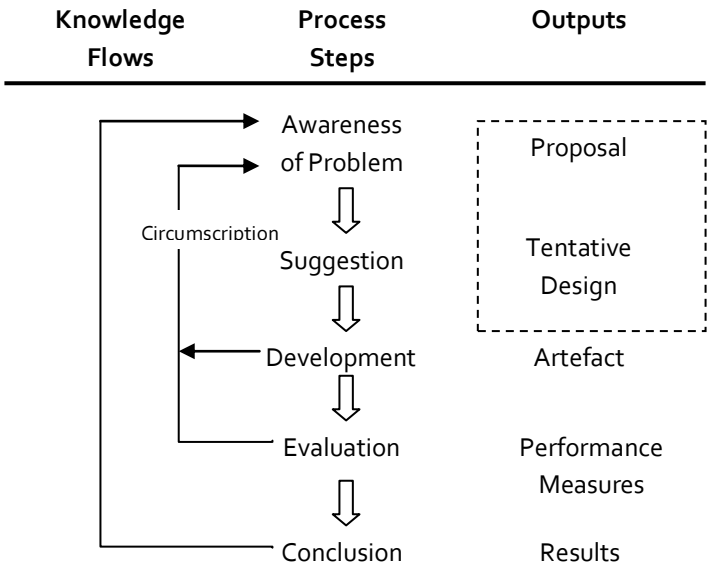

FIGURE 1 DESIGN SCIENCE RESEARCH PROCESS ADAPTED FROM TAKEDA (1990)

Firstly, *awareness of problem* came about through discussions with the industry partner and the academic supervisor. I identified that, while there are a number of theories and frameworks around Information Quality, none specifically addressed the question of valuing the improvements to information quality ie quantifying the "value-adding" nature of information quality to organisational processes. The industry partner was particularly keen to understand how to formulate a business case to identify, communicate and advocate for these improvements. The outcome of this step was an agreement between the University, supervisor, candidate and industry partner for an industry-sponsored doctoral research project.

The *suggestion* step was the insight that ideas (theories, constructs and measures) from the disciplines of Information Theory and Information Economics could prove beneficial in tackling this problem. These ideas are not readily transferable: it requires an understanding of the Information Quality literature, IS practice context, formalisation into an artefact and evaluation against some criteria. The output from this step was a doctoral proposal, accepted by the industry partner and academic institution as likely to meet the terms of the agreement.

The *development* and *evaluation* steps comprise the body of the empirical work in the project, and their rationale is outlined in this chapter. The output from the development step is the artefact for valuing information quality improvements. The output from the evaluation steps is the assessment of the artefact against recommended criteria.

Finally, the analyses and conclusions (including descriptions of the research process, empirical phases, the artefact itself and results of the evaluation) are embodied in the academic publications, including final thesis.

2.4 GOALS OF THE RESEARCH DESIGN

In order to tackle the customer information quality investment problem, it is important to understand what form a suitable response might take and how it might be used in practice. The over-riding consideration here is to *utility* rather than *truth*. That is, I am primarily concerned with producing a framework that is *useful* to practitioners and researchers as opposed to discovering an underlying truth about the world. The knowledge acquired is hence of an applied nature.

In this case, there must be a structured approach to building and evaluating the framework to ensure it has *rigour* and *relevance*. As Hevner et al. argue, IS research needs to be rigorous to provide an "addition to the knowledge base", and relevance allows for "application in the appropriate environment (2004)".

The question of whether IS research has favoured rigour at the expense of relevance has been discussed and debated widely throughout the IS research community. This debate was re-started most recently by commentary in the MISQ in 1999 by Benbasat and Zmud, arguing for increased relevance in IS research (1999). Their central thesis – that IS was too focused on gaining academic legitimacy through rigour, at the expense of practitioner legitimacy through relevance – was seized upon and other noted scholars joined the fray (Applegate 1999; Davenport and Markus 1999). Lee, for example, argued for the inclusion (and hence acceptance) of non-positivist approaches in IS research (1999). Robert Glass, writing an opinion piece in CAIS, reflects on his experiences to highlight the gulf between practitioners and academicians in the information systems world (2001).

Interestingly, Davenport and Markus argue that IS should model itself on disciplines like medicine and law to successfully integrate the rigour and relevance (1999). These are two examples of disciplines identified by Simon as employing the Design Science methodology (1996). In medicine and law (and related disciplines like engineering, architecture and planning), relevance and rigour are not seen as necessarily antagonistic and both goals may be pursued simultaneously through the two distinct "modes": develop/build and justify/evaluate. In this regard, Design Science picks up on an earlier IS specific approach known as *systems development* methodology (Burstein and Gregor 1999). Here, the research effort is centred on developing and evaluating a novel and useful information system, making a contribution to theory by providing a "proof-by-construction".

The main differences between the broader approach of Design Science and Information Systems Development are:

- Scope. Design Science is applicable to a much wider range disciplines than IS development. Indeed, Simon's conception of the Sciences of the Artificial spans medicine, architecture, industrial design and law (Simon 1996), in addition to technology-based fields.

- Artefact. Design Science takes a broader view of what constitutes an "artefact" for the purposes of research evaluation. Rather than just working instantiations, it also includes constructs, models, methods and frameworks.

In this case, the artefact is a framework for evaluating Information Quality improvements, in the context of Customer Relationship Management. So, where a Systems Development approach may be to build and test a novel system that identifies or corrects defects in customer information, a Design Science approach allows for focus on a more abstract artefact, such as a process or set of measures for evaluating such a system.

Some authors, such as Burstein and Gregor (1999), suggest that the System Development approach is a form of Action Research. It is reasonable to ask whether Design Science is also a form of Action Research. Here it is argued that this is not the case. Kock et al. propose a test for Action Research as being that where "intervention [is] carried out in a way that may be beneficial to the organisation participating in the research study" (Hevner et al. 2004; Kock et al. 1997).

Since I am not concerned with actually *intervening* in a particular organisation during this research, it should not be considered Action Research. Further, since there is no objective of *implementing* the method within the organisation, there is no imperative to trace the impact of the changes throughout the organisation – another aspect of Action Research (Burstein and Gregor 1999).

2.5 Employing Design Science in Research

The specific model of Design Science selected for use here is that presented by Hevner et al. (2004). This model was selected as it is well-developed, recent and published in the top journal for Information Systems. This suggests it is of high quality, accepted by researchers in this field and likely to be a reference source for a number of future projects. It also presents a number of criteria and guidelines for critically appraising Design Science research, which govern the research project.

This model makes explicit the two modes (develop/build and justify/evaluate) and links these to business needs (relevance) and applicable knowledge (rigour). This sits squarely with the applied nature of this project. I proceed by identifying the key elements from this generic model and map them to this specific project.

At this point it is useful to clarify the levels of abstraction. This project is not concerned with the information quality of any particular Information System (level 0). Neither is it concerned with methods, techniques or algorithms for improving information quality, such as data cleansing, data matching, data validation, data auditing or data integration (level 1). It is instead focussed on the description (or modelling) of such systems, techniques or algorithms *in a general way* that allows for comparison, appraisal, justification and selection (level 2). Lastly, in order to assess or evaluate this research itself, its quality and the degree to which it meets its goals, I employ Design Science. So, the prescriptions for evaluation within Hevner et al. pertain to *this research project* (level 3), not to the management of information quality (level 2). To recap the different levels of abstraction:

- Level 0. A particular Information System.
- Level 1. A specific method (or technique etc) for improving Information Quality within in Information Systems.
- Level 2. A framework for describing (and justifying etc) improvements to Information Quality within Information Systems.
- Level 3. A model for conducting (and evaluating) Design Science research.

With this in mind, I can proceed to map the elements in the model (level 3) to this research (level 2).

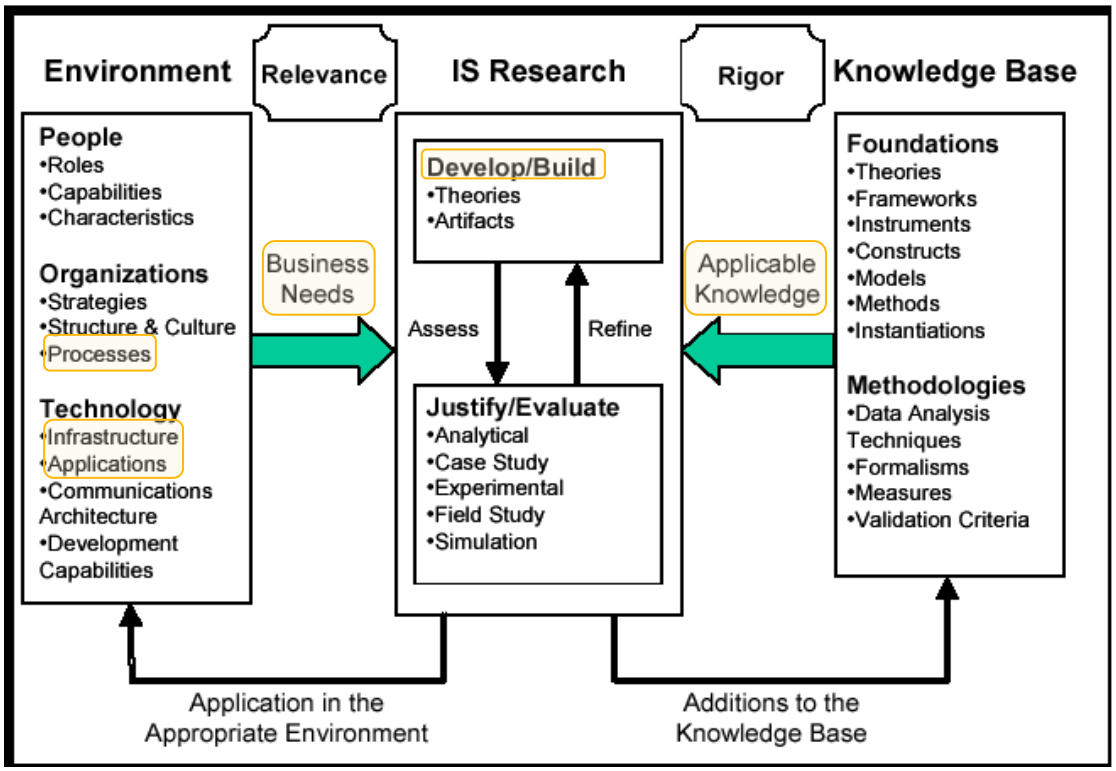

FIGURE 2 DESIGN SCIENCE RESEARCH MODEL (ADAPTED FROM HEVNER ET AL. 2004, P9).

2.5.1 BUSINESS NEEDS

I begin with the *business need*, which ensures the research meets the goal of relevance. Hevner et al. argue that the business need is "assessed within the context of organisational strategies, structures, culture and existing business processes". Hence, to understand the business need for an IQ evaluation framework I must examine these elements. If such a framework is developed but its assumptions or requirements are anathema to the target organisations then the framework will not be relevant. This also requires a careful definition of the "target organisations" to ensure that the scope is not so large that any commonalities in these elements are lost, nor so small that the research is too specific to be of wide use.

2.5.2 PROCESSES

From the research problem, it is clear that the target organisations must employ customer-level decision-making *processes* driven by extensive customer information. Examples of customer information include:

- information about the customer, such as date of birth, marital status, gender, contact details, residential and work locations and employment status,

- information about the customer's relationship with the organisation, such as histories of product purchases or service subscriptions, prior contacts (inquiries, complaints, support, marketing or sales), billing transactions, usage patterns and product/service preferences.

This information is sourced either directly from the customer, from the organisation's internal systems or from external information providers, such as public databases, partners or information service providers ("data brokers"). Of course, sourcing, storing and acting on this information is governed by the legal system (international treaties, national statutes and case law and local regulations), industry codes of practice, internal organisational policies and customer expectations.

Here, "customer-level decision-making" means that the organisation makes a decision about each customer, rather than treating all customers *en masse*. Examples of this include credit scoring and loan approval, fraud detection, direct marketing and segmentation activities. In each case, a business process is in place that produces a decision about each customer by applying business rules to that customer's information.

2.5.3 Infrastructure and Applications

The customer information is encoded and stored in large databases (data warehouses, data marts, operational data stores or other technologies), supported by computer *infrastructure* such as data storage, communication networks and operating environments. This infrastructure may be outsourced or provided in-house or shared between partners and suppliers.

The information is accessed (either stored or retrieved) by *applications* for Enterprise Resource Planning, Customer Relationship Management or Business Intelligence. These applications could be purchased "off-the-shelf" and customised or developed internally. People using these applications (and accessing the information) may be internal organisational staff, suppliers, partners, regulators or even the customers themselves.

Based on these key organisational and technological considerations, the IQ evaluation framework is targeted on IS-intensive, customer-facing service organisations. Examples of relevant service sectors include:

- financial services (personal banking, insurance, retail investment),
- telecommunications (fixed, mobile, internet),
- utilities (electricity, gas, water),
- government services (taxation, health and welfare).

Other areas could include charitable and community sector organisations, catalogue or subscription-based retailers and various customer-facing online business.

To ensure the IQ evaluation framework is relevant, the research design must include an empirical phase that seeks to understand the drivers of the business need (organisational and technological) in these target organisations.

2.5.4 Applicable Knowledge

In order for Design Science to achieve the objective of being *rigorous*, the research must draw on existing knowledge from a number of domains. "The knowledge base provides the raw materials from and through which IS research is accomplished … Prior IS research and results from reference disciplines provide [constructs] in the develop/build phase. Methodologies provide guidelines used in the justify/evaluate phase." (Hevner et al. 2004, p. 80)

Note that knowledge is drawn upon (in both phases) from prior IS research and reference disciplines. Design Science must also make "a contribution to the archival knowledge base of foundations and methodologies" (Hevner et al. 2004, p. 81). While this could conceivably include the reference disciplines, this is not required. There must, however, be a contribution to the IS knowledge base.

The point of access for this knowledge base varies with topic. In general, the IS research will be found in journal articles and conference papers as it is still emerging and being actively pursued by scholars. In addition, practitioner-oriented outlets may offer even more specific and current knowledge. The

reference discipline knowledge for this project, in contrast, is more likely to be in (older) textbooks as it is well-established, standardised and "bedded-in".

I begin mapping key elements of this model to the IQ evaluation framework by examining the specific IS research areas that form the knowledge base. From the research problem, it is clear that I am dealing with two sub-fields of Information Systems: Information Quality and Customer Relationship Management.

A number of Information Quality (IQ) models, frameworks, methods and theories have been proposed, analysed and evaluated in the IS literature (Ballou et al. 1998; Lee et al. 2002; Paradice and Fuerst 1991; Price and Shanks 2005a; Wang and Strong 1996). A solid understanding of existing IQ research, particularly for IQ evaluation, is required to avoid redundancy and misunderstanding. Fortunately, a large body of academic scholarship and practice-oriented knowledge has been built up over the past two decades or so. Importantly, the prospects of contributing back to this knowledge base are very good, as evaluation of information quality in the context of CRM processes is still an emerging area.

Customer Relationship Management (CRM) is a maturing sub-field of Information Systems, at the interface of technology and marketing. It has witnessed an explosion in research activity over the past ten years in both the academic and practitioner worlds (Fjermestad and Romano 2002; Romano and Fjermestad 2001; Romano and Fjermestad 2003). As a result, a significant amount of knowledge pertaining to theories, models and frameworks has accrued that can be drawn upon for this research project. Since customer information quality is flagged as a key determinant for CRM success (Freeman and Seddon 2005; Gartner 2003), it is likely that this research project will make a contribution to the knowledge base.

The next area to consider is the reference disciplines. This is the part of the knowledge base that provides a new perspective or insight to the problem that leads to 'building a better mouse trap'. Examples of Information Quality research employing reference disciplines include ontology (Wand and Wang 1996) and semiotics (Price and Shanks 2005a). In this research project, it is proposed that the reference disciplines include Information Theory (Shannon 1948) and Information Economics (Arrow 1984; Marschak 1974; Marschak et al. 1972; Theil 1967). These disciplines provide the foundational ideas for the "build phase", through their theories, models, formalisms (including notation) and measures.

Specifically, these reference disciplines provide very clear definitions of concepts such as entropy and utility. Additionally, these concepts can be communicated effectively to others through tried-and-tested explanations, representation and examples.

In light of the knowledge base, the research design must include a thorough review of existing knowledge in the IS research sub-fields (Information Quality and Customer Relationship Management) and the presentation of relevant material from the reference disciplines (Information Theory and Information Economics).

2.5.5 DEVELOP/BUILD

For a body of work to count as Design Science, it must produce and evaluate a *novel* artefact (Hevner et al. 2004). This has to be balanced by a need for IS research to be cumulative, that is, built on existing research where possible (Kuechler and Vaishnavi 2008). This project seeks to achieve this by taking the existing ontological IQ framework (Wand and Wang 1996) and extending it and re-interpreting it through the lens of Information Theory. In this way, it satisfies the requirement to be both cumulative and novel.

Also, I note that the artefact in Design Science does not have to be a particular system (Level 0, in the abstractions mapped out earlier) or technique (Level 1) but can be something more abstract (Level 2): in this case a framework for IQ valuation.

While March & Smith (1995) argue that constructs, models and methods are valid artefacts (March and Smith 1995), I need to be able to describe the proposed framework. To that end, I employ a modified form of the "Framework for Comparing Methodologies" developed by Avison & Fitzgerald (2002). While originally intended as a means for describing (and comparing) systems development methodologies, I argue that it is useful here for organising the ideas embodied in the valuation framework. The Avison & Fitzgerald framework can act as a "container" to describe the framework proposed here.

They outlined the following seven components:

1. Philosophy
    a. Paradigm
    b. Objectives
    c. Domain
    d. Target
2. Model
3. Techniques and Tools
4. Scope
5. Outputs
6. Practice
    a. Background
    b. Userbase
    c. Players
7. Product

Here, I will not use numbers six and seven since there is no practitioner group or instantiated product (the framework is still under development and evaluation). With this end in mind, the develop/build phase involves:

- synthesising a large body of knowledge (drawn from the IS research literature as well as the foundation or reference disciplines),
- acquiring a thorough understanding of the problem domain, organisational context and intended usage,
- assessing, analysing and extending the synthesised knowledge in light of this acquired understanding of the domain.

The next step is to subject the resulting artefact to the justify/evaluate phase.

2.5.6 JUSTIFY/EVALUATE

In order to ensure the artefact is both useful to practitioners (relevant) and contributing back to the IS knowledge base (rigorous), it must undergo stringent evaluation and justification.

Note that here I am not assessing the valuation of Information Quality improvements (Level 2), but rather assessing the artefact (framework) for doing this (Level 3).

Before I can justify/evaluate the framework, I need to clarify the nature of the claims made about it. For example, I could be stating that it is:

- necessarily the only way to value correctly IQ improvements,
- better – in some way – than existing approaches,
- likely to be preferred by practitioners over other approaches,
- may be useful to practitioners in some circumstances,
- is of interest to academics for related research.

These claims must be addressed in the formulation of the artefact during the develop/build phase, in light of the existing approaches and framework scope, and clearly stated.

While the precise claims cannot be stated in advance of the develop/build phase, the research problem make clear that the framework must satisfy two goals:

- Internal validity. It must allow for the modelling of a wide-range of organisational processes of interest. These models must conform to the foundational theoretical requirements, including representation, rationality assumptions and mathematical conventions.
- External validity. In order to be useful, the framework must be acceptable to the intended users in terms of its components (eg scope, outputs) but also explicable in its calculations, arguments and conclusions.

In other words, an artefact to help people quantify benefits must not only produce numerical results, but the users must have some confidence in those outputs and where they came from. Both of these goals must be met for this framework to be rigorous and thus likely to contribute to IS research.

With this in mind, I consider each of the evaluation methods prescribed by Hevner et al.

Evaluation Method	Description	Discussion
Observational	**Case Study:** Study artefact in depth in business environment.	Not possible since the IQ valuation framework has not been employed in an organisational setting.
	Field Study: Monitor use of artefact in multiple projects.	Would require deep access to IQ improvement projects, including to sensitive financial information (during business case construction) and customer information (during implementation). Not likely for an untested framework.
Analytical	**Static Analysis:** Examine structure of artefact for static qualities (eg complexity).	This approach would not meet the goal of external validity.
	Architecture Analysis: Study fit of artefact into a technical IS perspective.	This is method is not appropriate for an abstract framework.

	Optimisation: Demonstrate inherent optimal properties of artefact or provide optimality bounds on artefact behaviour.	This method relies on a clear optimality criterion or objective and accepted "figure-of-merit". This does not exist in this case.
	Dynamic Analysis: Study artefact in use for dynamic qualities (eg performance).	Again, performance criteria would need to be established as for optimisation.
Experimental	**Controlled Experiment:** Study artefact in controlled environment for qualities (eg. usability).	This is a promising candidate: I can generate evidence to support (or not) the artefact's utility. The results would also provide feedback to further refine the framework.
	Simulation: Execute artefact with artificial data.	Employing simulations (with artificial data) gets around the problem of access to real-world projects while still providing plausible evidence. Even better - for external validity - would be to use real-world data.
Testing	**Functional (Black Box) Testing:** Execute artefact interfaces to discover failures and identify defects.	The interfaces to the framework are not clearly defined and so this testing approach will not be sufficiently general.
	Structural (White Box) Testing: Perform coverage testing of some metric (eg. execution paths) in the artefact implementation.	Similarly, this approach suffers from a lack of a suitable metric for evaluating something as abstract as a framework.
Descriptive	**Informed Argument:** Use information from the knowledge base (eg relevant research) to build a convincing argument for the artefact's utility.	There is unlikely to be sufficient information in the knowledge base to convince practitioners and academics of the internal and external validity of the framework. It's more likely that practitioners would expect empirical evidence to be weighted against the claims.
	Scenarios: Construct detailed scenarios around the artefact to demonstrate its utility.	Another promising avenue to pursue since a contrived scenario grounds the artefact in a specific context without relying on an indefensible generalisation.

TABLE 1 POSSIBLE EVALUATION METHODS IN DESIGN SCIENCE RESEARCH, ADAPTED FROM (HEVNER ET AL. 2004)

2.6 Overall Research Design

With an understanding of the general Design Science approach and the particular needs of this research, I can now present the overall research design. I begin by outlining the philosophical stance I've taken (the nature of the world, how we acquire knowledge and our values in conducting research). Then, I show how each of the five empirical phases of the research project meet the requirements for doing Design Science. Lastly, I discuss the proposed research design in light of the research guidelines advocated by Hevner et al. to argue that this design is well-justified.

2.6.1 PHILOSOPHICAL POSITION

For this study, I have adopted "Critical Realism" (Bhaskar 1975; Bhaskar 1979; Bhaskar 1989). It's use in IS research has been advocated by a number of authors, including Mingers (Mingers 2000; Mingers 2004a; Mingers 2004b), Dobson (Dobson 2001), Smith (Smith 2006) and Carlsson (Carlsson 2003b; Carlsson 2005a; Carlsson 2005b), who has identified it as having a particularly good fit with Design Science. Similarly, Bunge posits that Design Science works best "when its practitioners shift between pragmatic and critical realist perspectives, guided by a pragmatic assessment of progress in the design cycle." (Vaishnavi and Kuechler 2004).

Broadly speaking, Critical Realism argues that there is a real-world, that is, that objects exist independently of our perception of them. However, it differs from so-called scientific realism (or naïve empiricism) in that it seeks "to recognise the reality of the natural order and the events and discourses of the social world." (Carlsson 2005a, p80). This is a very useful perspective, in the context of this research, as I outline.

Objects like Customer Relationship Management systems are complex socio-technical phenomena. At one level, they are manifestly real objects (composed of silicon, plastic and metal), whose behaviours are governed by well-understood physical laws (such as Maxwell's electromagnetic theory). At another level, they have been explicitly designed to implement abstractions such as microprocessors, operating systems, databases, applications and work flows. Lastly, CRM systems also instantiate categories, definitions, rules and norms – at the organisational and societal level. Examples include the provision of credit to customers, or the targeting of marketing messages.

It is not sensible to adopt a purely empiricist view to analyse such concepts as "customer", "credit" and "offer". Further, (social) positivism – with its emphasis on the discovery of causal relationships between dependent and independent variables through hypothesis testing – is not appropriate given the design-flavoured objectives of the research. In broad terms, the objective of positivism is *prediction*, whereas design science is concerned with *progress* (Kuechler and Vaishnavi 2008).

By the same token, it is important that the knowledge produced by the research is of a form acceptable to target users in the practitioner and academic communities. This means the IQ valuation framework will require a quantitative component, grounded in the norms of the mathematical and business communities. As such, philosophical positions that produce only qualitative models (such as hermeneutics, phenomenology and interpretivism in general) are unsuitable for this task. Critical Realism allows for the study of abstract phenomena and their interrelationships with both qualitative and quantitative modes of analysis:

"Put very simply, a central feature of realism is its attempt to preserve a 'scientific' attitude towards social analysis at the same time as recognising the importance of actors´ meanings and in some way incorporating them in research. As such, a key aspect of the realist project is a concern with causality and the identification of causal mechanisms in social phenomena in a manner quite unlike the traditional positivist search for causal generalisations." (Layder 1993).

I can now present the philosophical underpinning of the research project.

2.6.1.1 ONTOLOGY

The Critical Realist ontological position is that the real world (the domain of the real) is composed of a number of structures (called "generative mechanisms") that produce (or inhibit) events (the domain of the actual). These events are known to us through our experiences (the domain of the empirical). Thus, the real world is ontologically *stratified*, as summarised here:

	Domain of Real	Domain of Actual	Domain of Empirical
Mechanisms	X		
Events	X	X	
Experiences	X	X	X

TABLE 2 ONTOLOGICAL STRATIFICATION IN CRITICAL REALISM (ADAPTED FROM BHASKAR 1979)

Ontological assumptions of the critical realistic view of science (Bhaskar 1979). Xs indicate the domain of reality in which mechanisms, events, and experiences, respectively reside, as well as the domains involved for such a residence to be possible. (Carlsson 2003b, p329).

This stratification can be illustrated by way of example. Suppose that an experimenter places litmus paper in a solution of sulphuric acid. In this case, the event (in the domain of the actual) is the litmus paper turning red. We experience the colour red through our senses (domain of the empirical), but the generative mechanisms (ie the oxidisation of molecules and the resulting change in emission of photons) take place in the domain of the real. Bhaskar argues that:

[R]eal structures exist independently of and are often out of phase with the actual patterns of events. Indeed it is only because of the latter that we need to perform experiments and only because of the former that we can make sense of our performances of them (Bhaskar 1975, p13)

Here, the underlying mechanisms of chemistry would exist as they are without the litmus test being conducted. Since we cannot perceive directly the wavelengths of photons, we can only identify events in the domain of the actual. However, without the persistence of regularities within the domain of the real, it would not be possible to make sense of the experiments ie theorise about these generative mechanisms. The relationship between the domains of the actual and empirical are further expounded:

Similarly it can be shown to be a condition of the intelligibility of perception that events occur independently of experiences. And experiences are often (epistemically speaking) 'out of phase' with events - e.g. when they are misidentified. It is partly because of this possibility that the scientist needs a scientific education or training. (Bhaskar 1975, p13)

So, in this example, the experimenter must take into account that other events may interfere with the perception of the red colour on the litmus paper. Perhaps the experiment is conducted under (artificial) light, lacking a red component. Or maybe the red receptors in the experimenter's retina are damaged or defective.

It is the consideration of these kinds of possibilities that gives Critical Realism its "scientific" feel, while its rejection of the collapse of the empirical into the actual and the real (what Bhaskar calls the "epistemic fallacy") stops it being simply (naïve) empiricism. Similarly, Critical Realism differs from positivism in that it denies the possibility of the discovery of universal causal laws (invisible and embedded in the natural structure ie in the domain of the real) but instead focuses on the discernment of patterns of events (in the domain of the actual).

2.6.1.2 EPISTEMOLOGY

The epistemological perspective taken in this research could best be described as *coherence* during the build/develop phase and then *pragmatic* during the evaluation stage. This is not unexpected, as

[C]ritical realists tend to opt for a pragmatic theory of truth even though some critical realists still think that their epistemology ought to be correspondence theory of truth. Other critical realists prefer to be more eclectic and argue for a three-stage epistemology using correspondence, coherence and pragmatic theory of truth. (Kaboub 2002, p1)

The coherence theory of truth posits that statements are deemed to be knowledge (that is "justified true beliefs") if they are in accordance with ("cohere with") a broader set of knowledge, in this case from the reference disciplines of Information Theory and Information Economics. This fits well with the build/develop phase of Design Science, as applicable knowledge is drawn in from the knowledge base to construct the framework.

Later, during the justify/evaluate phase, the nature of knowledge claim shifts to a *pragmatic* theory of truth – in a nutshell, what's true is what works. Pragmatism, in epistemology, is primarily concerned with the consequences and utility (ie impact upon human well-being) of knowledge.

Pragmatism asks its usual question. "Grant an idea or belief to be true," it says, "what concrete difference will its being true make in anyone's actual life? How will the truth be realised? What experiences will be different from those which would obtain if the belief were false? What, in short, is the truth's cash-value in experiential terms?" The moment pragmatism asks this question, it sees the answer: TRUE IDEAS ARE THOSE THAT WE CAN ASSIMILATE, VALIDATE, CORROBORATE, AND VERIFY. FALSE IDEAS ARE THOSE THAT WE CANNOT. (James 1907, p201).

The emphasis here on utility, rather than truth, is appropriate given the goals of the evaluate/justify phase of Design Science: I seek to contribute back to the knowledge base a form of knowledge that is validated and useful (to practitioner and academic communities). From this perspective, justified true beliefs are knowledge that will work.

2.6.1.3 AXIOLOGY

The practice of research reflects on the underlying values of the various participants and stakeholders. In this case, the project is committed to conducting research ethically and in compliance with University statutes and regulations and the terms of the industry partner agreement. This means I must be ethical with all my dealings including research subjects, industry partners, academics and other stakeholders.

Further, I uphold the value of contributing to the knowledge base of the research community in an area with demonstrable need to practitioners, without consideration to potential commercial or other advantage to individuals or organisations. As such, the knowledge acquired must be placed into public domain, immediately, totally and without reservations.

2.6.2 BUILD/DEVELOP FRAMEWORK

In this section I present an outline of the research phases, why each phase is necessary and a rationale for each particular method's selection over alternatives. The goal is to show the overall coherence of the research design and how it fits with the requirements for Design Science of Hevner et al., as discussed in the preceding sections.

2.6.2.1 LITERATURE REVIEW

This first phase consists of gathering, assessing and synthesising knowledge through a review of literature. As discussed, rigour demands that Design Science research draw upon an existing knowledge base comprising the reference disciplines and accumulated knowledge in the IS domain. Further, the research project must be guided by the contemporary needs of IS practitioners in order to be relevant.

These requirements must be met by reviewing relevant literature from three broad sources:

- Current Information Systems research, comprising the top-rated scholarly journals, conference proceedings, technical reports and related publications. The authors are typically academics

writing for an audience of academics, postgraduate students and "reflective practitioners". This constitutes an important source of knowledge around methodology (Design Science for IS), Information Quality models and theories, and Customer Relationship Management systems and practices. This is also the knowledge base to which this project seeks to add.

- IS practitioner literature as found in practice-oriented journals, white papers, web sites and industry seminars. These authors are usually senior practitioners and consultants writing for others in their field. Knowledge from this source is useful for understanding the issues which concern practitioners "at the coal face", and how they think about them. It is important to understand their needs, as these people form one of the key audiences for the outcomes from this research project.

- Literature from the reference disciplines, in the form of textbooks and "seminal papers", is needed to incorporate that specific knowledge. While the authors and audience of these sources are also academics, it is not necessary to delve as deeply into this literature as with the IS research. This is since the reference discipline knowledge is usually much older (decades rather than years), has been distilled and codified and is now relatively static.

2.6.2.2 Interviews

The second phase is the development of a deep understanding of the business needs for IQ valuation. I argue that this is best achieved through a series of semi-structured interviews with analysts, consultants and managers in target organisations. This is since these people are best placed to explain the business needs around IQ that they have dealt with in the past, and how they have been met to date. They are also able to articulate the organisational strategies, cultural norms and business processes that will dictate the usefulness of any IQ valuation frameworks.

I considered and rejected two alternative approaches. Firstly, case studies would not be suitable owing to the "thin spread" of cases to which I would have access, combined with commercial and legal sensitivities involved in a very detailed examination of particular IQ valuation projects. I also wanted to maximise the exposure to different stakeholders (both by role and industry) given the time, resource and access constraints.

Secondly, surveys were deemed unsuitable for acquiring the kind of deeper understanding of business needs required for this research. A face-to-face conversation can elicit greater detail, nuance and context than a simple form or even short written response. For example, interviews allow the tailoring of questions to individual subjects to draw out their particular experiences, knowledge and perspectives; something that cannot be done readily with survey instruments.

2.6.2.3 Conceptual Study and Mathematical Modelling

The third phase is where the knowledge from the reference disciplines and IS domain (Literature Review) is brought to bear on the business needs elicited from the second phase (Context Interviews). The outcome is a conceptual model of Information Quality in organisational processes, amenable to mathematical analysis and simulation.

I argue that this can be characterised as a Conceptual Study since it involves the synthesis of disparate knowledge and key insights to argue for a re-conceptualisation of a familiar problem situation. Shanks et al. posit that:

Conceptual studies can be effective in building new frameworks and insights ... [and] can be used in current situations or to review existing bodies of knowledge. Its strengths are that it provides a critical

analysis of the situation which can lead to new insights, the development of theories and deeper understanding. (Shanks et al. 1993, p7)

This step is essential to the overall research design, in that it is where the framework is conceived and developed. The resulting artefact (a framework for IQ valuation) comprises a model (a set of constructs and the mathematical formulae defining and relating them) and some guidelines for practitioners to use analyse their particular system. This artefact must then be evaluated to understand its likely impact in practice and contribution to the knowledge base.

2.6.3 Justify/Evaluate Framework

2.6.3.1 Simulation Study

In order to evaluate the framework, I must put it to use to generate outputs that can be analysed. It is necessary to demonstrate that the framework can be employed and the results are intelligible.

I propose that computer simulations using *synthetic data* provide the best way of producing these results. By "synthetic data" I mean data from real-world scenarios made publicly available for evaluation purposes, which have had various kinds of information quality defects artificially introduced. The behaviour of the mathematical model (including the impact on outputs and relationships between constructs) can then be assessed in light of these changes.

Other methods considered included a field trial and mathematical proof of optimality. Problems with the former included the difficulty of getting access to a real-world IQ project (given the commercial and legal hurdles) and the scope (ie time and resource constraints would not allow examination of multiple scenarios). The second approach – formal proof – was considered too risky as it might not be tractable and such a proof might not be acceptable to the intended audience of practitioners and academics.

2.6.3.2 Evaluation by Argumentation

Lastly, the research process and resulting artefact must be evaluated against some criteria. It is not sufficient to rely on the statistical analysis of the simulation study as this will not take in a sufficiently broad view of the performance or suitability of the framework and ensuring that the research is indeed "Design Science" and not just "design". Of course, this stage hinges crucially on the selection of an appropriate set of criteria. Here, I've opted to use the guidelines published in MIS Quarterly (Hevner et al. 2004), perhaps the leading IS research publication, and heavily cited by other researchers in this field. Below is preliminary discussion on how the proposed research design meets these criteria.

An alternative evaluation method here would be a focus group of practitioners, as intended users. However, to seek practitioner opinions on likely use or adoption of the framework in the space of a few short hours would not be feasible. Providing sufficient knowledge of the proposed framework to elicit meaningful and thoughtful comments would require a large investment of time, not something that practitioners generally have in large amounts.

2.7 Assessment of Research Design

With an emerging Information Systems research approach, there is often some consternation about how one should assess the quality of the work. To go some way to meet this need, Hevner et al. were invited to develop some general guidelines for the assessment of Design Science. These guidelines are intended to be used by research leaders and journal editors.

This section describes their guidelines, and discusses how the research design presented here meets them.

Guideline	Description	Discussion
Design as an Artefact	Design Science research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation.	The IQ valuation framework produced during the development phase meets the criteria of an artefact, as it embodies a *construct* (conceptualisation of problem), a *model* (description of IS behaviour) and *method* (in this case, a socio-technical method for organisational practice).
Problem Relevance	The objective of Design Science research is to develop technology-based solutions to important and relevant business problems.	That the industry partner - and other practitioners - have provided time and resources to tackling this problem signals the extent to which they perceive the problem as important and relevant.
Design Evaluation	The utility, quality, and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods.	The artefact is evaluated by contriving scenarios with real data and decision processes with rigorous statistical analyses on results.
Research Contributions	Effective Design Science research must provide clear and verifiable contributions in the areas of the design artefact, design foundations, and/or design methodologies.	This research project identifies a clear gap in the existing IS knowledge base and seeks to fill it through the careful application of the appropriate research method (Design Science).
Research Rigour	Design Science research relies upon the application of rigorous methods in both the construction and evaluation of the design artefact.	While the construction process for Design Science artefacts is not widely understood (March and Smith 1995), this research design follows well-founded prescriptions from the IS literature (Hevner et al. 2004) for understanding business need (interviews) and the existing knowledge base (literature review).
Design as Search	The search for an effective artefact requires utilising available means to reach desired ends while satisfying laws in the problem environment.	Here, the artefact is bounded by organisational norms, assumptions and cultures and, to the extent practicable, seeks to understand these and operate within them.
Communication of Research	Design Science research must be presented effectively both to technology-oriented as well as management-oriented audiences.	Owing to the industry partnership and involvement with the wider IS practitioner community, the research outcomes are to be communicated to IS managers. Indeed, as information quality has visibility in the broader management world, these findings will be communicated more widely.

TABLE 3 GUIDELINES FOR ASSESSMENT OF DESIGN SCIENCE RESEARCH ADAPTED FROM(HEVNER ET AL. 2004)

2.8 CONCLUSION

This research project is concerned with developing and evaluating a novel instrument for valuing Information Quality in Customer Relationship Management processes. With this emphasis on a producing an artefact that is useful to practitioners, I argue that the most suitable research design is one employing Design Science. Critical Realism offers the best fit for a philosophical basis for this kind of research as it is "scientifically-flavoured", without being unduly naïve about social phenomena. The model of Design Science outlined by Hevner et al. is appropriate for my purposes and so I adopt their terminology, guidelines and assessment criteria.

Specifically, the build/develop phase employs a review of relevant literature (from academic and practitioner knowledge sources) and a series of semi-structured interview with key practitioners in target organisations. The framework itself is produced by a conceptual study synthesising this understanding of business need with applicable knowledge.

The justify/evaluate phase proceeds with a simulation study of the valuation framework using synthetic data, followed by a reflective evaluation examining the framework and simulation results.

Chapter 3

Literature Review

LITERATURE REVIEW

3.1 SUMMARY

This chapter reviews literature of relevance to the project, drawn from academic and practitioner sources. The purpose of the review is threefold:

- to identify the gaps in the existing Information Quality knowledge base that this project seeks to address,
- to present a specific organisational context for IQ valuation, in the form of Customer Relationship Management systems,
- to provide an overview of the reference disciplines which examine and measure value and uncertainty.

This kind of review is necessary in Design Science research to ensure that the research makes a contribution to the Information Systems knowledge base, is relevant to practitioners and makes correct use of the reference disciplines.

This chapter is organised into three sections, addressing the three goals outlined above: Information Quality, Customer Relationship Management and the information-centric reference disciplines.

3.2 INFORMATION QUALITY

Information Quality (IQ) is an Information Systems (IS) research area that seeks to apply modern quality management theories and practices to organisational data and systems. This involves building and applying conceptual frameworks and operational measures for understanding the causes and effects of IQ problems. Additionally, some research seeks to evaluate the impact of initiatives to improve IQ.

IQ is fundamental to the study and use of Information Systems. Yet it is not the principle focus of research or practice. Perhaps the most widely understood model of how IQ fits into IS more generally is the Delone and Mclean Model of IS Success (DeLone and McLean 1992; DeLone and McLean 2003; Seddon 1997).

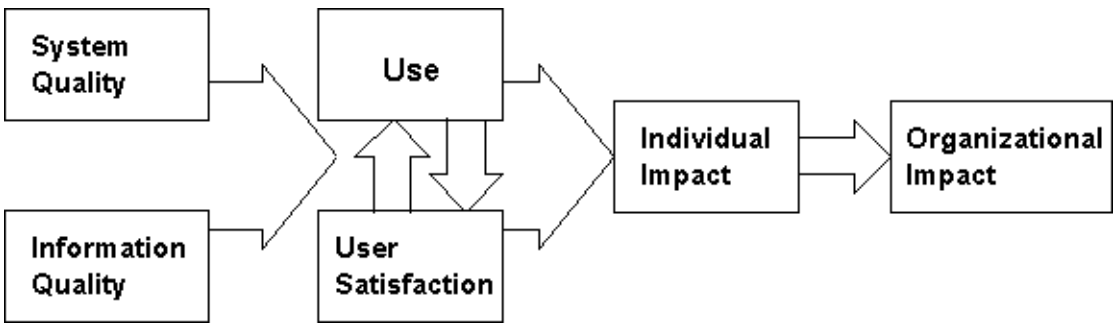

FIGURE 3 IS SUCCESS MODEL OF DELONE AND MCLEAN (DELONE AND MCLEAN 1992)

Here, IQ is understood to affect both *Use* and *User Satisfaction*, along with *System Quality*. This model's assumptions about the separation of content (Information Quality) from delivery (System

Quality), and about the individual vs organisational impact are discussed further below. However, it is a useful starting point owing to its widespread adoption and broad scope.

While IQ can be conceived as part of the IS Success sub-field, as an object of study it pre-dates Delone and Maclean's model. One notable general IQ researcher active during the 1980s is Donald Ballou (Ballou and Pazer 1985); (Ballou and Tayi 1985); (Ballou and Tayi 1989). Prior to this period, the research was either specific to certain fields such as auditing (Johnson et al. 1981) or related to specific techniques such as data-matching and integration (Fellegi and Sunter 1969).

Throughout the 1990s, IQ research increased with the proliferation of internet-based information-sharing, the deployment of enterprise systems such as data warehouses (DW) (Shankaranarayanan and Even 2004; Wixom and Watson 2001) and business intelligence (BI) and the importance of information-based business strategies such as enterprise resource planning (ERP) (Cai and Shankaranarayanan 2007) and customer relationship management (CRM) (Courtheoux 2003; Ishaya and Raigneau 2007; Miller 2005). During this period a number of authors (consultants and academics) wrote books and business journal articles for practitioners grappling with information quality problems (Becker 1998; English 1999; Huang et al. 1999; Marsh 2005; Orr 1998; Redman 1995; Redman 2008; Strong and Lee 1997; Tozer 1994)

Academic and practitioner researchers have produced several *generic IQ frameworks;* that is, they are intended to be applicable to a very broad class of information systems (Barone et al. 2007; Cappiello et al. 2006; Ge and Helfert 2008; Gustafsson et al. 2006; Joseph et al. 2005; Stvilia et al. 2007). Typically, these use a small number of components or dimensions of IQ to group a larger number of IQ criteria or characteristics. One early study listed 178 such IQ dimensions, criteria and goals (Wang et al. 1993), which illustrates the breadth of ideas encompassed within the Information Quality sub-discipline.

Some IQ research proceeds by examining one of these IQ concepts in isolation, such as *believability* (Pradhan 2005; Prat and Madnick 2008a; Prat and Madnick 2008b) or *timeliness* (Ballou and Pazer 1995; Cappiello et al. 2003). Another tack is to take a broader view of the concept of quality and how it relates to information (Batini and Scannapieco 2006; Fox and Redman 1994; Piprani and Ernst 2008; Sarkar 2002; Tayi and Ballou 1998; Welzer et al. 2007).

In contrast, another research stream examined IQ in the context of specific applications (Dariusz et al. 2007), such as accounting (Kaplan et al. 1998), security (English 2005; Wang et al. 2003), "householding[1]" (Madnick et al. 2004; Madnick et al. 2003) and undergraduate teaching (Khalil et al. 1999) as well as more traditional IS areas like conceptual modelling (Levitin and Redman 1995; Lindland et al. 1994; Moody and Shanks 2003; Moody et al. 1998), process design (Lee et al. 2004; Lee and Strong 2003; Strong 1997), metadata (Shankaranarayanan and Even 2004; Shankaranarayanan and Even 2006) and querying (Ballou et al. 2006; Motro and Rakov 1996; Parssian 2006; Wang et al. 2001).

Other researchers focused on the interaction between information quality and how it is used in decision-making by individuals, for example, in information-seeking behaviour (Fischer et al. 2008; Ge and Helfert 2007; Klein and Callahan 2007), decision quality (Frank 2008), information processing (Davies 2001; Eppler and Mengis 2004; Shankaranarayanan and Cai 2006) and visualisation (Zhu et al. 2007).

[1] *"Housholding" in the information quality context refers to the process of grouping related entities, for instance individuals who reside at the same house, or companies that fall under a shared ownership structure.*

The rest of this section is organised as follows. The next sub-section examines three important frameworks from the academic literature: the *Ontological Model* (Wand and Wang 1996), the *Semiotic Framework* (Price and Shanks 2005a) and the AIMQ (Lee et al. 2002). The first two are grounded in theory (ontology and semiotics, respectively) and adopt a "first-principles" approach to describe information systems (and deficiencies) in general. The third is empirically-based, drawing on the opinions of a pool of practitioners and researchers.

The subsequent sub-section addresses existing IQ measurement literature, including the different types of approaches endorsed by researchers (subjective and objective) and problems therein. Lastly, I consider a particular kind of measurement: valuation. Here I discuss the need for value-based (eg cost/benefit and investment-oriented) approaches to information quality assessment and critically examine past attempts at this.

3.3 EXISTING IQ FRAMEWORKS

3.3.1 AIMQ FRAMEWORK

The first framework I examine is the AIMQ (Lee et al. 2002). This framework has been selected as it is well-developed and a good exemplar of the empirical approach to IQ research. It also ties together a number of research projects arising from MIT's Total Data Quality Management (TDQM) project, lead by Professor Richard Wang. This program arose from Wang's group's view of information as a manufactured product (Ballou et al. 1998; Parssian et al. 1999; Wang et al. 1998) and that "total quality management" (TQM) principles – which had proved so successful in improving product quality for manufactured goods – could be applied to producing information goods (Dvir and Evans 1996; Wang 1998; Wang and Wang 2008).

The AIMQ paper proceeds with an analysis of academic and practitioner perspectives on IQ based on the four dimensions derived from the authors' earlier research (Wang and Strong 1996; Wang 1995): Intrinsic, Contextual, Representational and Accessibility IQ.

Intrinsic IQ implies that information has quality in its own right. Contextual IQ highlights the requirement that IQ must be considered within the context of the task at hand; it must be relevant, timely, complete, and appropriate in terms of amount, so as to add value. Representational and accessibility IQ emphasize the importance of computer systems that store and provide access to information; that is, the system must present information in such a way that it is interpretable, easy to understand, easy to manipulate, and is represented concisely and consistently; also, the system must be accessible but secure. (Lee et al. 2002, p135)

These dimensions are not grounded in any theory, but are derived empirically using market research methods. They argue that these dimensions – and associated criteria – are sufficient to capture the multi-dimensional nature of IQ. To support this, they cite content analyses from a number of case study projects where all issues raised by practitioners can be mapped onto these criteria.

Rather than grouping these criteria by the four dimensions above, they adopt the PSP/IQ (Product–Service–Performance/Information Quality) two-by-two matrix developed earlier (Kahn et al. 2002). Here, the columns represent two different perspectives of quality (conformance to specifications and meeting/exceeding customer expectations), while the rows represent two view of information (information-as-a-product and information-as-a-service).

	Conforms to Specifications	Meets or Exceeds Consumer Expectations
Product Quality	Sound Information IQ Dimensions: • Free-of-Error • Concise Representation • Completeness • Consistent Representation	Useful Information IQ Dimensions: • Appropriate Amount • Relevancy • Understandability • Interpretability • Objectivity
Service Quality	Dependable Information IQ Dimensions: • Timeliness • Security	Usable Information IQ Dimensions: • Believability • Accessibility • Ease of Operation • Reputation

FIGURE 4 - PSP/IQ MATRIX (KAHN ET AL. 2002)

The authors argue that while their four IQ dimensions offer complete coverage, this matrix is more useful for helping managers prioritise IQ problems. They go on to develop a survey instrument which assesses the quality of information by asking information consumers to rate each of these 15 dimensions on an eleven-point Likert scale. An average score for each quadrant is computed, and an overall IQ score is the simple average of the four quadrants.

These scores are used in two ways: firstly, they allow benchmarking against a best-practice referent (such as an industry leader). Here, the organisation can assess in which areas they are meeting best practices and in which there are "gaps", drilling down through quadrants to dimensions to survey items. Secondly, the survey instrument also records whether a respondent is an information consumer or IS professional. This allows analysis of another kind of "gap" this time based on the roles.

Organisations can target quadrants and dimensions where they are experiencing a best-practices gap. They can also determine whether this might be due to a role gap, where those using information and those responsible for managing it disagree about its quality. The authors conclude that the AIMQ method is useful for identifying IQ problems and areas for improvement, and tracking any improvements over time.

While this framework has a method for IQ assessment and prioritisation of improvements, it lacks a solid theoretical underpinning. The original research identified 16 constructs (Wang and Strong 1996), but as "value-added" was problematic it has been dropped without explanation. The remaining 15 constructs are not defined; instead the authors rely on diverse information consumers and IS professionals to interpret "near-synonyms". For example, to determine the accessibility dimension - part of the Accessibility IQ dimension in the original study and part of the Usability quadrant in the PSP/IQ model - respondents are asked to rate the following statements:

- The information is easily retrievable.
- The information is easily accessible.
- The information is easily obtainable.
- The information is quickly accessible when needed.

For this dimension, the authors report a Cronback's Alpha (construct reliability) of 0.92 – a very high score indicating that these items are indeed measuring a single latent variable. However, the authors offer no advice to respondents about the differences between the retrieval, access and obtainment of information. Additionally, further items assess currency and timeliness of information without regard to the "promptness of access" (in the fourth item above).

Other examples of the use of "near-synonyms" in items to assess dimensions include: *believable*, *credible* and *trustworthy*; *correct*, *accurate* and *reliable*; *useful*, *relevant*, *appropriate* and *applicable*; *current*, *timely* and *up-to-date*; and *understand* and *comprehend*. Relying on respondents to bring their own differentiation criteria to bear on these overlapping terms weakens their conclusions.

Further, the dimensions themselves suffer from "near-synonyms": it is not obvious how *interpretability* and *understandability* differ, nor *reputation* and *believability*. As a consequence, it is not surprising that scores on these dimensions have a very high cross-correlation of 0.87 and 0.86 respectively (Lee et al. 2002). Respondents are unlikely to give very different ratings to the statements "It is easy to interpret what this information means" (Interpretability) and "The meaning of this information is easy to understand" (Understandability).

Using overlapping dimensions, "near-synonymous" terms and relying on the individual to assign meaning is a result of using an atheoretic approach to understanding Information Quality. By this, I mean that the authors do not present a theory of the nature of information or how it is created, assessed and used. Rolling these 15 dimensions up into four quadrants (derived by a theory) is an improvement. However, the subsequent survey design relies on the initial conception of IQ and hence carries forwards its limitations.

3.3.2 Ontological Framework

An example of a theoretically-derived framework for information quality is the ontological model proposed by Wand and Wang (Wand and Wang 1996). Ontology is the branch of philosophy that deals with the structure and organisation of the world in the broadest sense. In this context, it is the body of knowledge concerned with constructing models of (parts of) the world.

Wand and Wang start with a very clear set of statements defining the real world, the subset of interest (the domain) and the information system in terms of states. Based on Wand's earlier work on ontological modelling (Wand and Weber 1990), they build up a set of postulates relating the state of the information system with the state of the real world.

Specifically, they conceive of the world as being made up of *things* with *properties*. The real world is a system, decomposable into sub-systems. Each sub-system may be described in terms of a set of *states* and *laws* governing how it may progress from state to state. A system exists in one state at a moment in time. An information system is clearly a type of system too, and also has a set of *states* and *laws*. The *representation* process is the creation of a view of the real world within the information system. The *interpretation* process is the inference of the real world by a user (human or machine) perceiving the representation. In this way, the states of real world and the information system should be "aligned". By analysing the relationship between these states, Wand and Wang offer a thorough analysis of *data deficiencies:* "an inconformity between the view of the real world system that can be inferred from a representing information system and the view that can be obtained by directly observing the real world system" (Wand and Wang 1996, p89).

They identify three deficiencies that occur at the time of information system design: incomplete representation, ambiguous representation and meaningless states. *Incomplete representation* is when states exist in the real world that cannot be represented in the information system. *Meaningless*

states are those in the information system that do not correspond to a real world state. *Ambiguous representation* is when an information system state corresponds to more than one real world state, making it impossible to correctly infer the state of the real world.

Note that these deficiencies refer to *sets* of states (statespaces) and *possible* mappings between them, rather than a particular system at a point in time. For example, with an incomplete representation, if the real world is not in that "missing" state the information system can provide a correct representation. Similarly, correct inference is possible for an IS with meaningless states (or ambiguous representation), as long the information system (real world) is not in the problem state. However, the *possibility* of a mis-mapping constitutes a design deficiency.

The fourth type of data deficiency Wand and Wang identify occurs at the time of operation: garbling. Here, a well-designed information system (ie complete, unambiguous and meaningful) may be in the "wrong" relative to the real world. That is, the information system's state (at a particular moment) does not correspond to the real world state. This may be due to erroneous data entry or failure to reflect changes in the real world. They label such situations as incorrect.

Based on this analysis of the deficiencies in mapping between the (perceived) real world state and the information system state, they describe four dimensions of data quality. These are: complete, unambiguous, meaningful and correct. They go on to show how a number of other frequently-cited attributes of data (or information) quality fall into these four dimensions. For example, "lack of precision" can be understood as an ambiguity problem. This can be seen when we consider a customer birth date: if the IS captures the year and month, but not the day then one IS state corresponds to (up to) 31 real world states: we cannot distinguish between them and so that mapping is deemed *ambiguous*. As alluded to above, currency (or timeliness) is understood as when the real world changes state but the IS fails to "keep up". This results in the operational deficiency of garbling (to an incorrect state).

So we can see that this ontological model – by virtue of its grounding in a well-constructed theory – provides assurance that it is reasonably exhaustive in its coverage of data deficiencies due to system design or operation. However, its drawbacks are two-fold: first, its narrow scope. By restricting it to what the authors term the "internal view" (that is, "use-independent" intrinsic properties of data) the model does not address oft-cited information quality concepts such as relevance, importance, usefulness or value. Secondly, while laying out a conceptual model, there is no guidance for how to formally analyse, assess, measure or value a specific implementation (planned or realised). These drawbacks are explicitly acknowledged by the authors, who call for further work to extend their model.

3.3.3 Semiotic Framework

Next, I present an example of a framework that builds on the ontological model presented above to tackle the usage and assessment aspects. This framework also employs another theory, this time of semiotics, so is known as the Semiotic Framework for Information Quality (Price and Shanks 2005a).

The analysis begins with the insight that the philosophical area of semiotics (the study of systems of signs and symbols, in a broad sense) provides a coherent lens through which information quality can be studied. While the philosophical aspects of language and meaning enjoy a long history, *semiotics* (or semiology) emerged as a distinct discipline around the start of the 20th Century through the work of early researchers like Swiss linguist Ferdinand de Saussure (1857-1913), the American philosopher Charles Sanders Peirce (1839-1914) and later Charles William Morris (1901-1979) (Chandler 2007). While their work influenced linguistics, philosophy and language-based studies, semiotics has also

found use within IS for systems analysis (Stamper et al. 2000), data model quality (Krogstie et al. 2006; Lindland et al. 1994) and later data model and content quality (Moody et al. 1998).

The key to understanding this framework is the equivalence of the semiotic notion of a sign and the IS conception of a datum. A sign is a "physical manifestation ... with implied propositional content ... that has an effect on some agent" (Price and Shanks 2005a), where an effect is either a change in understanding or action. The referent is the implied propositional content, or "intended meaning" of the sign while the process of effecting change on some agent (semiosis) is the interpretation or received meaning of the sign. Hence, a datum in a data store constitutes a sign and a semiotic analysis of the data store as a sign-system allows a rigorous theoretical description of the quality of information.

Specifically, Price and Shanks identify three levels that build on each other. The first is the *syntactic* level, which deals with relations between sign representations (ie data and meta-data). The second is the *semantic* level, concerned with relations between sign representation and its referent (ie data and external phenomena). Lastly, the third is the *pragmatic* level, addressing the relations between sign representation and its interpretation (ie data and task/context). So, loosely speaking, these three levels (and their corresponding quality criteria) describe data *form*, *meaning* and *usage*:

	Syntactic	Semantic	Pragmatic
Quality Question Addressed	Is IS data good relative to IS design (as represented by metadata)?	Is IS data good relative to represented external phenomena?	Is IS data good relative to actual data use, as perceived by users?
Ideal Quality Goal	Complete conformance of data to specified set of integrity rules	1:1 mapping between data and corresponding external phenomena	Data judged suitable and worthwhile for given data use by information consumers
Operational Quality Goal	User-specified acceptable % conformance of data to specified set of integrity rules	User-specified acceptable % agreement between data and corresponding external phenomena	User-specified acceptable level of gap between expected and perceived data quality for a given data use
Quality Evaluation Technique	Integrity checking, possibly involving sampling for large data sets	Sampling using selective matching of data to actual external phenomena or trusted surrogate	Survey instrument based on service quality theory (i.e. compare expected and perceived quality levels)
Degree of Objectivity	Completely objective, independent of user or use	Objective except for user determination of relevancy and correspondence	Completely subjective, dependent on user and use
Quality Criteria Derivation Approach	Theoretical, based on integrity conformance	Theoretical, based on a modification of Wand and Wang's (1996) ontological approach	Empirical, based on initial analysis of literature to be refined and validated by empirical research

TABLE 4 QUALITY CATEGORY INFORMATION (ADAPTED FROM PRICE AND SHANKS 2005A)

Syntactic quality – concerned with the relations between signs – is understood as how well operational data conform to IS design (embodied as meta-data). Integrity theory provides a ready-

made theory for determining this conformance to eg cardinality constraints and rules of well-formed data structures.

The semantic level naturally builds on the model presented by Wand and Wang, as it is to do with how the information system represents the real world; that is the mapping between states of the external world and the data that are intended to represent this world. However, Price and Shanks modify the Wand and Wang model in three significant ways: firstly, they introduce an additional criterion of "non-redundancy" in the mapping. They argue that, like meaningless states, the presence of redundant states (ie multiple states in the IS refer to the same state in external world) in the IS constitute a design deficiency because they introduce a "danger" of deficiency in operation. The result is that the both the representation and interpretation processes now require a bijective function (one-to-one and "onto"): all states the external world must map onto a unique state in the IS, and vice versa.

A subsequent refinement of the framework based on focus group feedback (Price and Shanks 2005b) recasts "non-redundancy" as "mapped consistency" ie multiple IS states are permitted as long as they agree with each other (or are reconcilable within an acceptable time). This allows for system designers to employ caching, versioning, archiving and other forms of desirable redundancy.

Price and Shanks also argue that *incompleteness* can arise at design time (one or more external state cannot be represented in the IS) or during operation (for example, a clerk fails to enter data into a field). Thirdly, Price and Shanks address the decomposition deficiencies outlined by Wand and Wang by introducing separate notions of *phenomena-correctness* (correct mapping to an entity) and *property-correctness* (correct mapping to an attribute value of an entity). In terms of conventional databases, this distinction corresponds to row and column correctness respectively.

At the pragmatic level, the Semiotic Framework abandons theoretical derivation and employs an empirical approach akin to the AIMQ Framework, based on literature analysis, to describe a list of pragmatic quality criteria. At this level, the *reliability* construct subsumes the semantic level criteria of mapped (phenomena/property) correctly, meaningfully, unambiguously, completely and consistently. The additional (revised) pragmatic criteria are: Perceptions of Syntactic and Semantic Criteria, Accessible, Suitably Presented, Flexibly Presented, Timely, Understandable, Secure, Type-Sufficient and Access to Meta-data. The last two are included based on focus group refinement (Price and Shanks 2005a; Price and Shanks 2005b): the former replaces "value" in requiring all types of data important for use, while the latter refers to the ability of users to assess the lineage, granularity, version and origins of data.

The authors suggest that the SERVQUAL theory (Parasuraman et al. 1985) provides a means for assessing the quality at the pragmatic level. Similar to the AIMQ Framework, a "gap" is identified through a survey instrument employing Likert scales – this time between a consumer's expectations and her perceptions.

The strengths of this framework include the use of semiotic theory to stratify information quality criteria into levels (form, meaning and use) and the successful integration of Wand and Wang's ontological model at the semantic level. The main weakness is the lack of theoretical basis for assessing quality at the pragmatic level, which introduces similar problems as found in the AIMQ Framework. These include inter-dependencies (as acknowledged by the authors for eg Understandability and Access to Meta-data), problems with "near-synonyms" (eg using the undefined terms "suitable" and "acceptable" to describe aspects of quality and "worthwhile" and "important" to describe aspects of value) and finally "value" in general (value-added, valuable). As with the AIMQ Framework, the concept was originally included but ultimately dropped owing to its

conceptual poor-fit and heavy inter-dependence: feedback showed that "*valuable* was too general and abstract to ensure consistent interpretation … and therefore not useful as a specific quality criteria" (Price and Shanks 2005b).

3.4 IQ MEASUREMENT

This section summarises existing research in Information Quality measurement. While dozens of papers propose and analyse aspects of IQ, surprisingly little has been written about the specific measurement and definitions of metrics for IQ-related constructs. Some examples include a methodology for developing IQ metrics known as InfoQual has been proposed (Dvir and Evans 1996), while the Data Quality Engineering Framework has a similar objective (Willshire and Meyen 1997). Measurement of particular aspects of IQ have been tackled, such as soundness and completeness (Motro and Rakov 1996) and accuracy and timeliness (Ballou and Pazer 1995), completeness and consistency (Ballou and Pazer 2003). In many cases, such measurements are combined through transformations using weighting, sums and differences to derive metrics that allow comparison of quality levels over time (Evans 2006; Parssian et al. 1999; Parssian et al. 2004).

There are, broadly speaking, three approaches to IQ measurement, based on the kinds of scores employed: percentages (ratio, 0-100%), Likert scales (ordinal, eg low, medium and high) and valuation (ordinal, eg Net Present Value). The third is addressed in the following sub-section, while the first two are discussed here.

The purpose of IQ measurement is largely managerial (Heinrich et al. 2007): selection and monitoring of existing information sources for tasks and the construction of new information sources (possibly out of existing ones). This may involve benchmarking within and between organisations (Cai and Shankaranarayanan 2007; Stvilia 2008), as well as before and after IQ improvement projects. Naumman and Rolker (2000) identify three sources (or perspectives) of IQ measurement in their comprehensive review of IQ measurement based on user/query/source model:

Perspective	User	Query	Source
Type of Score	Subject-criteria scores	Process-criteria scores	Object-criteria scores
Example Criteria	Understandability	Response time	Completeness
Assessment Method	User experience, sampling	Parsing	Parsing, contract, expert, sampling
Scale/Units	Likert	Percentage	Percentage
Characteristics	Varies between users and tasks	Transient, depend on each usage instance	Change over time but constant for each usage instance and user

TABLE 5 ADAPTED FROM NAUMANN AND ROLKER (2000)

The pattern of using a combination of percentages for objective measures of IQ and a Likert scale for subjective measures is repeated throughout IQ research. For example, the Semiotic Framework (Price and Shanks 2005a) employs objective (percentage) measures in their syntactic and semantic levels and subjective (Likert) measures in their pragmatic level. This is spelled out in Pipino (Pipino et al. 2002), who argue for the combination of objective and subjective measures on the grounds that "subjective data quality assessments reflect the needs and experiences of stakeholders", while "objective assessments can be task-independent … [which] reflect states of the data without contextual knowledge of the application, and can be applied to any data set, regardless of the task at hand."

Across a number of models and frameworks, there is widespread agreement that percentages are the obvious and natural way to measure at least some IQ aspects such as completeness (Ballou and Pazer 2003), currency (Cappiello et al. 2003) and correctness (Paradice and Fuerst 1991). The assumption is that an information source with a score of 75% is of better quality than one with a score of 70%. However, there are considerable difficulties in determining such a figure, which undermines the claim of objectivity.

For example, consider a customer database comprising many thousands of records, each with a several dozen attributes. In this case, 75% completeness could mean that 25% of the customer records are missing. Or that 25% of the attributes (columns) have blank values. Or – as an example of a design problem - that 25% of the allowed values for a certain attribute (say, Customer Title) may be missing (eg *Parson*, *Earl* and *Inspector*). More subtly, 75% completeness may mean any combination of these issues is extant. While some researchers distinguish between these issues (eg Semiotic Framework, Ontological Model), most do not. The fundamental problem is that an enormous number of issues could combine to yield a particular quality score; yet it's unlikely that all of these situations would be regarded as equivalent by any given user in a particular context.

By way of illustration, consider the use of *data quality tagging* (Chengalur-Smith and Ballou 1999; Fisher et al. 2003; Price and Shanks 2008), an application of IQ measurement of research interest and with practical import. Briefly, it is the process of presenting extra information about the quality of a dataset to data consumers, typically expressed as a ratio (or percentage) score. The idea is that information consumers will change their use of the data (ie decision outcome, time, confidence etc) based on the level of quality conveyed by the score. The experimental results indicate that information consumers will – under some circumstances – incorporate this extra information into their decision-making (to some extent).

However, the consumers' *interpretation* of a quality score of, say, 70% is not obvious: is this the probability that the data is correct? Or some sort of distance metric, confidence interval, measure of spread (like variance) or significance value? In the absence of any instructions (other than the unhelpful remark that 1 is perfect quality and 0 is no quality) consumers will make up their own mind based on their experience, education and expectations, given the task and its context. Data quality tagging provides one motivation for objective IQ measures, and also highlights the drawbacks of their use.

This motivates the use of specific task- or usage-oriented measures – addressing the contextual dimension in the TDQM (Pipino et al. 2002) and AIMQ frameworks and the pragmatic layer in the Semiotic Framework. These frameworks argue for the necessity of adopting subjective measures to address this. For example, Lee et al. (2000) state that considering the information consumer viewpoint "necessarily requires the inclusion of some subjective dimensions", while for Price and Shanks (2005), the use of data by consumers "is completely subjective", since the pragmatic quality criteria

are evaluated with respect to a specific activity and its context. That implies that the assessment of such criteria will be based on information consumer perceptions and judgements, since only they can assess the quality of the data relative to use. (Price and Shanks 2005a, p93)

This proposal – that information usage by consumers in context can only be assessed by subjective measures – seems appealing, at least in the general case. After all, who would suppose that an objective measure of the quality of information for any arbitrary task could exist, given the problems with the objective measures in the comparatively simpler case of assessing semantic (or inherent) data quality?

However, this does not imply that surveying users with a Likert scale (or letter grade or nominal percentage) is the only possible approach. There is an important category of subjective assessment of IQ that employs an entirely different elicitation approach based around *user preferences*. This approach, while still subjective, allows for considerable sophistication in derivation and analysis. The next sub-section addresses this approach - the valuation of Information Quality.

3.4.1 IQ VALUATION

Information Quality valuation can be understood as a special case of IQ assessment, whereby the goal is to place a value on the quality level associated with an information source. In other words, the assessment methods are financial and the units of measurement are money (dollars or other currency). Some authors advocate a resource or asset view of an organisation's information resource (Levitin and Redman 1998; Moody and Walsh 2002; Solomon 2005). Frequently, IQ frameworks address value by considering cost as a factor or quality item. This is understood in at least two different ways: the *cost of reaching a level of quality* (through checking and correction procedures), where it is considered a factor to trade-off against other factors (Ballou et al. 1998); and the *cost of reaching a level of non-quality* through errors, mistakes and "information scrap and re-work" (English 1999). The former detract from value, while avoiding the latter contributes to value.

Other frameworks directly explicate the cost/benefit trade-off (Ballou and Tayi 1989; Eppler and Helfert 2004; Mandke and Nayar 2002; Paradice and Fuerst 1991), while others have applied decision-theoretic approaches – employing probabilities and pay-offs – to understand the impact of poor quality data (Kaomea 1994; Michnik and Lo 2009). One sophisticated analysis uses the financial engineering concept of "real options" to price data quality (Brobrowski and Soler 2004). Some examine trade-offs in particular applications contexts, such as data warehousing (Rao and Osei-Bryson 2008).

The importance of valuing IQ has been recognised by both academics and practitioners (Henderson and Murray 2005; Jacaruso 2006). For example, at the Data Quality workshop hosted by the National Institutes for Statistical Sciences in 2001, one of the key recommendations was that "Metrics for data quality are necessary that ... represent the impact of data quality, in either economic or other terms" (Karr et al. 2001). Earlier, practitioner/researcher Thomas Redman made an estimate for the typical organisation of about 25% of revenue (Redman 1998). Estimating the costs involved is a difficult accounting challenge owing to the diffused and intangible nature of poor information quality. In particular, opportunity costs (eg earnings forgone due to poor decision) are notoriously difficult to capture.

Even and Shankaranarayanan are amongst the few researchers to have tackled explicitly the notion of value-driven information quality (Even and Shankaranarayanan 2007a; Even and Shankaranarayanan 2007b; Even et al. 2007), using models that subjectively weight the benefit associated with data values across a number of familiar IQ dimensions, before aggregating up to get a total utility estimate for data assets.

The concept of value – incorporating costs and benefits in the broadest sense – faces two significant problems within Information Quality research. The first is conceptual: most researchers recognise its importance, but are unsure or inconsistent in its handling. The second is practical and concerned with the process of making a reasonable valuation subject to resource constraints. Despite these problems, valuation remains an important (albeit under-explored) area within IQ.

Examples of the conceptual problems with the concept of value were introduced earlier in the context of both the AIMQ and Semiotic frameworks. To recap, the "value-added" attribute of the contextual dimension was originally a part of the TDQM model (Strong et al. 1997) but was then

dropped for the PSP/IQ model without explanation (Kahn et al. 2002). As a consequence, value was not added to the AIMQ Framework (Lee et al. 2002).

With the Semiotic Framework, value was originally included (Price and Shanks 2005a), but as a context-specific "placeholder" item at the pragmatic level. Feedback from a focus group of practitioners identified its inclusion as a weakness, and the item was removed altogether. Further, "near-synonyms" and tautologies around value are used throughout the paper, adding to the lack of clarity. For example, value, value-added and valuable are, at different points, equated with or defined as worth, importance, usefulness and sufficiency (Price and Shanks 2005a; Price and Shanks 2005b).

The second difficulty with valuation of information quality concerns the tractability of valuation processes. One example such example is presented by Ballou and Tayi (1989) who prescribed a method for periodic allocation of resources to a class of IQ proposals (maintenance of data assets). It assumes a budgetary approach (that is, a fixed budget for IQ to be shared among a set of proposals), rather than an investment approach (evaluation of proposals based upon expected value returned). It further assumes that the data managers have sought and won the largest budget they can justify to their organisation. Based upon statistical sampling, a parameter estimation heuristic and an iterative integer programming model, the method arrives at an optimal dispersal of resources across proposals.

The method requires data analysts to understand the appropriate level of data granularity (fields, attributes, records) for the analysis and the expected costs of errors in these data sets. In general, the problem of estimating the costs of IQ defects is extremely complex. Earlier work (Ballou and Pazer 1985) employs differential calculus to estimate transformation functions that describe the impact of IQ defects on "down-stream" decision-making. This functional approach was later combined with a Data Flow Diagram method (Ballou et al. 1998).

Gathering information on the parameters required for these methods is likely to be very costly and fraught with technical and organisation difficulties. Further, there is little empirical evidence to support the feasibility of industry analysts undertaking the sophisticated mathematical analyses (ie the differential calculus and integer linear programming) as described.

Regardless of the valuation perspective or process, it can only be undertaken within a specific organisational process: the same information source or dataset will introduce (or remove) different costs depending on the purpose for which it is being used.

3.5 Customer Relationship Management

During the "dot-com boom" era, there was considerable academic interest in Customer Relationship Management (CRM) strategies, applications and processes, with some 600 papers published by the "bust" (Romano and Fjermestad 2001). CRM is the natural context to examine customer information quality, as it provides an academic framework and business rationale for the collection and use of information about customers. While quality data (or information) about customers is identified as key to the success of CRM initiatives (Messner 2004; Missi et al. 2005) it is not clear exactly how one should value this. Indeed, even the real costs of poor customer data are difficult to gauge due to the complexities of tracing causes through to effects. This is part of the much larger data quality problem. At the large scale, The DataWarehousing Institute estimated that – broadly defined - poor data quality costs the US economy over $US600 billion per annum (Eckerson 2001).

3.5.1 CRM Business Context

Customer Relationship Management can be understood as a sub-field of the Information Systems discipline (Romano and Fjermestad 2001; Romano and Fjermestad 2003), to the extent that it is a business strategy that relies on technology. Alter suggests that we can conceive of such systems as *work systems* (Alter 2004; Alter and Browne 2005). As such, the relationship between CRM and IQ is bi-directional: CRM systems *require* high quality customer information to succeed; and improving the quality of customer information can be a beneficial *outcome* of deploying CRM (Freeman et al. 2007; Jayaganesh et al. 2006).

One example of the latter is the study by Freeman and Seddon on CRM benefits (2005)(Freeman and Seddon 2005). They analysed a large volume of qualitative data about reported CRM benefits to test the validity of an earlier ERP (Enterprise Resource Planning) benefits framework. Some of the most significant benefits to emerge from this study related to quality of information: *improved customer-facing processes* and *improved management decisions*. Indeed the key "enabler" of these benefits was identified as "the ability to access and capture customer information".

Other studies highlight the importance to high quality customer information for CRM success. For example, industry analysts Gartner reported that "[CRM] programs fail, in large part, because the poor quality of underlying data is not recognized or addressed." (Gartner 2004, p1). Gartner stresses the link between poor quality customer information and CRM failure in their report "CRM Data Strategies: The Critical Role of Quality Customer Information" (Gartner 2003).

In light of the importance of quality information on CRM success, practitioners and researchers involved in CRM are frequently concerned with information quality. Similarly, CRM processes represent a significant source of value for practitioners and researchers dealing with information quality. That is, customer processes (undertaken within a CRM program) afford information managers with an opportunity to examine how high quality information can impact upon value-creation within the firm.

Certainly, we should not regard CRM processes as the only means by which quality customer information is translated into value: regulatory functions, strategic partnerships, market and competitor analyses and direct sale (or rent) of information assets can also contribute through cost reduction and revenue increases. Further, obtaining and maintaining high quality customer information is not a guarantee of a successful CRM strategy. However, the relationship between the two is sufficiently strong as to warrant a closer look at how information is used within CRM processes.

3.5.2 CRM Processes

Meltzer defines, a CRM process is seen an organisational process for managing customers (Meltzer 2002). He identifies six basic functions:

Cross-sell: offering a customer additional products/services

Up-sell: offering a customer higher-value products/services.

Retain: keeping desirable customers (and divesting undesirable ones).

Acquire: attracting (only) desirable customers

Re-activate: acquiring lapsed but desirable customers.

Experience: managing the customer experience at all contact points

At the core of these processes is the idea of customer classification: a large set of customers is partitioned into a small number of target sets. Each customer in a target set is treated the same by the organisation, though each may respond differently to such treatments. This approach seeks to balance the competing goals of effectiveness (through personalised interaction with the customer) and efficiency (through standardisation and economies of scale).

For example, a direct mail process might require partitioning a customer list into those who are to receive the offer, and those excluded. In this case, there are four possible outcomes from the treatment dimension "Offer/Not Offer" and the response dimension "Accept/Not Accept". The objective of the process is to correctly assign all customers to their correct treatment (ie accepting customers to "offer", not accepting customers to "Not Offer").

Clearly, organisations require hiqh-quality customer information in order to be able to execute these processes. Further, the need to correctly place a particular customer into the right group constitutes the (partial) customer information requirements of the organisation: the types of information collected, the levels of granularity, timing and availability and other characteristics depend, in part, on the usage. Obversely, the design of the customer processes themselves will depend on what information is (in principle) available, suggesting an interplay between information managers and process designers.

Hence, at its core, this *segmentation task* is a key point at which high-quality customer information translates into value for the organisation. As discussed above, this is not to say that CRM processes constitute the entirety of the value-adding effects of customer information; rather, that a sizeable proportion of the value amenable to analysis may be readily found therein. This is due the existence of a widely employed valuation method underlying many CRM strategies: the idea of the Customer Lifetime Value.

3.5.3 Customer Value

Customer Value is sometimes called Lifetime Value (LTV) or Customer Lifetime Value (CLV) or Future Customer Value. It is widely used as the basis for evaluating CRM and Database Marketing initiatives (Hughes 2006). There is a related notion of Customer Equity, which could be considered the sum of Customer Value over all customers. The idea is that the worth of a customer relationship to an organisation can be evaluated by adding up the revenues and costs associated with servicing that customer over the lifetime of the relationship, taking into account future behaviours (such as churn) and the time value of money (Berger and Nasr 1998). As such, it represents the Net Present Value of the customer relationship; that is, "the sum of the discounted cash surpluses generated by present and future customers (within a certain planning period) for the duration of the time they remain loyal to a company" (Bayón et al. 2002, p18).

Customer Value is used to evaluate the impact of CRM processes on an organisation's bottom line and takes the role of the "target variable" for controlling the design and operation of these processes. For example, a cross-sell process that focussed just on immediate sales over the duration of a particular campaign is not a suitable measure since it will fail to take into account follow-up purchases, referrals and "channel cannabilisation" (whereby sales from one channel, such as a website, may be transferred to another, say a call centre, without a net gain). Using Customer Value aligns operational marketing efforts with the longer-term interests of investors (and other stakeholders).

3.6 Decision Process Modelling

In this section, I introduce some important conceptual tools for understanding the role of information in representing the world (meaning) and making decisions (usage). Firstly, I look at some ideas from *information economics* – the study of the value of information. I then narrow that to examine *decision-theoretic* models that can be used to describe Customer Relationship Management processes, and the kinds of quantitative evaluation that they employ. Finally, I examine an engineering-oriented model, known as *information theory*, widely used to understand and measure information flows. Throughout, the relevance to customer information quality is highlighted through examples involving common CRM processes.

3.6.1 Information Economics

There are two basic concepts underpinning information economics: *uncertainty* and *utility*. Firstly, uncertainty refers to the absence of certain knowledge, or imperfections in what an observer knows about the world. It can be characterised by a branch of applied mathematics known as Probability Theory (Cover and Thomas 2005). While other approaches have been proposed (eg Possibility Theory), Probability Theory has by far the widest reach and most sophisticated analysis. The idea is that an observer can define a set of mutually-exclusive outcomes or observations, and assign a weight to each outcome. This weight reflects the chance or likelihood that the (as-yet-unknown) outcome will materialise. Originally developed to help gamblers calculate odds, it is now so embedded in all areas of science, statistics, philosophy, economics and engineering that it is difficult to conceive of the world without some reference to probabilities.

That said, there is some dispute and consternation about the interpretation of these weights. The so-called *frequentists* argue that the weights correspond to the long-run frequencies (or proportion of occurrences). So, to say "the probability of a fair coin-toss producing a heads is 50%" means that, after throwing the coin hundreds of times, 50% of the throws will result in a head. The objection, from rival *Bayesians*, is that this interpretation falls down for single events. For example, to state that "the probability of the satellite launch being successful is 50%" cannot be interpreted in terms of frequencies since it only happens once. These discussions aside, Probability Theory remains the single most comprehensive theory for understanding and reasoning about uncertainty.

The second key concept is utility. This refers to a measure of the happiness or net benefit received by someone for consuming or experiencing a good or service. Its role in economic theory is to capture (and abstract) the idea of "value" away from psychological or cognitive processes. We can thus reason about how a particular decision-maker's utility varies under different circumstances. As such, utility has underpinned economic theory for several hundred years (Lawrence 1999), allowing theorists to posit *homo economicus*, the so-called "rational man", to describe and predict the behaviours of large groups of people.

However, there have been many debates within the economics community about the nature of utility (eg whether or not it is subjective), how it is measured and so on. Despite these latent problems, a sophisticated edifice was constructed throughout the 19th century in a theoretical body known as "neoclassical microeconomics". This explains and predicts decision-making around economic production and consumption through concepts such as supply and demand curves, marginal (or incremental) utility, production possibility frontiers, returns to scale and so on.

Following important developments in Game Theory after World War II, two mathematicians, Morgestern and von Neumann, set about recasting neoclassical microeconomics in terms of this new mathematical model (Neumann and Morgenstern 2004). Their resulting work is often known as the "game-theoretic reformulation of neoclassical microeconomics", or more loosely, Utility Theory.

Morgestern and von Neumann's key insight was to link utility with preferences. In their model, actors have a *preference function* that ranks different outcomes, or possibilities, associated with "lotteries" (taking into account chance). They showed that, mathematically, micro-economics could be reconstructed "from the ground up" using this idea of ranked preferences. What's more, preferences can be observed indirectly through people's behaviour (ie their preferences are revealed through their choices), allowing experimental research into decision-making.

3.6.2 Information Theory

Throughout the 1960s and 1970s, the field of information economics integrated Game Theory with another post-war development: Information Theory. Working at the Bell Laboratories on communications engineering problems, the mathematician Claude Shannon published a modestly-entitled paper "A Theory of Mathematical Communication" for an in-house research journal (Shannon 1948). When the full import of his ideas was grasped, it was re-published (with Warren Weaver) as a book entitled *The Mathematical Theory of Communication* (Shannon and Weaver 1949).

The key quantity Shannon introduced was *entropy*, a measure of the uncertainty of a random variable. By measuring the changes in uncertainty, Shannon's theory allows analysts to quantify the amount of information (as a reduction in uncertainty) of an event.

Conceptually, Shannon's innovation was to explain how the communication process is, at its heart, the selection of one message from a set of pre-defined messages. When the sender and receiver select the same message (with arbitrarily small probability of error), we are said to have a *reliable* communication channel. This simple precept – combined with a rigorous and accessible measurement framework – has seen information theory (as it is now known) continue development through dozens of journals, hundreds of textbooks and thousands of articles. It is widely taught at universities in the mathematics and engineering disciplines.

From this grand theoretical foundation a large number of application areas have been developed: communications engineering, physics, molecular biology, cryptography, finance, psychology and linguistics (Cover and Thomas 2005). Economics – in particular, information economics – was very quick to adopt these new ideas and integrate them with Game Theory. The object of much of this work was to understand the interplay between value and information – how economics can help place a value on information and (in turn) how information can shed new light on existing economic theories (Heller et al. 1986). Notable economists tackling these ideas during this period included Henri Theil (Theil 1967), Jacob Marschak (Marschak 1968; Marschak 1971; Marschak 1974a; Marschak 1974b; Marschak et al. 1972; Marschak 1980) and Joseph Stiglitz (Stiglitz 2000) and George Stigler (Stigler 1961).

3.6.3 Machine Learning

One application area of particular interest to this research is *machine learning*. This branch of applied mathematics examines methods for sifting through large volumes of data, looking for underlying patterns. (For this reason, there is a large overlap with the data mining discipline.) Specifically, the focus is on algorithms for building computer models for classifying, segmenting or clustering instances into groups. For example, models can be used to estimate how likely it is a customer will default on a loan repayment, based on the repayment histories of similar customers. Other example applications include direct marketing (where the task is to identify customers likely to respond to an offer), fraud detection (flagging suspect transactions) and medical diagnostic tasks.

The primary interest of the machine learning research community is not the models themselves, but the algorithms used to build the models for each application area. When evaluating and comparing the performance of these algorithms, researchers and practitioners draw on a range of measures.

In terms of outcomes, the recommended measure is the *cost of misclassification* (Hand 1997). That is, when making a prediction or classification (or any decision), the cost arising from a mistake is the ideal success measure, in terms of getting to the best decision. Standard Economic Theory requires decision-makers to maximise their expected utility (Lawrence 1999), which in Information Economics is used to develop a sophisticated approach to valuing information based on so-called pay-off matrices (a table that captures the costs and benefits of different decision outcomes).

Evaluating performance in this way is highly context-specific: the costs of misclassification for one decision-maker might be very different than for another. In other cases, the costs might not be known *a priori* or even be entirely intangible. To deal with these scenarios, decision-theoretic measures of outcome performance are used. As these are independent from the consequences of decisions, they evaluate the ability of the models to guess correctly the preferred outcome only.

The most widely used measures for binary decisions like medical diagnoses are *sensitivity* and *specificity* (Hand 1997) and derived measures. Essentially, these measure the probabilities of "false positives" and "false negatives" and are known as *precision* and *recall* in the document retrieval literature. In the direct marketing literature, it's more common to describe the success of a campaign in terms of the ratio of true positives ("hits") to false positives ("misses"). This generalises to the ROC[2] curve, which plots out these two measures on a graph. The area under the curve (AUC) is frequently used to compare between models (Fawcett 2006; Provost et al. 1997). Other research by has extended the ROC concept to include costs, where they are available (Drummond and Holte 2006).

A marketing-specific version of this concept is found in "lift", or the proportional expected improvement in classifying prospects over a random model. This idea is further developed in the L-Quality metric proposed by (Piatetsky-Shapiro and Steingold 2000).

While these approaches are independent of costs and as such allow evaluation of the models in a general sense, they do not naturally extend to cases where there are more than two outcomes. For example, a CRM process that categorised each customer into one of four different groups, depending on their likely future spend, cannot be characterised neatly in terms of false negatives/positives. A further problem is that these approaches do not take into the prior probabilities. For instance, suppose a process correctly categorises customers' gender 97% of the time. That might sound high-performing, but not if it's actually being applied to a list of new mothers in a maternity hospital!

One approach to both of these situations is to use measures based on entropy, or the reduction of uncertainty, as first proposed by Shannon (Shannon and Weaver 1949). The machine learning community makes extensive use of a set of measures proposed by Kononenko and Bratko (1991). The "average information score" and "relative information score" measure how much uncertainty is reduced by a classifier, on average. Being theoretically-sound, it elegantly takes into account both non-binary outcomes and prior probabilities, allowing performance comparison between different decision tasks as well as different contexts.

CRM processes (which, at their core, involve segmenting customers into different groups for differentiated treatment) can be characterised as classifiers. From a classifier perspective there are three approaches to measuring their performance: cost-based (which is context-specific and to be preferred in real situations, if costs are available), decision-theoretic (useful for common cases

[2] *The term "ROC" originated in communications engineering, where it referred to "Receiver Operating Characteristic".*

involving binary decisions) and information-theoretic (useful for multiple outcome decisions with uneven prior probabilities).

Conceived as classifiers, the impact of information quality on the performance of these CRM processes can be understood in terms of decision-making: how do IQ deficiencies result in mis-classification of customers? The methods and measures used for quantifying CRM performance (including scoring and valuing) can be brought to bear to answer this question, indirectly, for customer IQ.

3.7 CONCLUSION

The information quality literature is replete with frameworks and definitions, few of which are theoretically-based. These conceptual difficulties mean that measurement of IQ deficiencies is weak, especially in the area of valuation. A large and growing body of knowledge relating to quantifying value and uncertainty is established in the fields of information economics, decision-making and information theory which have seen little application to IQ. Customer Relationship Management provides a customer-level focus for IQ and, through machine-learning models, provides a natural and obvious context for employing this established knowledge.

Chapter 4

Context Interviews

CONTEXT INTERVIEWS

4.1 SUMMARY

This chapter presents the rationale, process and key findings from field interviews. These semi-structured interviews were undertaken with Information Systems practitioners with a view to understanding current practices and their understanding of Information Quality measurement and valuation in large-scale customer-focused environments.

It was found that while IQ is regarded as important, there is no standard framework for measuring or valuing it. Further, the absence of such a framework hampers the ability for IS practitioners to argue the case for investing in improvements as access to organisational resources is dependent on such a case being made.

4.2 RATIONALE

Before developing a framework for customer IQ valuation, it is important to determine the existing "state of the art". This is to avoid wasted effort and to ensure that any contribution is cumulative, in the sense that it builds on existing knowledge established during the Literature Review. Further, for such a framework to be acceptable to practitioners, it is necessary to understand their expectations: which assumptions are valid, which elements are present, what prior skills or knowledge are required, who is intended to use it and how the results are to be communicated.

4.2.1 ALTERNATIVES

Two other data collection methods were considered before settling on the use of practitioner interviews. The first was an analysis of industry texts ("white papers"), while the second was a practitioner survey. The merits and drawbacks of these approaches - along with the rationale for their rejection – follow.

White papers are an important part of information systems vendor marketing. While they may include pricing and other sales-specific information, more generally, they seek to show that the vendor understands (or even anticipates) the needs of the prospective buyer. Taken at face value, these texts could provide an accurate and current picture of the availability of information quality products and services in the market. Further analysis could draw-out the requirements, norms and practices of IS buyers – at least as seen by the vendors. What makes this approach appealing is the ready availability of voluminous sources, published online by a variety of market participants ranging from hardware vendors to strategic consultancies to market analysis firms.

The principal drawback to using white papers to assess the current situation in industry is that, as marketing and pre-sales documents, they are unlikely to be sufficiently frank in their assessments. It is expected that problems that a particular vendor claims to fix will be overstated while unaddressed problems will be glossed over. Similarly, undue weight may be given to a particular vendor's strengths while their weaknesses are downplayed. Further, as sales documents, they are usually generic and lack the richness that comes with examining the specific contexts in which IQ is measured and valued. Contrived case studies – even those purporting to relate to creating and arguing a business case – may say more about how the vendor *hopes* organisations invest in IQ than the actual experiences of practitioners.

The second data collection method considered was a practitioner survey. By directly seeking the opinions of practitioners, the problem of cutting through marketing agendas is dealt with. A survey would potentially allow the opinions of a large number of practitioners to be gathered from across industry. Further, such a dataset would be amenable to a statistical analysis, allowing for rigorous hypothesis-testing, trending and latent-variable discovery.

Gathering data about respondent qualifications, experience and role would be straightforward and is common enough in this type of research. However, designing a meaningful set of questions about IQ measurement and investment would be fraught, even with piloting. This is because the terminology and even concepts are not standardised, while accounts of organisational structures and processes do not lend themselves to the simple measurement instruments used in surveys, such as Likert Scales. A compounding problem lies in the recruitment of respondents: very few organisations have a single "point-person" responsible for IQ and the low response rate and selection bias may undermine any claims to statistical significance.

4.2.2 SELECTION

Practitioner interviews offered a number of benefits over text analysis and surveys (Neuman 2000). The face-to-face communication means that terms relating to both IQ and organisational issues can be clarified very quickly. For example, position descriptions, internal funding processes and project roles are broadly similar across the industry but variations do exist. The flexible nature of interviews means that a subject – or interviewer – may guide the discussion based on the particular experiences or understandings of the subject. Generally, this tailoring cannot be planned in advance as it is only at the time of the interview that these differences come to light. Lastly, there is a richness of detail and frankness that comes only through people speaking relatively freely about specific experiences (Myers and Newman 2007).

At a practical level, practitioner interviews are cheap, quick and comparatively low risk. It is reasonable to expect that workers in the IS industry would have exposure to a range of technologies and work practices due to their high mobility and the industry's immaturity. Thus, interviewing a fairly small number of practitioners can glean insights across a large number of organisations.

Such an approach is not without its limitations. Primarily, there is a risk of a getting a biased or unrepresentative view of the wider industry through problems with subject recruitment (sampling). A secondary problem is with the testimony from the subjects: faulty recollection, self-censorship and irrelevant materials were identified as concerns.

4.3 SUBJECT RECRUITMENT

To conduct this phase of research, IS practitioners had to be contacted with an offer and agree to participate. This section outlines the goals and methods used. The terminology used here comes from the sampling chapter in Neuman's text on social research methods (Neuman 2000).

4.3.1 SAMPLING

The purpose of the field interviews was to understand the "state of the art" of IQ measurement and valuation by decision-makers within the IS industry, particularly focusing on those dealing with mass-market, multi-channel, retail customer management. The idea is to find a group that – collectively – spans a suitable cross-section of that industry sector. This is not achieved via statistical (random) sampling, but through a process called *stratification*. In short, a number of criteria are identified and at least one subject must meet each criterion. These criteria are outlined and discussed subsequently.

The question of sample size is understood in terms of *saturation*[3]: the point at which the incremental insight gained from further interviews becomes negligible. Of course, this raises the problem of defining negligibility in this context. The approach taken was to begin the analysis process in tandem with the data collection. This meant that the incoming "new" data from each interview could be compared with the entirety of the existing data, so that a view on the novelty of insights for each interview could be formed.

This sequential approach makes a very good procedural fit with the mechanism of recruitment: *snowballing*. This refers to asking subjects to suggest or nominate new subjects at the end of the interview. The reason for this is that each subject, through their professional network, may know dozens or scores of possible subjects. However, at the end of the interview, they have a much better idea about the research project and can quickly nominate other practitioners who are both likely to participate and have something to offer the study.

The key to making snowballing work is trust and rapport. After spending time in a face-to-face context with the interviewer, subjects may be more willing to trust the interviewer and so make a recommendation to their contacts to join the study. By the same token, an approach to a new subject with a recommendation from a trusted contact will be more likely to succeed than "cold-calling" from an unknown person.

The snowball was "seeded" (ie the initial recruitment) with two sources: subjects drawn from the professional network of the researcher and those from the industry partner. By a coincidence, these were centred on the one company, Telstra Corporation, the incumbent and (at the time) partially-privatised Australian telecommunications carrier. However, owing to its vast size and fragmented nature, there was only one subject in both "seed lists".

All stages of the field study, including subject approach, obtaining of permission and consent, question and prompt design and the collection, analysis and storage of data were governed by an appropriate university Human Research Ethics Committee. Given the intended subjects and the types of data collected, the project was rated as being low-risk.

The following strata (or dimensions and criteria) were identified for ensuring the sample is representative of the target decision-makers in the IS industry. (That is, those who operate within large-scale customer management environments involving significant amounts of complex customer data deployed across multiple channels.)

- **Industry Sector.** The two types targeted by this research are (retail) *telecommunications* and *financial services*. Since most households in the developed world have an ongoing commercial relationship with a phone company and a bank, organisations operating in these two sectors have very large customer bases. They also operate call centres, shop fronts, web presences in highly-competitive sectors and are sophisticated users of customer information.
- **Organisational Role.** There are three types of roles identified: *executive*, *managerial* and *analytical*. By ensuring that executives, managers and analysts are represented in the sample, the study will be able to draw conclusions about decision-making at all levels of the organisation.
- **Organisational Function.** The sample should include representatives from both *business* and *technology* functional groups. This includes marketing, finance or sales on the business side, and research, infrastructure and operations on the technology side. These groups may

[3] *"Saturation" is sometimes referred to as "adequacy" in the social sciences.*

have different terminology, priorities and understandings of IQ and to omit either would leave the sample deficient.

- **Engagement Mode.** This refers to the nature of the relationship with organisation: *full-time employee*, *contractor/consultant* and *vendor*. People working in these different ways may offer different perspectives (or levels of frankness) about the organisational processes or projects.

A sample composed of representatives across these four strata (meeting all ten criteria) would maximise the collection of disparate views. It is worth stressing that the sample is not intended to be *calibrated*, that is, with the respective proportions in the sample matching those in the wider population. Instead, it should achieve *sufficient coverage* of the population to allow inferences to be drawn about current practice.

Also, it is not necessary to find representatives in each of the possible combinations of strata (ie 2 x 3 x 2 x 3 = 36). For example, the absence of a financial services sector technology analyst from a vendor firm should not be taken as invalidating the sample. As long as each criterion is met, the sample will be considered to capture the viewpoints in the target population.

Finally, the interviews asked subjects to reflect on their experiences in their career, across many roles and employers. Given the high-mobility of the IS workforce, many of the more experienced subjects have worked for different employers in a range of sectors and with different roles. The explanations, insights and anecdotes gathered represent their views from across these disparate rolls and organisations.

4.3.2 DEMOGRAPHICS

The final sample consisted of fifteen subjects, interviewed for an average of 90 minutes each. In accordance with the ethical guidelines for this research project, the subjects' names are suppressed as are their current and past employers. Pseudonyms have been used for employers, except for this project's industry partner, Telstra Corp, where approval was obtained.

ID	Organisation	Sector	Role	Function	Mode	Experience in years	Qualifications (highest)
S1	ISP	Telecom	Exec	Business	FTE	30+	PhD
S2	Telstra	Telecom	Exec	Business	FTE	35+	BA
S3	Telstra	Telecom	Analyst	Business	FTE	5+	BE, BSc
S4	DW	Telecom	Analyst	Tech	Vendor	5+	MBA
S5	DW	Telecom	Mgmt	Tech	Vendor	25+	MBA, MEng
S6	Telstra	Telecom	Mgmt	Business	FTE	15+	MIS
S7	Telstra	Telecom	Analyst	Tech	FTE	15+	PhD
S8	Telstra	Telecom	Exec	Business	FTE	35+	Trade Cert.
S9	Telstra	Telecom	Exec	Business	FTE	15+	Grad.Cert.
S10	Telstra	Telecom	Mgmt	Business	Consult	20+	MBA
S11	Telstra	Telecom	Mgmt	Business	FTE	15+	BSc
S12	OzBank	Finance	Exec	Business	FTE	20+	High School
S13	OzBank	Finance	Analyst	Tech	Consult	20+	Dip. (Mktg)
S14	Telstra	Telecom	Mgmt	Tech	FTE	30+	Unknown
S15	Data	Finance	Exec	Tech	FTE	20+	Unknown

TABLE 6 SUBJECTS IN STUDY BY STRATA

Note that for the purposes of these classifications, a subject's role is not as self-reported owing to differences in terminology. Subjects were deemed "executive" if they had board-level visibility (or, in

one instance, had a large equity stake in the business), while "management" meant they were accountable for a number of staff or significant programs of work.

All 15 subjects reported that IQ was an important factor in their work and of interest to them, two had "Information Quality" (or similar) in their job title and a further two claimed significant expertise in the area. All subjects have experience with preparing, analysing or evaluating business-cases for IS projects.

The subjects are unusually-well educated compared with the general population, with most having university qualifications and half having postgraduate degrees. Their combined IS industry experience exceeds 300 years, with four reporting more than 30 years in their careers. Only four subjects have had just one employer while another four indicated significant work experiences outside of Australia. Eight subjects have staff reporting to them, and four have more than 30.

While there were still "leads" available to pursue, after the fifteenth interview each of the designated strata (sector, role, function and mode) was adequately represented. After some 25 hours of interview data, no new IQ measures or investment processes were identified. Novel anecdotes around poor IQ were still emerging; however, gathering these was not the intent of the study. As such, it was deemed that "saturation" had been reached and additional interview subjects were not required.

4.3.3 LIMITATIONS

Owing to practical constraints around time, travel and access, the final sample has some limitations. Here, the most significant are addressed.

- **Geography**. The sample consists of subjects from metropolitan Australian cities. There may be reason to think that viewpoints or practices vary from country to country, or even city to city, and that subjects should be selected from different locations. However, it is argued that the high-mobility of the IS workforce – along with the standardising effect of global employers, vendors and technologies – means that geographical differences among the target population are minimal.
- **Gender.** While the sample consists only of men, it is argued that the absence of female subjects in the study is a limitation rather than a serious flaw. The reason is that subjects were asked to provide responses to how they have seen IQ measurement and valuation employed in practice. Given that the sampling processes targeted managers and analysts in larger organisations, it is considered unlikely that women will have worked on significantly different IS projects than men. While there may be instances of women being "streamed" into certain projects or management processes (thus affording very different experiences), it is unlikely that such practices could remain widespread and persistent in the face of labour market changes and regulatory frameworks. If there are separate "male" and "female" ways of understanding IQ, it is not in the scope of this study to determine these.
- **Culture.** While their ethnic backgrounds varied, the subjects were all drawn from Australian workplaces. Similarly to geography and gender, sampling from a range of cultures was not a goal. It is argued that the global nature of the industry (reflected in the sample) tends towards a standardisation of norms and values to the dominant culture, in the case a Western business perspective. Again, this is not to rule out the possibility of cultural (or linguistic or ethnic) differences in understanding of IQ; rather, it is outside of the scope of this study to ascertain these differences.
- **Organisation.** All subjects were employed in the corporate sector at the time of the interview, with none from the government or small-business sector. The absence of public

or government sector experience is ameliorated somewhat by the fact that Telstra was, until the mid-90s, a government organisation. Over one third of the collective experience in the sample (i.e. more than 100 years) was from this time. The unique understanding of IQ by small business practitioners was not sought given the goal of large-scale customer-focused IS environments.

These limitations in the final study sample suggest areas of possible further research but do not substantially undermine the purposes of sampling for this study.

4.3.4 Summary of Recruitment

Subject sampling consisted of using "snowballing" from within the industry partner organisation to ensure representation across four identified strata. As the interview series progressed, demographic information and further "leads" were examined with a view to obtaining sufficient coverage within each stratum. Response data were examined after each interview to ascertain when "saturation" was achieved.

4.4 Data Collection Method

In this section, the method for collecting and organising the qualitative data from the interviews is discussed in some further detail. It covers the general approach, including a description of the process and settings, and the materials, covering the specific questions and prompts used in the interviews.

The objective was to gather evidence of existing practices and norms for justifying investments in improving the quality of customer information. Of particular interest was the collection or use of *measures* (of any kind) to support, evaluate or test initiatives that have an impact on customer IQ.

4.4.1 General Approach

Subjects were recruited and guided through a "semi-structured" interview at their workplaces using some pre-written questions and prompts (lists of words). All interviews were tape recorded (audio-only) and notes were taken by the investigator during and after the interview.

The interview was conducted in three phases. The first phase gathered demographic information and a brief history of the subject (including qualifications, work experience and current position). The second phase looked more closely at the subjects' awareness, use and selection of IQ-related measures. The final phase was more open-ended and provided an opportunity for subjects to share information and perspectives that they though relevant, in light of the preceding two phases.

Interview subjects were recruited using "snowballing" and approached via email. In some instances, telephone calls were used to organise locations and times for the interview. The subjects determined their own workplace approval (where needed). Ethics consent for the study was obtained in accordance with University processes.

All interviews took place at the subjects' workplaces, either in their offices or designated meeting rooms. The locations were the central business district (or adjacent business areas) of Melbourne, Australia. One interview took place in the central business district of Sydney, Australia.

The interviews took place between 10am and 7pm on working weekdays, with the majority occurring after lunch (between 1pm and 5pm). Subjects were advised that the interviews were open-ended, but that they should schedule at least one hour. Most gave considerable more than that, averaging one and half hours. The shortest was one hour and ten minutes, while the longest was just over two hours.

In all cases, there was one researcher and one subject present and the setting was "corporate": dress was formal i.e. suit and tie, the language was business English and the prevalent mood could be described as relaxed. The subjects' attitude to the research project was positive.

These observations are important because they indicate the subjects were willing participants, in their usual work environment, speaking to familiar topics in a tone and manner which was comfortable. By allowing them to control the time of the interview, the subjects were not rushed or interrupted. Comments made before and after the tape recording (that is, "off the record") were not markedly different in either tone or substance. This suggests that subjects were speaking reasonably freely and that the presence of recording equipment and note-taking did not make them more guarded in their remarks.

4.4.2 MATERIALS

During the interviews, two sheets of A4 paper were used as prompts. The first described a series of questions, including demographic and context questions to lead discussions. This information would help assess if the study met the sampling criteria; that is, obtaining reasonable coverage of industry practice. It would also allow analysis of interactions between an individual's education, role or experience and their views about IQ justification.

The second sheet comprised of a list of measures that may relate to investment in IQ. These measures were grouped into three lists: system, relationship and financial and were taken from academic literature in information quality, customer relationship management and IS investment, respectively. Further, these lists were not static but evolved over interviews.

At the commencement of each interview, the investigator pointed out these prompts and explained that the contents of the second one would be progressively revealed during that phase of the interview.

The context questions used are as follows, with further prompts in parentheses:

- *What is your professional background?* (Qualifications and work experience.)
- *What is your current role within your organisation?* (Title, position, current projects/responsibilities.)
- *What experiences have you had with Customer Information Quality?* (Projects, systems, methods, tools, roles.)
- *How would you describe your perspective or view on Customer Information Quality?* (Operational, analytical, managerial, strategic.)
- *How does your organisation generally justify investments in your area?* (Business case, investment committee, ad hoc.)

Generally, this last question (investments) prompted the lengthiest exposition, as it involved explaining a number of corporate processes and required frequent clarification by the subject of their terminology around roles and titles. This discussion frequently involved reference to measurements, which were drawn out more fully in subsequent stages. Further, this question also afforded an opportunity for subjects to describe some non-quantitative approaches to IQ investment justification.

The second last question (perspective) was the most confusing for subjects, frequently requiring prompting and clarification by the interviewer. In many cases, the second question about the current role informed responses to this question and it was largely redundant.

In the course of conducting these interviews, explanations and clarifications were streamlined and, given the bulk of subjects were drawn from one large company (albeit in quite different areas), the time taken to convey descriptions of corporate hierarchies and processes reduced.

Sources of potential confusion for the subjects were anticipated as the investigator gained experience with the question set. For example the third question (experiences) was historical in nature; this on occasion caused confusion about whether subsequent questions were asking about current or historical perspectives and processes. By explicitly stating that during the question, this saved the subject from either seeking clarification or answering an unintended question and being re-asked.

The second sheet comprised three columns: *system measures*, *relationship measures* and *financial measures*. The system measures related to "the technical quality of the repository" and initial instances were drawn from the IQ literature. The relationship measures described "outcomes of customer processes" and came from the Customer Relationship Management literature. The investment measures were selected from the IS investment literature and were characterised as measures describing "the performance of investments".

During this phase of the interview, subjects were told of these three broad groups (with explanation) and the table was covered with a blank sheet in such a way as to reveal only the three headings. As each group was discussed, the blank sheet was moved to reveal the entire list in question.

For each of the three groups, subjects were asked questions to ascertain their *awareness*, *use* and *selection* of metrics as part of IQ evaluation, assessment and justification. Firstly, for *awareness*, subjects were asked to nominate some measures they've heard of (unprompted recall). These were noted. Next, subjects were shown the list and asked to point out any that they *haven't* heard of (prompted recall). Finally, they were asked to nominate additional, related measures that they think should be on the list. In this way, their *awareness* of a variety of measures was established.

The next series of questions related to their *use* of measures in IQ evaluation, assessment and justification. Subjects were asked which of the measures they had a) heard of other people actually using and b) they had used directly themselves. Follow-up questions related to the nature of the usage; for example, whether it was retrospective or forward-looking, formal or informal, ongoing or *ad hoc* and the scope (in terms of systems and organisation).

Lastly, for the measures they had used, subjects were asked to explain why those particular measures were *selected*. Was it mandatory or discretionary? Who made the decision? What kinds of criteria were employed? What are the strengths and weaknesses of this approach? These questions helped establish an understanding of what drives the *selection* of measures for IQ evaluation, assessment and justification.

As subjects moved along the awareness, use and selection phases, the set of measures under discussion rapidly diminished. In most instances, it didn't progress to the selection phase since subjects had not used directly any measures. No subject expressed first-hand experience of selecting measures across the three domains of system, relationships and investment.

During the awareness phase, subjects were asked to nominate additional measures that they thought should be included. As a result, the list of measures grew over the course of the study. Further changes came from renaming some measures to reduce confusion, based on subject feedback and clarification. The three sets of measures at the start and the end of the study are reproduced here.

System Measures	Relationship Measures	Investment Measures
Validity	Response Rate	Payback Period
Currency	Churn	Internal Rate of Return
Completeness	Cross-Sell / Up-Sell	Share of Budget
Latency	Credit Risk	Economic Value Added
Accuracy	Lift / Gain	Net Present Value
Consistency	Customer Lifetime Value	
Availability		

TABLE 7 INITIAL MEASURE SETS

At the end of the study, 14 of the initial 18 measures were unchanged and a further six measures had been added. Note that "accuracy" was renamed "correctness", while "latency", "availability" and "churn" were augmented with near-synonyms. The other changes comprised of additions.

System Measures	Relationship Measures	Investment Measures
Validity	Response Rate	Payback Period
Currency	*Churn / Attrition / Defection*	Internal Rate of Return
Completeness	Cross-Sell / Up-Sell	Share of Budget
Latency / Response	Credit Risk	Economic Value Added
Correctness	Lift / Gain	Net Present Value
Consistency	Customer Lifetime Value	*Accounting Rate of Return*
Availability / Up-Time	*Share of Wallet*	*Profitability Index*
	Time / Cost to Serve	*Cost / Risk Displacement*
	Satisfaction / Perception	

TABLE 8 FINAL MEASURE SETS (NEW MEASURES IN ITALICS)

This list was stable for the last five interviews, providing a strong indication that saturation had been reached as far as the awareness, use and selection of measures were concerned.

The final phase of the interview involved asking more open questions of the subjects, in order to elicit their perspectives in a more free-flowing dialogue. It also provided a means to garner further references and participants. Many subjects availed themselves of this opportunity to relate anecdotes, grievances and "war-stories from the trenches". The specific questions were:

- *Do you have any other views you'd like to express about criteria, measures or models justifying IQ initiatives?*
- *Do you have any anecdotes or maxims you'd like to share?*
- *Do you have any references to authors or publications you think would be of benefit to this research?*
- *Can you recommend any colleagues you think may be interested in participating in this research?*
- *Would you like to receive a practitioner-oriented paper summarising this stage of the research?*

It was important to ask these questions after the main body of the interview, since subjects would have a much better opinion about the kinds of information (anecdotes, references, colleagues) the study was seeking. They would also be comfortable with the investigator and more likely to make a recommendation or endorsement to their colleagues.

Over the course of the study, the views and anecdotes continued at a constant pace, while the (new) nominated colleagues tapered off. In particular, subjects from the industry partner suggested the same people repeatedly, so once these subjects were either interviewed (or confirmed their unavailability), the pool of leads diminished.

This indicates that the study was speaking to the "right" people, in terms of seniority and organisational group, at least as far as the industry partner was concerned. It also provides more evidence that saturation had been achieved, in that few new candidates with specific knowledge were being nominated.

4.4.3 Summary of Data Collection

Data were collected by interviewing subjects at their workplaces and making audio recordings and taking field notes. The interviews were semi-structured and had three phases. First, subjects were asked about their experience and current roles. Second, their awareness, use and selection of measures pertaining to IQ were ascertained (across system, relationship and investment domains). Finally, more open-ended discussion and leads to further resources were sought.

As the interview study progressed, the sets of measures used in the second phase were updated to reflect either better definitions or additions to the list. Also, subjects were asked to nominate colleagues to participate in the study, as part of the "snowballing" recruitment process. The latter interviews generated no new measures and very few new "leads", indicating that saturation had been reached.

4.5 Data Analysis Method

This section describes the analytical method applied to the data collected in the interviews. The approach taken – and its rationale - is outlined first, followed by a description of the three analytical phases undertaken.

The first phase involves immersion in the data and distillation of key points as individual narratives (narrative analysis) using "open coding" (Neuman 2000). The second phase is the grouping and re-aggregation of key points by topic and theme (topic analysis) using "axial coding" (Neuman 2000). The third phase is the specification and evaluation of the emerging propositions (induction) using "selective coding" (Neuman 2000).

4.5.1 Approach and Philosophical Basis

The primary goal of analysing the interview data collected in the study is to ascertain a summary of the IS industry's "state of the art" in IQ assessment, evaluation and justification within large-scale customer processes. By collecting data about subjects' experiences and roles, the intent is to establish the scope over which such summarisations may hold valid.

In keeping with the over-arching Design Science research method, the secondary goal is to uncover the unstated or implicit *requirements* (or *constraints*) of analysts and decision-makers working in this field. In particular, the organisational importance and acceptability of methods and measures used for justifying investments in IQ is sought.

When constructing a framework for investing in IQ improvements (involving a method and measures), it is necessary to understand both current practice in this regard *and* the likely suitability of tentative new proposals to practitioners. An understanding of currently used measures provides great insights into how measures could be used in a new framework and how they could inform existing practice.

The form of the output of such an analysis is a set of *propositions* induced from the data. To be explicit, it is *not* the goal of the study to build a theory – or comprehensive theoretical framework - of how organisations currently justify their IQ investments. The set of propositions instead constitute a distillation or summary of events and processes to be used in subsequent stages of this research.

Given the use of inductive reasoning to produce propositions, it is worth establishing what is meant by "proposition" here by re-visiting the philosophical basis for this research: Critical Realism. While the deeper discussion of the ontological, epistemological and axiological position adopted in this research project is in Chapter 2 (Research Design), it is appropriate to re-cap and apply those ideas here, in light of this study.

An analysis of how organisations justify investments in IQ is not amenable to the kind of rigorous controlled laboratory experimentation popular in investigations of natural phenomena. We are talking about socially constructed objects like organisational processes, business cases and corporate hierarchies. Hence, lifting wholesale Humean notions of scientific rigour – naturalist positivism, for a want of a better description – would be inappropriate for this task (Bhaskar 1975).

For example, the interview subjects are conceptualising and describing these objects (intransitive dimension) as well as navigating and negotiating their way through them (transitive dimension). The two are inexorably linked: the way an individual manager conceives of a corporate funding process will also reinforce and perpetuate the structure. In Bhaskar's terms, these subjects are operating in an open system.

This mixing of object and subject leads to a mixing of facts and values, in that it is not possible for participants to state "value-free" facts about their social world. Even seemingly factual content, like descriptions of professional qualifications or current organisational role, necessarily contain value judgements about what is included or excluded.

Critical Realism (CR) acknowledges these complexities while recognising that there are still "patterns of events" that persist and can be described. Rather than insisting on the positivist purists' "constant conjunction" of causes and their effects (unachievable in a non-experimental or open system), CR offers "CMO configurations", or Context-Mechanism-Outcome propositions. The analyst seeks to determine regularities (loosely, causality) in a particular context (Carlsson 2003a). Additionally,

certain extraneous factors may "disable" or inhibit this mechanism from "firing" and the analyst's role is to determine these.

The resulting descriptions may be referred to as propositions, but they bear two important distinctions to the classical positivist meaning of this term. Firstly, they are not "facts" as commonly understood in the natural sciences: the entities to which they refer are highly contingent and situated. They are not elemental, atomic, universal nor eternal. Secondly, it is not essential to describe the causal relationships between entities in terms of necessity or sufficiency. Instead, there is a pattern or regularity at the actual level that is observable in the empirical. This regularity may be grounded in the real world, but its action may not be apparent to us.

Consider an example proposition like "Approval must be obtained from the Investment Panel before vendors can be engaged". The entities – approval, panel, vendor – are not real world phenomena in the way that atoms or wildebeest are. The processes – approval and engagement – are similarly contrived and cannot be tested in a closed system, that is, experimentally. While the relationships between them can be characterised in terms of contingency ("if ..., then ..."), this would be to misstate the nature of the causality here.

For example, a sufficiently senior executive may be able to over-ride this configuration (perhaps through by-passing the approval or blocking the engagement). To a naïve positivist, just one instance of this happening would invalidate the proposition and require it to be restated in light of the executive's capacity.

But what if no executive has actually done this? From a CR perspective, if such a possibility exists in the minds of the participants, then whether or not anyone has observed it happening (or has been able to produce this under experimental conditions!) does not undermine the validity of the proposition. To the positivist, the absence of such an observation would render an extension invalid since there is no empirical basis for its support.

From this point of view, we can consider CR to be more robust than positivism and more accommodating of situational contingencies. Alternatively, we could characterise CR as being upfront about the kinds of assumptions that are needed to make positivist inquiry sound and practicable in the social realm.

Following Layder's stratification of human action and social organisation (Layder 1993), this investigation into organisational IQ justification is primarily concerned with the *situated activity* level. That is, how individuals navigate social processes like shared understanding, evaluation and collective decision-making. During the interviews, phenomena at this level are described by these same individuals, so it will involve consideration of the lowest level - *self* - the tactics and "mental models" employed by individuals as they engage in this situated activity.

This situated activity takes place at the *setting* level of large-scale corporate environments, with all the norms and values embedded therein. The top level, *context* (which encompasses macro-level phenomena like political discourse, cultural participation and economic production and consumption), is outside of the scope of this study.

The use of Critical Realism to underpin this investigation means that the "state of the art" of the IS industry in IQ justification can be couched as propositions - CMO configurations in CR terms. Summarising knowledge in this way is not intended to be treated as positivist propositions, with the associated operationalisation into hypotheses and resulting empirical testing. Nor is the set of CMO configurations intended to form a comprehensive theoretical framework.

Instead, these propositions, rigorously established and empirically grounded, can be used to provide guidance in the systematic construction of a framework for investing in IQ.

4.5.2 Narrative Analysis

The data were considered as a succession of narratives, taken one participant at a time, and ranging across a wide variety of topics. Some topics were pre-planned (as part of the interview question design) and others were spontaneous and suggested by the participant.

The first analysis of the data took place during the actual interview. Empirically, this represented the richest level of exposure since the face-to-face meeting facilitated communication of facial expressions and hand gestures which could not be recorded. The audio from the entire interview was recorded, while hand-written notes were taken about the setting (including time of day, layout of the meeting room or office and so on). Notes were also taken of answers to closed questions, unfamiliar (to the interviewer) terminology and some key phrases.

After each interview, a "contacts" spreadsheet was updated with the key demographic information (organisational role, education and so on). Additional "leads" (suggested subjects for subsequent interviews) were also recorded. This information was used to keep track of the "snowballing" recruitment process and to ensure that the sampling criteria were met. This spreadsheet also provided a trail of who suggested each subject, times, dates and locations of the interview, contact details (including email addresses and phone numbers) and notes about whether they'd been contacted, had agreed to the interview and signed the release form.

The second pass was the paraphrasing and transcribing of each interview. Typically, this took place some weeks after the interview itself, by playing back the audio recording and referring to the hand-written notes. Rather than a word-for-word transcription of the entirety of the interview, a document (linked to the contacts spreadsheet) was prepared containing a dot-point summary and quotes. Approximately half of this material was direct quotations with the remainder paraphrased. The direct quotes were not corrected for grammar, punctuation was inferred, meaningless repetition was dropped and "filler words" (eg "um", "ah" and "er") were not transcribed.

For example

"So um you see you see the way we ah tackled this was …"

becomes

"So, you see, the way we tackled this was …"

During the paraphrasing and transcription process, most of the audio was listened to three or four times. This is due to the listen/pause/write/rewind/listen/check cycle associated with transcription from audio. For direct transcription, the length of audio that could be listened to and retained in short-term memory long enough to type reliably was between five and ten seconds. For paraphrasing, it was up to 20 seconds. In some cases, this cycle itself had to be repeated as the audio was poor (with hissing and clicks), some subjects spoke either very quickly, with a non-Australian accent or both. The variable playback feature of the audio device used was extremely helpful here, allowing the audio to be sped up during replay or slowed down

Preparatory remarks and explanations from the researcher that differed little from subject to subject were copied from prior interviews and modified where needed. Some sections of discussion – often lasting several minutes – concerned the researcher's prior work experience, common acquaintances, rationale for undertaking doctoral studies and future plans. While such "small talk" is important for

establishing rapport and helping the subject understand the context and purpose of the interview (in particular, the understanding of organisational processes and terminology), it sheds very little light on the subjects' view of the matters at hand. As such, much of this discussion was summarised to a high level.

The third pass of the data consisted of analysing the textual summaries without the audio recordings. This involved sequentially editing the text for spelling and consistency (particularly of acronyms and people's names) to facilitate text searches. The layout of the document was also standardised. For example, interviewer questions were placed in bold, direct quotes from the subjects were inset and put into italics and page breaks were introduced to separate out discussion by question (as per the interview materials).

The result of this sequential analysis was that some 25 hours of interview data plus associated hand-written notes were put into an edited textual form with a standardised format, linked to a spreadsheet that tracked the subjects' demographic details and recruitment history.

4.5.3 Topic Analysis

Here, the analysis was undertaken on a topic-by-topic basis, rather than considering each subject's entire interview. The units of analysis ("topics") were created by a process of dividing the text data into smaller units and then re-linking related but non-contiguous themes. That is, discussion on one topic – such as the role of vendors – would typically occur at several points during the interview.

This process cannot be characterised as true *open coding* (Neuman 2000) since there was a pre-existing grouping of concepts - the groupings for the set of semi-structured questions initially prepared:

- Context
- System Measures
- Relationship Measures
- Investment Measures
- Conclusion

Within each of these categories, three to five questions were asked (as outlined above in the Materials section at 4.4.2). Along with the "contacts" spreadsheet, these questions formed the basis of a new spreadsheet template ("topics spreadsheet") for recording the subjects' responses, with 21 fields.

The next phase was to consider each of these fields as candidate topics by reviewing each in turn, corresponding to *axial coding* (Neuman 2000). This involved manually copying the relevant text from each subject on that topic and highlighting (in a standout colour) keywords or phrases.

Phrases were highlighted as being significant if they seemed "typical" or "exemplary" of what a large number of subjects reported. Other times, they were selected because they stood out for being unusual, unique or contrarian. These keywords/phrases were then isolated and put into the topics spreadsheet.

In light of this, the topics spreadsheet became a table: each column related to a question while each row described a subject. Each cell contained a list of these topics that arose in the course of the discussion by each subject. The summarisation of the topics in the columns was very straightforward as there was a high degree of similarity between subjects. For example, discussion of Service Level Agreements (SLAs) by the subjects occurred in response to the same questions in many cases.

Thematic consolidation of keywords/phrases (*codes*) between topics was more arbitrary and required more interpretation. An example might illustrate. Discussion of "business case" came up towards the start (Context question 5: *"How does your organisation generally justify investments in your area?"*) and the end (Investment Measures questions relating to Net Present Value and Internal Rate of Return). Quotes (as assertions or explanations) about business cases and their role in the organisational decision-making were found as a response to both questions. As a topic, it is independent of either question and so can reasonably be separated from them. However, the site of its emergence determines its link to other topics. For example, its relationship to process-oriented topics like "approval" and "accountability" lie in the Context question, whereas the discussion of it in the abstract (eg "discounted cash flow") is found in the Investment Measures questions.

The outcome of this topic analysis was a set of recurrent themes or topics (*codes*) that are interlinked and span a number of questions asked of the subjects. These topics relate to information quality and valuation in the abstract as well as particular steps or artefacts used by particular organisations. They form the building blocks of the proposition induction that follows.

4.5.4 Proposition Induction

The propositions – or CMO configurations, in Critical Realist terms – are induced from data using the topics or themes identified during the topic analysis phase (*selective coding*). The elements of context, mechanism and outcome are proposed and evaluated (Pawson and Tilley 1997). In particular, "blockers" are identified (that is, when the CMO was "triggered" but the regularity did not emerge). Supporting evidence in the form of direct quotes is sought as well as any counter-evidence, to enable the balancing of their respective weights.

When specifying the context in which the configuration occurs, it is important to describe the scope, or level. Layder's stratification of human action and social organisation (Layder 1993) is used for this. It is not possible to completely describe the norms and practices from first principles, so as a summary, a large amount of "background information" about corporate processes and IT must be assumed.

The mechanism identified is a regularity or pattern, frequently observed in the specific context and associated with certain outcomes. It should not be understood as a formal causal relationship (that is, as a necessary or sufficient condition), since the events and entities under description have not been formally defined. From a Critical Realist perspective, what we observe in the workplace (through the explanations given by observers and participants) may have underlying causes that are "out of phase" with these observations. The approach of systematic observation through controlled experimentation can reveal these underlying causes in a closed system (eg laboratory), but it is simply not possible in an open system.

It can be difficult to isolate the outcome of interest from the range of possible consequences described in a CMO configuration. This is made all the more challenging when analysing verbal descriptions of events by subjects who themselves were participants and who will undoubtedly apply their own criteria of interest through imperfect recollections. This suggests that to more objectively determine the outcomes, the study could pursue techniques from case study research, such as having multiple participants describing the same events or incorporating supporting documents (such as financial or project reports).

However, the subjects' subjective selection of certain outcomes as worth reporting in the interview (and, implicitly, leaving out others) has significance in itself. The subjects (typically with many years or even decades of experience), when asked to share their views on their experience, are implicitly drawing on their own mental models of "how the world works". Drawing out these patterns is more

useful to the task of understanding existing practice than running an objective (yet arbitrary) ruler over past projects.

The expression of the CMO configurations follows a simple format – a "headline" proposition followed by an explanation of the context, mechanism and outcome identified. Supporting evidence – with disconfirming or balancing evidence – is provided in the form of direct quotes, where suitable.

4.5.5 Summary of Data Analysis

The study uses Critical Realism to identify regularities in the data, in the form of propositions (CMO configurations, in CR terms). These propositions are summarisations of the norms and practices of IS practitioners in IQ justification; as such they capture the "state of the art" in industry. They do not constitute a comprehensive descriptive theoretical framework, but can be seen as an expression of the implicit requirements for the construction of a normative framework for IQ justification.

The propositions were induced from the data in three phases: the first considered each subject's interview data in its entirety, distilling key themes (narrative analysis with *open coding*). The second examined each theme in turn, grouping and re-aggregating the summarisations (topic analysis with *axial coding*). The last pass involved constructing and evaluating the propositions with reference to the original data (*selective coding*).

4.6 Key Findings

This section outlines the key findings from the interview study expressed as Context/Mechanism/Outcome configurations, a technique from Critical Realism. Each section addresses a high-level mechanism pertaining to customer information quality investments (evaluation, recognition, capitalisation and quantification). The configurations are supported by direct quotes from subjects and analysis and interpretation.

4.6.1 Evaluation

P1: *Organisations evaluate significant investments with a business case.*

Context: Organisational decision-making about planning and resource allocation takes place at the situated activity level of stratification. Individual managers and executives, supported by analysts, prepare a case (or argument) for expenditure of resources, typically in the form of initiating a project.

Mechanism: A separate entity (typically a committee) evaluates a number of business cases either periodically or upon request. The criteria for evaluation are specified in advance and are the same across all cases. The process is competitive and designed to align management decisions with investors' interests.

Outcome: A subset of proposals is approved (perhaps with modifications) and each is allocated resources and performance criteria.

All subjects described in some detail the evaluation process for initiatives to be developed and approved. Regardless of their role, experience, organisation or sector, there was an extensive shared understanding about the concept of a "business case" and how it justifies expenditure.

Whether a junior IT analyst or a senior marketing executive, this shared understanding consisted of a number of common features. Firstly, business cases are initiated by a group of employees who see

the need and then pass the request for funding to their superiors. In this, business cases are "bottom-up".

Almost equally widespread was the view that the initiative needs to be driven by the "business side" of the organisation, that is, the marketers, product managers and sales units rather than the technology side. One Telstra executive explained:

In terms of where we spend in data integrity and quality, it predominantly is business initiative driven as well as the ones that we ourselves identify and therefore ask for money – funding – to improve the quality of data. The ones that are business initiatives, it's predominantly an individual business or group – let's say marketing decided to industry code all of their customer base – they would put a business case. So these are the logical steps: the initiator of a business group identifies the business requirements for their own functional group [...] Once they put that business idea together then it's presented to a business unit panel, who then determines whether the idea itself has any merit and if it has then it's approved [...] the resources are approved to go ahead with the idea. So all projects that relate to data integrity or data conformity or data quality go through this process. (S2, Telstra)

This view was confirmed by a marketing analyst:

In a marketing driven organisation, marketing takes the lead on this [investing in IQ initiatives] and it comes back to what benefits we can deliver to customers" (S3, Telstra)

Secondly, the initiators have a belief that the proposal should be undertaken before they prepare the business case. They approach the preparation of the business case as a means of persuading funding authorities to prioritise their (worthy) initiative rather than as a means for determining for themselves whether it should proceed. From the proponents' perspective, the business case is a means for communicating with senior management rather than a decision-making tool for themselves.

Thirdly, there is a large degree of consensus about the way to appraise a business case regardless of its subject matter: financial measures of future cash flows under different scenarios.

The way we structure it is we have a bucket to spend on IT and each group will put up a case of why we need to do this. Projects are driven from a marketing side to deliver improvement but also IT need to be involved [...] There's a bucket of money and people just have to put up their hands to bid for it and senior management will make an assessment on what's the costs and benefits and which ones get priority. [...] In terms of financial measures, it's pretty standardised across the company because everybody has to go through the same investment panels (S3, Telstra)

The role of financial measures in the formulation of the business case – and some alternatives – is discussed further below in some detail. No evidence was found that contradicted this configuration, that is, instances were significant investments were undertaken without even a cursory (or implicit) business case.

Given the sampling strategy for selecting organisations, it's not surprising that there is such a large conformance of views on funding initiatives. After all, financial discipline in public companies stems from a legal obligation of board-members to seek to maximise the value of the shareholders. Approving funding (or appointing executives to investment panels) is perceived as an effective way to discharge this obligation and ensure discipline:

The days of just being able to say 'well, if you give me $5 million bucks and we'll increase the take-up rate on x by 5%' and that's what gives us the money … it really doesn't work like that. You've got to get people to really hone the number that they're claiming. (S10, Telstra)

Interestingly, these funding methods seem to have been replicated in the two organisations that are privately held (ISP and Data). Presumably, this is because the private investors regard these methods as best practice. An alternative explanation is that they may wish to retain the option to take their firms public later, so adopting the methods of publicly-listed companies would increase the value of their shares in the eyes of public investors.

4.6.2 RECOGNITION

P2: *Organisations recognise Customer Information Quality as important.*

Context: Organisations structure their organisation, processes and technologies at the setting level (organisation-wide values, norms and practices). These values, in part, drive resource allocation and prioritisation.

Mechanism: The importance or value of Customer Information Quality is recognised by the organisation through the deployment of resources: appointing managers and creating organisational units, undertaking projects, engaging with service and technology vendors and training employees.

Outcome: Customer Information Quality is conceived as a capital good expected to justify its use of resources in terms of its costs and benefits to the organisation through its flow-on impact on other initiatives.

In order to get a sense of how "Customer Information Quality" (CIQ) is conceived by industry practitioners, it is worth considering how the phrase is used. For instance, there were three executives (one with board visibility) from two organisations with that phrase (or near-synonym, such as "Customer Data Quality") as part of their job title. This suggests that CIQ – in some form or other – must be a principal activity and responsibility for these senior people.

One such interviewee led a team of over 30 analysts (S2, Telstra) while another had in excess of 12 (S8, Telstra). Both had been in their current role for over five years. These staffing levels alone suggest a multi-million dollar commitment by Telstra to Customer Information Quality. Add to this the costs of software, hardware, services and other infrastructure costs related to CIQ operations across the business and we see it is an area of significant expenditure.

This is only the starting point when you consider the groups' respective work in devising training programs for call centre staff to properly collect information, for technologists in processing the information and for managers in credit, finance and marketing in using the information.

Expending a large amount on resources towards CIQ or ensuring accountability for it rests in senior staff is not the only way that organisations recognise the importance of CIQ; awareness throughout the organisation by non-specialists is also an indicator of a pervasive recognition.

Subjects and prospective subjects invariably responded positively to a request to participate in "research about Customer Information Quality". There were no questions about what that is or uncertainty about whether their organisation "did" it - or even why this topic should be the subject of academic research. It is fair to say that recognition of this abstract concept as a topic in its own right was 100%.

Analysts, vendors and researchers all reported some experience with CIQ, indicating when prompted that even if they do not have responsibility for it they regard it as a defined concept that is important to their job. This is not to say that the subjects agreed about the specifics of the definition of CIQ.

Of course, it may be the case that subjects who were confused about the topic, had never heard of it, regarded it as a waste of time or a "buzz-word driven fad" would self-select out of the study. Certainly, if anyone regarded the topic in such a light they did not reveal it during the interviews.

In the 25 hours of interviews, the strongest "negative" statement about the importance of CIQ came from a vendor working on data warehousing:

"I find data quality is an area that no one really cares about because a) it's too hard, it's too broad a field and involves a lot of things. Quality itself is one of those nebulous concepts [...] quality is a dimension of everything you do [...] It's almost, I reckon, impossible to justify as an exercise in its own right. " (S5, Telstra)

This remark suggests that the subject's difficulty is with how to analyse the concept rather than either complete ignorance of the topic or an outright rejection of it.

4.6.3 CAPITALISATION[4]

P3: *Organisations regard Customer Information Quality as a capital good.*

Context: In order to justify use of rivalrous organisational resources like capital and employees, investments are expected to create value for stakeholders (the setting level). Customer Information Quality is not a valuable good in itself, but does create value when used in organisational processes.

Mechanism: This value creation is unspecified, though its effects are observed through increasing future revenues or decreasing future costs associated with servicing customers.

Outcome: The financial performance of the organisation improves through customer process performance.

A strong theme to emerge from the study was the capacity of Customer Information Quality to create business value. Many subjects were at pains to explain that CIQ is not valuable in itself, but that it plays a supporting role in contributing to value-creation in other initiatives, particularly organisational processes that focused on customers:

It's difficult to envisage seeing a business proposal for a project that improves the quality of information by itself – it would most likely be wrapped up in another initiative. (S6, Telstra)

A vendor, with substantial experience selling and implementing data warehouses in corporate environments, makes a similar observation:

It's a means to an end and a lot of people carry on like it's an end in itself. Which is why, I think, it's hard for quality projects to get off the ground because there is no end when presented that way. (S5, Telstra)

[4] *Here, we refer to "capitalisation" as the process of conceiving of a good as a means of further production rather than a directly consumable good. This is not to be confused with the notion of "market capitalisation" as a measure of the market value of a firm.*

Even managers primarily focused on systems and systems development regard CIQ as a value-creating element.

One of the reasons I was interested in the role [of data quality manager] was that time and time again the success of projects depended to a large degree on the quality of the data in the application being delivered. And so often we found that we either had poor quality data or poor interfaces to existing systems. And basically it comes down to: no matter how good the CRM system is it's only as good as the data you present to the operator, user or customer. (S12, OzBank)

A more concrete example of linking CIQ with processes comes from another executive who explains it in terms of the "value chain" concept.

The most successful way of engaging with our business partners is to translate factual data statistics into business impacts and business opportunities. So what we've had to do is engage our customers on the basis that they don't have 50,000 data errors in something - it's that you have data errors and this means you've been less than successful in interacting with that customer and that in turn translates into lost sales or marketing opportunities which translates into reduced revenue which is a key driver for the company. […] The other outcome is reduced cost.

The quality of the information itself does not get much traction in the company it's that value chain that leads to how the company can be more successful in its profitability. […] When we make changes at the data level we can now track that to a change in the business impact. (S8, Telstra)

The focus on couching CIQ initiatives in terms of contributing to or enabling value creation in other initiatives is a necessary consequence of the decision-making process (business case) and the treatment of customer information as a capital good (ie it is a factor of production for other goods rather than one in its own right).

Within process improvement, I'm not sure you'd ever actually go forward with a proposal to improve information quality by itself. It would obviously be part of a larger initiative. The end game is to make a significant process improvement or a breakthrough process improvement and maybe one of things you have to do to do that is to improve the quality of the information you are collecting including there being a complete absence of information in the first place.(S6, Telstra)

It's through this process improvement that business owners (and project sponsors) expect to see value created and are ultimately interested in. This is summarised by an analyst working at the same company:

By doing this what do we get? In practice it's hard to do always, but it's what the business is after. (S3, Telstra)

As he notes, the difficulty lies in translating information systems events into business outcomes to facilitate rational investment. A consultant flags this as a gap in research.

There's a lot of value in looking not just at the data itself in isolation, but looking at how the data is used at the end of the day by people to make decisions and that's an area of […] research that is lacking. (S5, Telstra)

4.6.4 QUANTIFICATION

P4: *Organisations expect to quantify their Customer Information Quality investments.*

Context: Investments are prioritised and tracked and managers are made accountable for their performance. Numbers that quantify business events and outcomes are collected to support the appraisal and scoring of individual initiatives and the management processes that govern them.

Mechanism: Investments are evaluated (beforehand and afterwards) against objective, quantitative, value-linked criteria. These criteria, expressed as financial metrics, are driven by underlying organisational processes.

Outcomes: Financial models of Customer Information Quality investment are created, including assumptions and predictions, to allow comparison between investments and to guide decision-making.

(Blocker): There is no accepted framework for quantifying the value of Customer Information Quality in customer processes.

(Outcomes): Customer Information Quality investments are approved on an ad hoc basis by intuition and personal judgement.

Where a business case exists and is accepted, the usual organisational decision-making applies. One executive with responsibility for information quality (S2) was asked to nominate a project that went "by the book" (that is, with a business case):

That was a program of work which had a business case, it had a driver, it had the sponsors from all the most senior sales people in the company and the resources were allocated. At the end of the day they actually saw the benefits. (S2, Telstra)

This was the exception. Others struggled to recall an instance they were prepared to say followed the normal organisational processes and was, in hindsight, successful.

Sometimes, the rationale offered for an initiative is not explicitly couched in financial terms. Regulatory compliance is an oft-cited example, where there seems to be little discretion for firms. However, the threat of fines or litigation provides a means for cash flows to be considered:

When they're looking at an initiative, if it reduces costs to the business by improving the quality of data you're reducing the cost to the business ongoing. If a company is exposed by the quality of data is not right – meaning, we have so many regulators in our market place these days – that in some cases, if the data is not accurate and the customer complains, then we can get fined from $10K to $10M. So it's removing that risk by making sure that you have taken care of the quality of data and it's at the highest level of integrity. (S2, Telstra)

Another possibility is the loss of reputation (or brand damage) associated with adverse events:

I have heard of one example, not in my direct experience, but there was one at a bank where letters went out to deceased estates, offering them an increase in their credit limits. It's kind of offensive to the people who got those letters that start out by saying 'we notice you haven't been making much use of your credit cards the last three months'. (S1, ISP)

In a similar vein, anecdotes about particular difficulties in conducting normal business operations can percolate upwards and be lodged in the minds of executives:

The local branch managers get a list of their top 200 customers. What they're supposed to do is contact them and 25% of those people don't have a phone number. So how are they supposed to contact and develop a relationship with those customers if you don't even have the bloody phone number? (S12, OzBank)

While there's an awareness of these issues, a reluctance or inability to quantify the damage wrought stifles the response. One senior executive posits a disconnection between operational and strategic management that he attributes to the difficulty in quantifying and reporting on CIQ:

From a managerial point of view it would be great to have those management systems that help you manage your data quality … that you had reports that actually helped you write your business case […] From a strategic point of view, it's making sure that all those things are understood and worked. The strategic point of view has suffered a lot I think because the managerial stuff hasn't happened. […] From a strategic point of view you can say 'well, we need to make sure that we've got the right measures, that we've got our managers looking at data quality, that they're sending through the right feedback to the places that matter so they can quantify it.' Once they've done that, all of a sudden you've got a much bigger strategic issue since all of a sudden it's quantified. (S9, Telstra)

Quantification of the financial impact of CIQ is essential for efficient investment. Its absence makes justification difficult as participants expect comparative performance of alternatives to be expressed numerically. The difficulty in providing such numbers frustrates the ability to get CIQ initiatives approved. In the wider organisational context of projects-as-investments, there is an expectation that CIQ initiatives can prove their worth in a financial sense:

One of the most difficult things from a data quality standard point is when people say to me: Can you prove that this data quality improvement … show me what the business benefit is. And that's difficult because there's no model where you can put the figures in and pop out the answer. There isn't a model that … it depends on the data attributes. (S2, Telstra)

This frustration was expressed by senior executives, in particular. This is because justifying investments through business cases is a principal activity at this level. For example, a senior manager in another part of the organisation echoed this frustration:

I often find it really frustrating that information quality things are made to show a financial benefit … my view is that if you're building quality information that you should take that as a necessary cost and say 'If I don't do it, okay, it's going to be really bad for us in terms of maintaining our market position'. We're prepared to wear a cost here, and then we'll see what benefits are derived in the follow up work. (S11, Telstra)

In terms of the cost and benefit sides of the ledger, there was a consensus that the benefits – particularly increasing revenue – are the most difficult part to anticipate. One very experienced data warehousing manager was explicit about this:

When you get into the space of increasing revenue, that's much harder to justify. That's when you start to get into "guess". You can fairly easily assume that if you knock off ten people there's a cost saving. That's where the problem is with this sort of expenditure [information quality] and justifying it and trying to work out how do you do it from an investment point of view. And that's hard. (S5, Telstra)

With CIQ conceived as a capital good whose value comes from its ability to impact upon a diverse range of business activities, it's to be expected that benefits in particular will be diffused throughout the organisation. Business case discipline ensures that costs are concentrated and visible, but benefits will be intangible. An analyst on the marketing side made this point about "flimsy benefits":

We have got so much data that we need to nail data how we're going to segment customers. What usually happens is that it's all very costly. And sometimes with this type of project it's harder to justify all the costs and returns. This makes senior management a bit hesitant to commit the money to actually do it. As a result, what usually happens is that a lot of the scope went into place and the project is trimmed down to really the bare minimum. It's not ideal but that's the reality, because it's very difficult to quantify the projects. [...] Is the project really worth $4M, or is it $2M? That's really hard to justify on the benefits side. It's a very flimsy benefit. (S3, Telstra)

The point about the difficulty of quantifying benefits is also made by senior people on the technology side. This vendor argues that such quantification is simply not possible and a "strategic" approach needs to be taken:

To justify data quality is a very hard exercise. Mainly because the benefits are invariably intangible, so therefore not very concrete, and tend to be a bit fluffy as well. For instance, if you decide to make date of birth more accurate, you could spend ... for the sake of the argument, say a million dollars doing that. But how do you justify the benefit? To me, you can't. It very much needs to be done I think perhaps at the portfolio scale where you say 'this is a strategic investment, we are doing marketing around xyz space therefore as part of our marketing program we need to have a high level of accuracy with regard to the date of birth field' and drive it that way. (S5, Telstra)

Throughout the discussions with senior executives, it was repeatedly asserted that projects or initiatives need to happen but that the lack of financial metrics for CIQ hampered or frustrated this. The assumption underpinning these views is that the value must be there, it's just that it cannot be articulated. This suggests a problem with the valuation mechanism.

All respondents reported that they were familiar with at least some of the "system-level" Information Quality metrics (validity, currency, completeness and so on). All reported that they were familiar with "investment-level" metrics (Net Present Value, Internal Rate of Return, Return on Investment and so on). There was a high-degree of familiarity with the "customer-level" metrics (cross-sell rates, retention rates, customer lifetime value etc). Many respondents, including all executives on the business side, reported using these in business cases they've either seen or prepared themselves.

For example, a senior analyst working for a data warehousing vendor indicated that:

Business cases I've dealt with tended to deal with this category of [customer] relationship measures rather than system measures as justification for the work. (S4, Telstra)

However, in not one instance did anyone report seeing a business case that explicitly linked "system-level" CIQ measures with investment measures. These system-level CIQ measures were cited as being involved in contract management (S4, S5) or intra-organisational agreements (S11). Two respondents working in direct marketing indicated some specialised metrics impacted on commercial rates charged by information broking business (S13, S15).

The only subject to even mention system-level measures in this context was a marketing analyst. He offered only this vague passing remark and couldn't provide further detail:

… in terms of actual [customer] relationship measures there's a lot of room for individual projects or people as to how they link that back to financial measures. And systems measures usually get popped in there. (S3, Telstra)

This disconnection between CIQ quantification and investment quantification means alternative approaches must be found if initiatives are to proceed. Some respondents (eg S5 above), advocate taking a "strategic" approach. This seems to mean bypassing the discipline of the formal business case and relying on the intuition or judgement of senior people:

A lot of our data quality business cases have been approved because they've been seen to be strategically important and they haven't had value written around them. The strategic side is going to be pretty hit and miss unless you get the managerial side happening. (S9, Telstra)

The impact of the "hit and miss" nature of this approach on resource allocation was discussed by another senior executive at a retail bank:

When I say the bank's not very mature in allocating scarce resources, the prioritisation is a good example of how that works because a lot of it is to do with who yells loudest … it's the 'squeaky wheel' and how much effort's gone into preparing what needs to be done … I haven't seen much science go into 'let's evaluate this proposal' … So that's the whole bunch of competing proposals going forward and data quality being one of those and then 'saying okay, where does all this fit? Who's going to benefit from this?' … Data quality is not sexy, it is just not sexy. (S12, OzBank)

Note that the subject is indicating here that the shortcomings of this approach are two-fold: an inability to compare between CIQ initiatives and an inability to compare CIQ initiatives with other proposals.

Unsurprisingly, one executive reported that the "strategic" or intuitive approach finds less acceptance amongst finance professionals. The explanation offered is that people working directly with the data are better placed to understand the costs:

It's interesting that it's the marketing people now who get the message who have the best understanding of what it [data quality] is costing in terms of bad data and what is achievable if they improve the data. So it's marketing people who are driving this [data quality improvement]. The [credit] risk people are not far behind. The finance people [shakes head] … very hard, very hard. (S1, ISP)

When asked about methodologies or standards or generally accepted principals for capturing or articulating the benefits of CIQ initiatives, none was nominated by any of the subjects. While this isn't proof that none exists, it does suggest that, if it does, it is not widely known.

This was confirmed by one manager who reported a view that his organisation is not suffering under a competitive disadvantage because the problem is widespread throughout the industry:

No company's doing this [IQ investments] really well. It's a problem that's always been there. Some are just devoting more time and money – and probably intellectual capacity – to fixing up these issues because of down-stream dependencies. (S11, Telstra)

Again, the suggestion is that differences in investment (time, money and "intellectual capacity") are due to the visibility of problems "down-stream".

4.6.5 The Context-Mechanism-Outcome Configuration

Based on the above analyses, a tentative explanation for how practitioners understand and approve Customer Information Quality improvement initiatives is presented.

The account is presented in two parts. Firstly, there is the *normative* model of how decisions should generally be made in larger private-sector organisations (the *context*). Essentially, this is the view that decisions are made about initiatives by evaluating them as investments (P1) using a business case with financial metrics (the *mechanism*). The outcome is an optimal set of proposals going forwards, with uneconomic ones rejected.

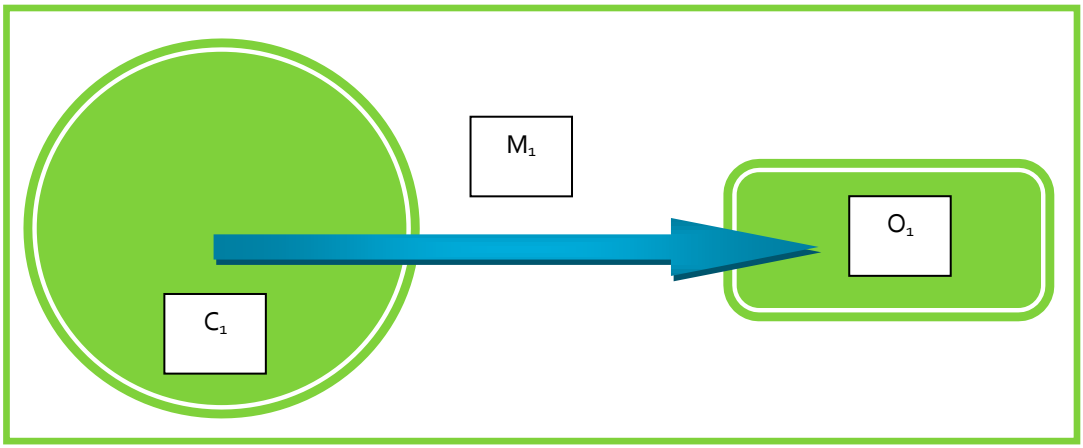

FIGURE 5 NORMATIVE CMO CONFIGURATION

Element	Description
C_1	At the *setting* level, this context includes shared values about the organisation's goals, responsibilities to stakeholders and what constitutes good organisational practices. Also present is recognition that CIQ (like many other goals) is important (P2).
M_1	The mechanism is the funding approval process that determines which initiatives proceed to implementation. This mechanism relies on a comprehensive quantitative assessment of the financial costs and benefits of each proposal in a way that allows direct comparison (P1).
O_1	The outcome is a set of all initiatives or proposals which will optimally meet the organisation's goals. (That is, their benefits exceed their costs.) Initiatives which would detract from this optimality are excluded.

TABLE 9 NORMATIVE CMO ELEMENTS

The second part is a *descriptive* model of what happens in practice for CIQ improvement initiatives in particular. Participants have an expectation that the general normative model can be used as a template. (This expectation forms part of the *context*.) However, the *mechanism* (assessment of costs and benefits) cannot always operate in this case. As a result, an alternative mechanism may be employed resulting in a less-than-optimal *outcome*.

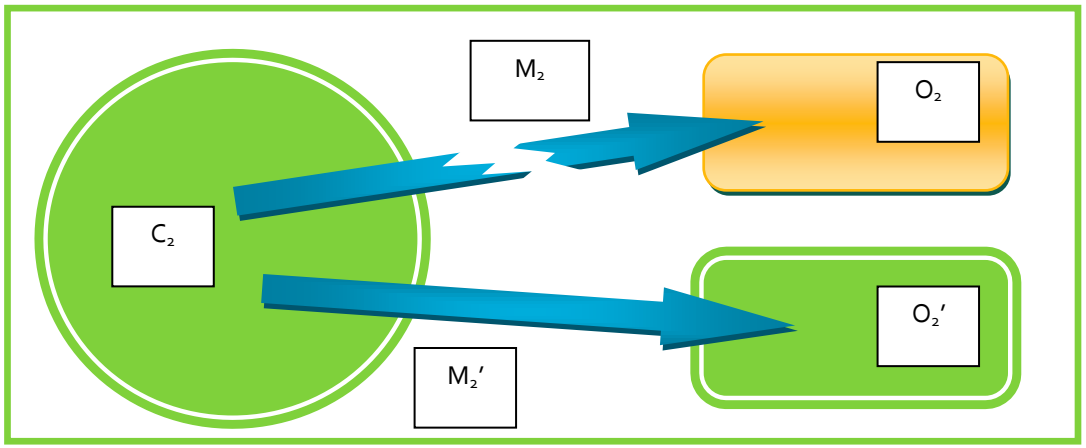

FIGURE 6 DESCRIPTIVE CMO CONFIGURATION

Element	Description
C_2	At the *situated activity* level, this is the context of individuals navigating a complex social realm of organisational processes and hierarchies to advance their agendas. It encompasses the normative configuration (C_1, M_1, O_1) outlined above as well as beliefs about the best course of action for managing the organisation's systems and processes.
M_2	This mechanism is to conceive of CIQ as a capital good (P3) and then apply metrics to systems and processes to quantitatively inform the business case of how candidate CIQ initiatives' costs (direct and indirect) impact on the organisation's financial position, including benefits of increasing revenue and reducing costs (P4). As is shown, this mechanism cannot operate ("blocked") because the quantification of the financial impact is not possible, not undertaken or not accepted by decision-makers.
O_2	The outcome is a set of initiatives to proceed (and others that are declined) where the formally estimated portion of total benefits that can be articulated exceeds the cost.
M_2'	This mechanism is an alternative to the financial-metric led business case described in M_2. Here, the commercial judgement (based on experience and intuition) of a sufficiently senior manager is used to approve (or deny) CIQ proposals without reference to quantitative assessments of costs and benefits.
O_2'	This outcome results from the use of the alternate "strategic" approval mechanism, M_2'. It is the set of initiatives which are perceived by the manager as having sufficient worth to proceed. This includes the set of initiatives where the (unspecified) benefits are judged to exceed the (unspecified) costs, as well as others.

TABLE 10 DESCRIPTIVE CMO ELEMENTS

The outcomes in the descriptive model are likely to be sub-optimal. That is, the set of initiatives approved (and declined) will be different to the outcome if full knowledge of the costs and benefits were available. (It's worth noting that perfect knowledge of such things will always be unavailable; however, some estimates are better than others.)

Under the first outcome (O_2), an organisation is likely to experience under-investment in CIQ initiatives. This is because initiatives that would have created value are declined due to a soft benefit. Under this circumstance, actors may proceed to use the second mechanism if it is available.

The second outcome (O_2') relies on a fiat decree that an initiative should proceed. This introduces three potential problems: firstly, the wrong ones may be approved by virtue of their visibility (characterised by one executive as "the squeaky wheel") rather than a cool assessment of the pros and cons of competing alternatives. This could lead to either over or under-investment.

Secondly, the absence of a quantitative financial basis hampers industry standard project governance practices (such as gating) and management performance evaluation (eg key performance indicators).

Finally, bypassing the established organisational norms about how important decision-making *should* proceed (C_1) undermines confidence in the approval mechanism (M_1). Individual workers in the organisation may feel resentment or distrust of more senior figures if they are seen to exercise power arbitrarily, opaquely or capriciously. Further, in the private sector, this will impact upon shareholders, who rely on the discipline of business cases to ensure their interests are aligned with management's.

4.6.6 CONCLUSION

The application of standard approval mechanisms to Customer Information Quality improvement initiatives requires a comprehensive quantitative assessment of the financial costs and benefits associated with undertaking.

CIQ, as a capital good, is not a valuable end in itself. It creates value for an organisation through its capacity to improve customer processes. The causal relationship between improvements in customer information quality (as measured in the organisation's repositories) and creation of organisational value (as measured by financial metrics) is not well understood. There is no widely-accepted method or framework for undertaking this analysis.

The potential value of CIQ improvement initiatives are understood (and proposed) by people "on the ground" who deal with the systems, processes and customers in question. The lack of clear measures of organisational benefits is of particular import. The inability to articulate this perceived value to senior decision-makers means that valuable initiatives are declined. Alternatively, an initiative without a financially-supported business case may still proceed by fiat.

The outcome is a significant risk of resource misallocation. This can arise from under-investment (when benefits are seen as "flimsy" or "guesses") or over-investment (the "squeaky wheel" is funded). Additionally, these mechanism failures can undermine confidence in the organisation's collaborative decision-making processes.

Chapter 5

Conceptual Study

CONCEPTUAL STUDY

5.1 SUMMARY

This chapter is a conceptual study of Information Quality (IQ) undertaken in order to develop a framework for IQ valuation. It evaluates and synthesises concepts from theoretical reference disciplines (including Information Theory, semiotics and decision theory) from the Literature Review (Chapter 3), motivated by the requirements from the practitioner Context Interviews (Chapter 4).

As part of the Design Science methodology, this constitutes artefact design, where the artefact is a framework comprising of a conceptual model, measures and methods for analysing the quality of information in Customer Relationship Management (CRM) processes.

The outcome is a target framework for valuing IQ improvements, with a view to organisational uses including business case development, performance evaluation and inter-organisational agreements. Subsequent chapters will evaluate this framework for rigour, relevance and usefulness.

5.2 PRACTICAL REQUIREMENTS

This section addresses the context, intended use and goals of the framework under development. These are motivated by the findings from the practitioner interviews (Chapter 4), which identified a gap in Information Systems (IS) practice when it comes to IQ. The interviews also provided insights into the problem domain, for which a properly-conceived framework may prove useful. The role of design in the framework development is through the selection, composition and evaluation of abstract theoretical concepts for practical ends.

Thus, the development of this framework (including evaluation) is a Design Science research project. The creation of organisation-specific information-value models by business analysts is also a design science activity, in the sense that it involves the development of an artefact (model). However, the development, use and evaluation of these concrete models is not within the scope of this research project. The framework for creating such models is the object of analysis.

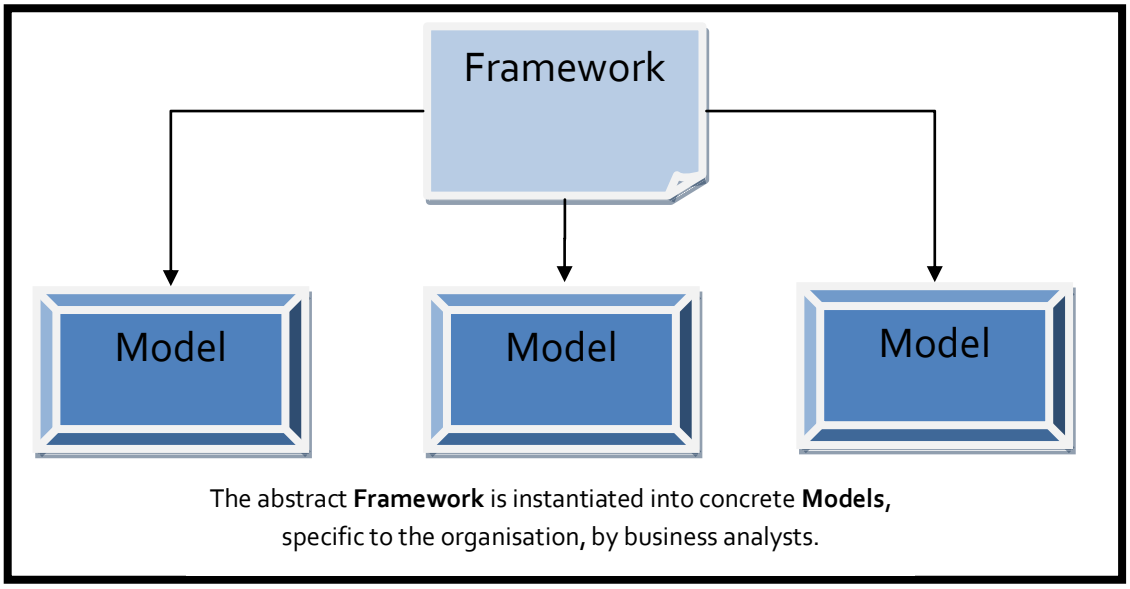

Figure 7 Use of the Designed Artefact in Practice

5.2.1 Organisational Context

The following points are a distillation of the analysis of this organisational context from the practitioner interviews (Chapter 4). Large organisations which maintain complex relationships with customers rely on high-quality customer information to operate successfully. To acquire and manage this information requires significant expenditure on systems, including capital and operational items. These can include information technology, staff training, auditing and testing activities, customer communication and vendor and supplier management.

In these organisations, significant resources are typically allocated to projects on a competitive funding basis, whereby projects compete for access to a capital budget, assessed by an investment review panel or similar decision-making body. The project owner develops a business case for the expenditure of resources, couched in investment terms and supported by financial models and metrics. Return on Investment and Net Present Value (along with other discounted cash flow approaches) are used the most.

While IQ is recognised by organisations as being important, it is difficult for IQ projects to compete for access to resources when the business case cannot be articulated. The problem stems from an inability to quantify the impact of IQ improvement, and to express this impact financially. As a result, there is likely to be significant and widespread under-investment in IQ, contributing to inefficiencies in resource allocation, customer dissatisfaction and competitive risks.

In some instances, IQ projects are approved by support from a sufficiently senior executive, relying on judgement. In addition to the potential for misallocation, this can also result in undermining confidence in the capital investment program in the organisation, characterised by one executive as "the squeaky wheel gets the oil".

This investment problem does not arise from a lack of IQ measures: practitioners are aware of and use a number of measures to describe their systems. Nor is it due to a lack of financial sophistication among practitioners, with many managers, analysts and consultants experienced in preparing IS business cases. The key gap identified lies in conceptually linking the quality of customer information to financial outcomes in a way supports organisational decision-making. Specifically, in linking the

two in a way that allows consideration of alternative courses of action: diagnoses of key value drivers, speculative "what-if" scenario testing and the evaluation and prioritisation of interventions.

5.2.2 PURPOSE

The framework is designed to be used by analysts to prepare a model of how IQ impacts on outcomes in customer processes. This model is to be used for organisational decision-making, primarily business case development. Further, the model may be useful for designing and setting service level agreements (for suppliers) and key performance indicators (for staff). In each case, an understanding of how IQ impacts on financial outcomes will better align the interests of managers with those of the organisation.

The use of the framework to develop such a model will also help the organisation better understand its own information supply chain. That is, it will foster an understanding of how customer information is used to create value within the organisation, the relative importance of different information elements for different decision tasks and the true costs associated with low-quality customer information.

Based on the extensive practitioner interviews, it seems that managers and analysts "close to the data" generally have firm views on where the IQ problems lie and how to go about fixing them. From this perspective, the development of such information value models is not seen as supporting their decision-making about diagnostics or intervention design; rather, it's a means of communicating the problems and opportunities to key decision-makers "higher up" in the organisation.

These models can also help further the shared understanding between the owners of customer processes ("business") and the managers of the supporting infrastructure ("IT"). By addressing the so-called alignment problem, prioritisation of work and planning should be facilitated, as well as improvements in working relationships.

5.2.3 OUTPUTS

The framework, as an artefact in itself, is instantiated as a collection of concepts, formulae, measures and tasks for describing and modelling aspects of customer processes. As such, it is necessarily abstract and highly theoretical. The framework is to be used by analysts to prepare an information value model tailored to the target organisation.

This output model has a number of components that describe and link:

1. Information elements.
2. Customer processes.
3. Quality interventions.
4. Organisational outcomes.

Depending on the scope of the analytic effort, these may be mapped at a level of great detail or more superficially, by focusing on just the key aspects of the organisation.

As a bridging model spanning IT, operations and finance, the terms and quantities should be familiar to professionals in those areas, where possible.

5.2.4 PROCESS

The framework is employed to produce information value models for the organisation. There are precedents for developing such artefacts within large organisations that invite comparison. For example, on the business side, most organisations produce and use cash flow models of their business activities. These capture the (expected) flow of cash over time from customers and to

suppliers across different organisational units and are used to support planning and evaluation activities. On the IS side, many organisations conduct data modelling, where they document (and sometimes mandate) enterprise-wide definitions of entities, relationships and processes. Another example is the statistical or data mining models developed to support some aspect of operations, such as logistics, credit or marketing.

In each case, these artefacts are valuable organisational assets. They require a non-trivial effort to generate and a commitment from the organisation to adhere to or use them to realise that value. Further, they require ongoing maintenance or review to allow for changes in the operating environment or the organisation's strategy. Lastly, they can be shared across organisational units (such as subsidiaries) or even outside the organisation, with trusted partners.

In common with these types of models, the proposed information value models would follow a similar high-level lifecycle of scoping, development, deployment and maintenance. Responsibilities and resources could be allocated in a similar fashion to other enterprise projects. Expertise would be required from the information management function, the customer process owners and the finance unit.

The theoretical concepts that underpin the model are, necessarily, complicated and not widely available. This is because the key to the model lies in the quantification of information, which inherently demands the use of comparatively advanced statistical techniques. However, a thorough understanding of these concepts should not be required to construct and interpret models.

In terms of technology, the models themselves can be expressed using spreadsheets or similar programmable calculation environments; no specialised software or hardware is required. As artefacts, these models would represent the distillation of knowledge of how the organisation acquires and uses customer information to create value. The effective sharing and security of such an asset must also be carefully managed.

5.3 Theoretical Basis

The framework under development must be grounded on a sound theoretical basis. This is because, for the artefact to be useful, it must generate models that describe what they purport to describe: the relationship between IQ in customer processes and organisational value.

This study draws on four reference theories, discussed in detail during the literature review (Chapter 3). As I have previously argued (Hill 2004) these reference theories provide sufficient conceptual and quantitative rigour for modelling of information and value.

- **Semiotics**. This is the formal study of signs and symbols and provides an over-arching hierarchy for organising discussion of data and information.
- **Ontological Model of IQ**. This maps the relationship between information systems and the external world they are intended to represent.
- **Information Theory**. This mathematical theory is used to quantify the amounts of information within the models.
- **Information Economics.** This theory is used to value the use of information for decision-making.

With the exception of the Ontological Model, these theories have their own long-standing traditions and conventions and have been applied to a wide variety of situations. In this context, semiotics has been used to tackle IQ from a purely conceptual perspective (Shanks and Darke 1998), while the Ontological Model (Wand and Wang 1996) is a rigorous general theory of IQ. The Ontological Model

comprises the semantic level in the Semiotic Framework for Information Quality (Price and Shanks 2005a). The semiotic framework provides the starting point for this analysis.

The inclusion of Information Theory (Shannon and Weaver 1949) is necessitated by the practitioner requirement for quantification of IQ. Information Theory has enjoyed widespread success in this task in other applied disciplines, such as communications engineering, psychology, economics and genetics (Cover and Thomas 2005). Further, Information Economics has been included to enable practitioners to explain and characterise their models in financial terms, an identified gap in accepted methods for valuing IQ.

5.3.1 SEMIOTICS

Semiotics, the study of signs and symbols in the most abstract sense, is a philosophical discipline that underpins linguistics, critical theory and related fields. At the core of the modern theory is the concept of a *sign* (Chandler 2007). This is a very general notion: a sign could be a traffic light used to control the flow of traffic, an article of clothing worn in a particular way or text written on a sheet of paper. Semiotics, in the Peircean tradition, is the study of the triadic relations between the sign's (physical) *representation*, its *referent* (intended meaning) and *interpretation* (received meaning). As Price and Shanks note:

Informally, these three components can be described as the form, meaning, and use of a sign. Relations between these three aspects of a sign were further described by Morris as syntactic (between sign representations), semantic (between a representation and its referent), and pragmatic (between the representation and the interpretation) semiotic levels. Again, informally, these three levels can be said to pertain to the form, meaning, and use of a sign respectively. (Price and Shanks 2005a, p218)

The authors use this stratification into syntax (form), semantics (meaning) and pragmatics (use) as an organising principle for collating and rationalising a number of commonly-used IQ goals and criteria. The syntactic level is the domain of integrity constraints and conformance rule checking. Here, I focus on the semantic level (correspondence between the information system and the external world) and the pragmatic level (the impact of the information system upon organisational decision-making).

The reason for this is two-fold: firstly, the semantic level subsumes the syntactic in the sense that flaws in the syntactic level will manifest in the semantic level. For example, a syntactic problem like a malformed expression of a date ("2005-20-a3") will result in a semantic (meaning) problem. The second reason is due to the scope of the study. With the emphasis on organisational value, the framework focuses on how meaningful customer information translates into action.

Following the Semiotic Framework for IQ, the semantic level is analysed in terms of the earlier Ontological Model for IQ to derive the criteria. However, here the Ontological Model is quantified using Information Theory and extended to include the pragmatic level. This is further analysed through an economic analysis of the (value) impact of information upon decision-making and action-taking within customer processes.

5.3.2 ONTOLOGICAL MODEL

This model of IQ, proposed in 1996 by Wand and Wang, is a clear expression of the relation between an information system and the external world it purports represent. It is a rigorous and theoretically sound approach to analysing this relation, based upon the idea of "states of nature" (Wand and Wang 1996). In this model, both the information system (IS) and the external world (EW) are taken as two (related) sub-systems of the physical world, each of which is governed by laws and must assume precisely one state (out of many) at every point in time. The model captures the essential nature of

an IS in that the IS "tracks" the EW in some significant way. An IS user must be able to infer the underlying state of the EW based on observing only the IS. Based on this key insight, the authors proceed to establish the technical conditions under which this is possible.

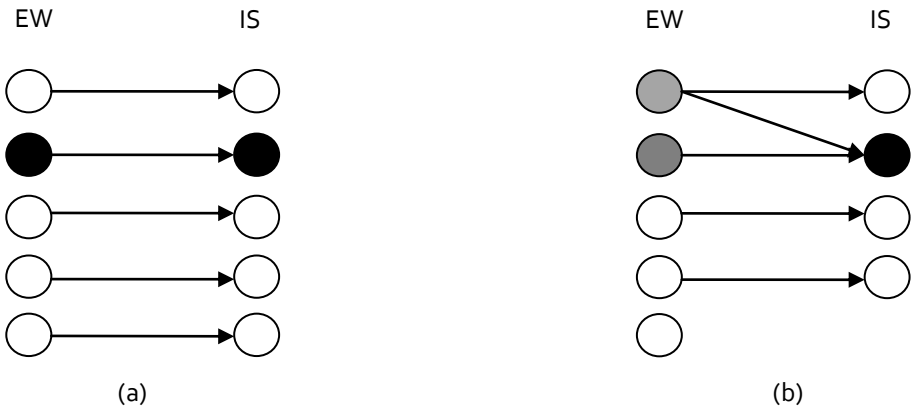

FIGURE 8 ONTOLOGICAL MODEL (A) PERFECT (B) FLAWED.

The simple examples here illustrate the concept. The columns of circles represent the five possible states of interest, forming a state-space σ. "EW" refers to the external world, which has five possible states. In the context of customer processes, this could be the preferred title of address (eg "Mr", "Ms", "Dr" and "Prof"). At any point in time, each customer has precisely one title – not two, not zero. The purpose of the IS is to capture this by providing five states and maintaining a mapping between the possible external world states and the internal states of the IS. Wand and Wang refer to this process as representation (*rep*). The inverse – interpretation (*int*) – is the process of inspecting the IS state and inferring what the original external world state is.

So, for (a) this works perfectly: all states of interest are captured and the IS represents perfectly the EW. However, for (b), a number of deficiencies or flaws have been introduced. Firstly, there is a "missing" state: the fifth EW does not have a corresponding state in the IS. This is a design problem. Secondly, there is ambiguity during interpretation, arising from a representation problem. Observing that the IS is in the second state does not conclusively inform us of the original EW state: it could have been the first or second state since both could result in the IS being the second state. Based on similar considerations, the Ontological Model identifies four possible deficiencies in the mapping between the external world and information system.

- **Incompleteness**. An EW state of interest cannot be represented by the IS.
- **Ambiguity**. An IS state maps to more than one EW state.
- **Incorrectness.** An EW state maps to an IS state such that the inverse mapping cannot recover the original EW state.
- **Meaninglessness.** An IS state that does not map to a valid EW state.

Note that the Price and Shanks Semiotic Framework adds the concept of "redundancy" as a deficiency, on the grounds that multiple IS states mapping to a single EW state introduces the potential for other IQ problems (Price and Shanks 2005a). However, subsequent focus groups with practitioners suggest that redundancy is not necessarily an issue for them, equating it with the notion of "replication" (Price and Shanks 2005b). This suggests the practitioners did not understand the distinction between multiple instances of a data set (that is, replication in the database sense) and multiple equivalent possible states in an IS.

In the example used here, making a copy of the file is redundancy in the informal "replication" sense. This may or may not be a good idea, as the practitioners reported. However, in terms of the

Ontological Model under discussion here, redundancy would be adding a sixth state, "Mister", which is semantically equivalent to "Mr". Regardless of which one is chosen, it will always be possible to infer what the EW state is. Since these two are semantically equivalent, the states can be treated as one "merged state" by the IS, without loss.

In general, if one state is a perfect synonym for another state under all conditions and possible uses, then its existence cannot introduce semantic errors. If there is a meaningful distinction to be made between them, then it is not a redundant state and its inclusion is a valid design choice. Accordingly, non-redundancy is not required here as a semantic criteria.

Also, the original definition of "completeness" used by Wand and Wang was restricted to mappings where the EW state *cannot* be represented by the IS (as above), indicating a design problem. The modified definition in the Semiotic Framework states that it arises where the EW *is not* represented, expanding to include operational problems such as "when a data entry clerk manually entering data into the IS accidentally omits an entry" (Price and Shanks 2005a).

Here, "missing data" is regarded as an operational correctness problem. If the IS is in the "null" state (that is, accepts the empty field) so it is not possible to infer the original EW state, then it meets the definition for being incorrect.

Consider the example of customers' preferred titles above. If a customer is filling in an enquiry form displaying the five options and ticks multiple boxes, then this is ambiguous. If a customer leaves no box ticked, then this is a valid (but incorrect) state of the information system (enquiry form). If there is no option for "Fr" when this is a state of interest, then the enquiry form is incomplete. The completeness (or otherwise) of the relationship between the states is determined by the *possibilities* (design), not the *actuality* (operation).

Using the term "completeness" to characterise the expressive power of a language has a long history in theoretical computer science and meta-mathematics (for example, Gödel's Second Incompleteness Theorem). The definition of incompleteness does not need to be expanded to incorporate the informal meaning of "missing data", as long as the context makes it clear.

From the earlier discussion of semiotics, it is clear that this Ontological Model falls in the semantic level as it addresses the *meaning* of signs. That is, the relationship between the representation (IS) and its referent (EW). The four criteria for semantic quality are that the relationship is complete, unambiguous, correct and meaningful. Departing from the Semiotic Framework, the model adopted here uses the original definition of *completeness* and excludes the additional concept of *redundancy*.

Conceptually, these definitions are clear, concise and sufficient for explaining the semantic deficiencies of an IS. However, not all flaws are equivalent and it is not clear how best to quantify them for comparisons. The next section proposes how Information Theory can be used for this purpose.

5.3.3 Information Theory

The Ontological Model of IQ provides a very clear basis for modelling the relationship between an IS and the EW. The definitions for semantic quality criteria that arise from it are logical, not quantitative. That is, a mapping is either complete or its not, or it's correct or not. The definitions do not allow for degrees of ambiguity or grades of meaninglessness. In practice, no mapping is going to be perfect in all criteria so this raises the issue of how to compare deficient (potential) EW/IS mappings.

Most obviously, we could count the deficiencies. Using these logical definitions, it may be possible to identify one IS as missing three states (incompleteness) while another is missing four. Or one IS has four ambiguous states while another has six. This simple counting strategy may work with one criterion (such as completeness), but it's not clear how this approach could be adapted to make comparisons across multiple criteria, for example, to make design trade-offs.

Further, consider two EW/IS mappings of the same external world, both with a (different) single meaningless state. On a naïve counting basis, these might be regarded as equivalently deficient. Suppose that, in operation, the first IS never gets into its meaningless state while the second one is frequently found in it. It seems reasonable at a common-sense level to infer that the second deficiency is worse than the first.

Lastly, comparison between mappings when the underlying EW is different is fraught too. If one mapping has 25% of its states incorrect and another has 20%, is that worse? What if the former has hundreds of states while the latter just ten? What's needed is a reliable, objective and quantitative method for assessing and comparing the semantic quality of the EW/IS relationship.

The most natural approach to try is Information Theory. Developed by Shannon in the context of communications engineering[5] after World War II, it has evolved into a significant body of rigorous mathematical research, underpinning a range of applied activities spanning economics to linguistics to genetics (Cover and Thomas 2005).

Building on earlier mathematical ideas from Fischer, Hartley and Nyquist, Shannon showed how to quantify the amount of information conveyed in a message through a channel (Shannon and Weaver 1949). His first key insight was that a message is a selection (choice) from a range of possible alternatives. When the same selection is made by the sender (source) and the recipient (receiver), the message is deemed to have been communicated. Shannon's second key insight was that information is the reduction of uncertainty. As a result of receiving the message, the recipient's beliefs about the world change. (In semiotic terms, this process is called *semiosis*.)

Shannon's remarkable achievement was to develop and present a unified and coherent model to quantify both of these insights: the amount of information in a source and the amount conveyed in a channel. As his basic model is isomorphic to the Ontological Model outlined above, this approach is applicable here too.

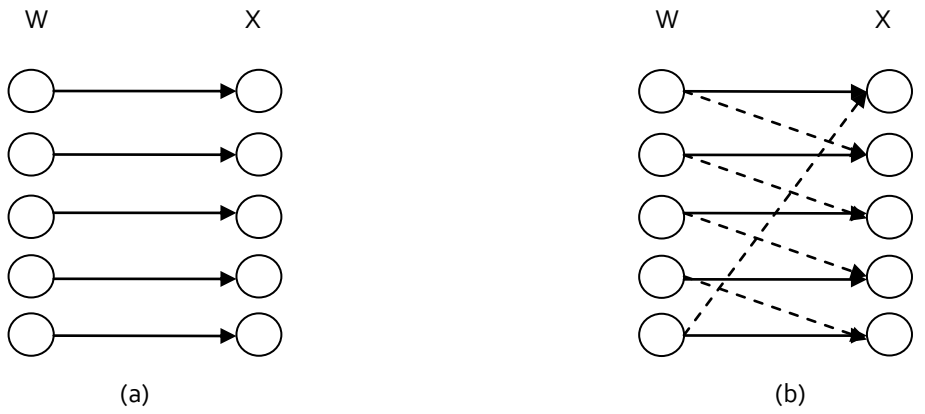

FIGURE 9 SIMPLIFIED SOURCE/CHANNEL MODEL PROPOSED BY SHANNON

[5] *The paper that launched the field was originally called "A Mathematical Theory of Communication" in 1948.*

In this simplified model (the channel encoding and decoding stages have been omitted), the sender (W) selects one of five possible messages (w_1, w_2 ... w_5). The receiver (X) must determine which of the five possibilities was sent. If both W and X agree, then the message was successfully transmitted. If X selects a different message, then we say the message was *garbled*. In the case of the "noisy channel", (b) above, garbling is indicated by the dashed arrow. For example, w_1 could result in x_1 being received, or x_2. If a w_1 was sent, then either x_1 or x_2 could be received. Conversely, if x_2 is received, it could be the result of either w_1 or w_2 being sent.

At a conceptual level, the specific medium of transmission (the channel) does not matter; it only matters whether the source and receiver agree on what was sent. This is purely a function of the *probabilities* of a particular message being garbled. Mathematically, a channel is characterised as a transition matrix, where the elements are the conditional probabilities. For the case of a perfect channel (Figure 9a), the transition matrix is the identity matrix ie ones on the diagonal and zeroes elsewhere.

$$\begin{array}{cc} & \text{X} \\ \text{W} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

FIGURE 10 CHANNEL AS A TRANSITION MATRIX

For a particular set of messages to be sent, some channels will be better than others, with none being better than the perfect case (a). (That is, more of the information in the source will be able to reach the receiver.) The figure of merit for assessing different channels is called the *mutual information*, or *transinformation*.

But how much information is in the source initially? Information Theory states that the amount of information is not determined by the number of symbols in the message, but how *likely* it is that a message is selected. When all options are equally likely, the amount of information in the source is at a maximum. Based on mathematical arguments, Shannon presents the *entropy function* as the appropriate function to quantify the amount of uncertainty (Shannon and Weaver 1949). The amount of information in the source, W, is given by the entropy (or self-information):

$$H(W) = -E[\log p(w)]$$

$$= -\sum_i p(w_i) \log\ p(w_i)$$

This quantity reaches a maximum when all source symbols w_1 through w_5 are equally likely to be sent (ie with a 20% chance). This quantity is log 5 ≈ 2.32 bits[6]. The derivation below uses logarithm laws to show this.

$= -\sum_i \frac{1}{5} \log \frac{1}{5}$

$= -5(\frac{1}{5} \log \frac{1}{5})$

$= \log 5$

So what is this the meaning of "2.32 bits"? One natural interpretation is that it must take, on average, at least 2.32 well-formed "Yes/No" questions to deduce which of the five messages was being sent. At the other extreme, if the only message ever sent was w_3 (probability of one, all the others are probability zero), then the amount of information in W is zero. (Here, the convention that 0 log 0 = 0 is adopted, following arguments by limit.) This satisfies our intuition that a deterministic source must have zero information.

Using this definition of entropy, we can also define the mutual information of the transition matrix $p(W=w_i|X=x_j)$, as used to characterise the channel:

$I(W;X) = H(W) - H(W|X)$

In words, the mutual information between two random variables W and X is the uncertainty about W minus the uncertainty about W given X. It is the difference between the uncertainty about W before observing X and afterwards. That is, how much uncertainty about one variable (W) is "left over" after observing a second variable (X). Mutual information reaches a maximum of H(W) when H(W|X) = 0. That is, when observing X is sufficient to extinguish all uncertainty about W.

Similarly for the correlation co-efficient, if W and X are statistically independent then their mutual information must be zero (observing X cannot tell us anything about W).

Armed with these definitions, I can systematically and objectively quantify the Ontological Model of IQ. The essence is to conceive of the external world as being the *source* and the information system as the *receiver*. The external world sends information about its state to the information system (representation), while users of the information system are able to infer the external world's state by inspecting the information system (interpretation). The semantic quality of the IS is determined by how well it mirrors the EW. Now, with Information Theory, I can quantify this as follows.

Firstly, the source messages in W (w_1 through w_5) correspond to the states of the EW, while the received messages in X (x_1 through x_5) correspond to the states of the IS. The probabilities of these states occurring are known *a priori*, perhaps through historical observation. The transition matrix, T, describes the probabilities of each state being garbled (that is, incorrect) upon receipt. The amount of information in W is the entropy in W, H(W). This is the amount of uncertainty in W resolved upon observing W (hence the term "self-information"). The amount of information in X is similarly defined as H(X). The measure of semantic quality of the information system is defined as the normalised mutual information between W and X, which is dubbed here "fidelity":

[6] *Note that all logarithms used here are base 2, unless otherwise stated. This means the unit for information is bits.*

$$F = \frac{I(W;X)}{H(W)}$$
$$= 1 - \frac{H(W|X)}{H(W)}$$

Conveniently, this measure ranges from 0% to 100%. When F=0%, it implies "perfect failure", that is, H(W|X) = H(W) so that observing the IS tells us nothing at all about the EW. (The IS is formally useless.) When F=100%, it implies "perfect information" and H(W|X)=0 so that observing the IS reduces our uncertainty about the EW to zero. (The IS is formally infallible.) Real ISs are likely to be somewhere within this range.

The name "fidelity" is chosen here because it means "faithfulness" (from the Latin *fide*), in the sense of how "faithfully" the IS tracks the EW. In using the IS as a proxy for the EW, users are trusting the IS to be a "faithful" substitute. It is in this sense that audiophiles use the terms "hi-fi" and "lo-fi" (high fidelity and low fidelity, respectively).

This measure captures the operational semantic criteria of ambiguity and incorrectness. An increase in ambiguity entails a dispersion of the probabilities in the transition matrix, resulting in a decrease in fidelity. Similarly, an increase in incorrectness entails moving probabilities off the main diagonal of the transition matrix, also decreasing fidelity. It is possible to decrease the ambiguity and increase the correctness, or vice versa, while increasing the fidelity.

The design semantic criteria of completeness and meaningfulness relate to the presence of all and only all state in the IS needed to capture states of interest in the EW. Suppose there is an "extra" state in the EW (w_6, implying incompleteness). If there is a zero probability of this state being occupied, then it – literally – doesn't count. If there's a finite chance of the EW getting into this "extra" state then the IS must be in *some* state (because it's always in precisely one state), which, by definition, must be the wrong one. Hence, incompleteness during design necessarily results in incorrectness. An "extra" EW state may map to just one IS state (making it consistently and predictably incorrect), or it may map probabilistically to a range of IS states (making it ambiguous as well as incorrect).

Suppose instead that there was an "extra" state in the IS (x_6, implying meaninglessness). As above, a non-zero probability of the IS ever getting into this state means that it is the same as that state not existing. However, if it is ever occupied then, again by definition, the EW must be in some other state, which also results in incorrectness. Furthermore, with the IS in a meaningless state, the EW cannot consistently be in the same state – if it were then it would no longer be meaningless because it would become perfectly synonymous with another meaningful (yet redundant) IS state.

So the magnitude of design errors – incompleteness and meaninglessness – can be assessed by their effect (if any) during operation. The way they are manifested is through ambiguity and incorrectness which, as shown, are characterised by the fidelity measure. Therefore fidelity is an appropriate general measure of semantic information quality.

With the semantic level of IQ quantified in this way, it is now possible to tackle the pragmatic level, which is concerned with the use of information. First, we need to understand how information is "used" and how we can quantify that. For that, I turn to the analysis of decision-making and the value of information as formulated in the discipline of information economics.

5.3.4 Information Economics

The concept of "value" as a criterion, measure or goal in IQ research has often been poorly understood. This difficulty is explicitly acknowledged in the Semiotic Framework for IQ:

[V]aluable relates to the overall worth or importance of the data with respect to the use of that data. Of all the quality criteria listed, this is the most problematic in that it has inter-dependencies with all of the other quality criteria. That is, data which is not highly rated with respect to other criteria (e.g. not complete, not reliable) will necessarily be less valuable as a result. However, in accord with most information quality researchers, we believe that a comprehensive understanding of quality requires the inclusion of such a criterion, sometimes termed value-added or value.

…

In essence, the quality criterion valuable acts as a generic place-holder for those aspects of quality specific to a given application rather than universally applicable. Thus, other than replacing the generic quality criterion with the appropriate domain-specific terms for each individual application, the only other option is its inclusion despite the resulting inter-dependencies. The problems and significance of this particular quality criterion has not, to our knowledge, previously been acknowledged in the literature. (Price and Shanks 2005a, p222)

The distinction between value and quality is an important but difficult one. In particular, whether value is a quality characteristic or quality is a determinant of value is difficult to reconcile. In both cases, it's clear that some sort of comparison is being made. Here, it is proposed that the *comparator* is what is different: a quality assessment is made using an "ideal" as the yardstick, whereas a value assessment compares against (potential) alternatives.

Consider the example of something mundane, like ice-cream. When determining the quality of an ice-cream, a consumer may compare a range of characteristics of the ice-cream (flavour, texture, quantity, safety etc) against an "ideal" ice-cream, which is conceivable but does not exist. This ideal will vary from person to person, and perhaps even across time. The quality assessment of an ice-cream is expressed in terms of shortcomings against these criteria, perhaps as a star-rating, a percentage score or text description (review).

By contrast, a value assessment involves ranking the ice-cream against candidate alternatives in terms of how much *utility* (satisfaction, pleasure) the consumer derives from it. (In the formalism of Utility Theory, this ranking is called a partially-ordered preference function.) The specific reasons why the ice-cream is ranked in its position are not considered, and they too vary from person to person and over time. This ranking process can be operationalised experimentally by using observations of people's behaviour to reveal their preferences.

These two kinds of assessments have their own strengths and weaknesses, and may be useful in different situations. Both have subjective elements, such as the weighting given to quality criteria or the specific ranking of an alternative. Both have objective elements too: the quantity of ice-cream, for example, can be measured objectively. So too can people's preference for a vanilla over cabbage flavoured ice-cream.

Quality assessments have two advantages over value assessments. Firstly, quality assessments can be used to gain insights into which aspects or characteristics consumers care about the most. With ice-cream, for instance, by explicitly setting up a list of criteria and evaluating each in turn, it is easier to see how changes could be made to improve the product. In this sense, quality is more important for design activities where trade-offs must be made.

The second advantage is for situations when no comparable alternative is available. A nation's tax system is, in general, not discretionary and taxpayers do not have the option of using another one. Assessing the value of such a system (and hence, their preferences for tax systems) is going to be

fraught when no such choice is possible. (Governments, which do have such a choice, may examine the value of their tax system.)

Value assessments, however, do have advantages of their own. Primarily, value assessments – being more abstract – allow comparisons between unalike things. A child faced with "You can have two more rides on the merry-go-round or an ice-cream" will determine which is more valuable to them. For business, the most powerful aspect of this is the comparison with cash amounts, which allows *pricing* of goods and services. If someone ranks their preferences for a $10 note, an ice-cream and a $5 note in that order, we can conclude they value the ice-cream at between $5 and $10. Using smaller and smaller amounts of cash, we could drill down to a specific price (cash equivalent) at which they are indifferent. This quantity of cash is their price[7].

These different valuations by buyers and sellers is what drives transactions, and hence markets, and explains the concept of "value-added". If the seller is willing to accept $6 for the ice-cream while the buyer is willing to pay $9, this $3 gap is called the "consumer surplus". In general, a proportion of this $3 will go to both parties; the specific amount depends on market conditions such as competition and regulation.

I use this distinction to address the pragmatic quality of information and the value of information.

While the Semiotic Framework conceives of the pragmatic level as incorporating quality criteria such as *useful*, *relevant* and *valuable*, it does not prescribe a quantitative model for measurement of these constructs. Instead, the authors propose the use of consumer-centric, context-specific instruments such as surveys to understand this dimension (Price et al. 2008). In contrast to the semantic level, the pragmatic one is necessarily subjective.

This approach is not sufficient for the purpose of this framework. During the practitioner interviews, practitioners reported that they understood where "the points of pain" were in their systems and – broadly – how to fix them. What they said they required was a solid business case that credibly stated, in particular, the benefits side of the equation. They were adamant that organisational funding processes dictated that this had to be quantitative, expressed in financial terms and commensurate with other capital projects. In essence, this is a call for a *de*-contextualisation of the information systems and associated customer processes, where all factors are reduced to future expected cash flows.

As a measurement approach, pricing like this could be characterised as subjective-quantitative. It is quantitative, in the sense that real ordinal values are used to describe the phenomenon and logic and mathematics drive the calculations. However, it is also subjective in that, at root, preferences are innate and ultimately cannot be explicated. In this way, prices differ from measurements of natural phenomena in science and engineering. For example, the price of a single share in a listed company may be subject to a huge amount of mathematical analysis and modelling yet it is determined by the aggregate opinions of thousands or millions of actors. This is not to say the prices are entirely arbitrary or set on a whim, but constitute a shared understanding.

Prices may ultimately be subjective, but if an organisation has an agreed valuation method for final outcomes, then intermediate prices may be derived objectively (in the sense that different people can arrive at the same answer). For example, suppose a retailer has a standard method for valuing stock in various locations (warehouses, retail shops etc). These prices may be set using a mixture of

[7] *Economic Theory describes two prices: Willing To Pay (WTP, or "buy price") and Willing To Accept (WTA, or "reservation price"). In general, WTP ≤ WTA, a phenomenon known as the endowment effect.*

estimated market prices from competitors and suppliers, precedent, custom and judgement. But given these (subjective) prices, it is possible to price objectively the costs and benefits of changes to the retailer's logistical system, such as changing shipping frequencies, trucking routes or warehouse capacity. It is in this sense that the quality of information can be objectively and quantitatively valued.

Information Economics, as broadly conceived, examines the value of information in widely different contexts. The starting point is that information is both costly and – potentially – beneficial. People can be observed behaving in ways that suggest they have preferences for different information for different tasks. Managers, for instance, will expend organisational resources on acquiring information in the belief that the benefits outweigh the costs.

Approaches to quantify and valuing information are incorporated into microeconomics, which deals with supply and demand, individual decision-making and Utility Theory. In the von Neumann and Morgenstern game-theoretic reformulation of neo-classical microeconomic theory (Neumann and Morgenstern 2004), very general assumptions are made about how rational people deal with uncertainty. Specifically, the Expected Utility Hypothesis assumes that (groups of) people, when faced with multiple possible outcomes, will assign a utility ("benefit") to each outcome and then weight each utility by the probability of its occurrence.

This "weighting" approach to risk can be described as a "rational gambler's perspective", in that it involves calculating the probabilities and pay-offs for possible outcomes. Indeed, it was in that context that it was first proposed by Daniel Bernoulli in 1738. The Expected Utility Hypothesis is a normative theory of behaviour and, while it stacks up quite well in practice (Lawrence 1999), more sophisticated descriptive theories take into account nuances of irrationality and cognitive biases and other deviations from this ideal. One such alternative is Prospect Theory (Kahneman and Tversky 1979).

Another key concept in Information Economics is the definitions of semantic and pragmatic information. These definitions correspond with those described in the Semiotic Framework for IQ, albeit in a more formal mathematical sense. In particular, a message or event contains *semantic information* if and only if it changes someone's *beliefs*; while a message or event contains *pragmatic* information if and only if it changes someone's *actions*.

Consider this explanation from an Information Economics perspective:

Pragmatic information *involves the application of the statistical [semantic] information[8]; it concerns the potential impact of the statistical information; it concerns the potential impact of the statistical information on choice and pay-off in a specific decision problem. This distinction separates nouns commonly associated with statistical attributes of information, such as coherence, format, and accuracy, from pragmatic attributes such as relevance, completeness, and timeliness. Statistical information affects what the individual knows; pragmatic information affects what the individual does. (Lawrence 1999, p5).*

It is clear that improving semantic information quality is necessary, but not sufficient, for improving pragmatic information quality. For example, if you know your friend's birth date, then finding out the precise hour of his birth (thus reducing ambiguity) does not have any bearing on your decision about when to send a birthday card. Further, improving pragmatic information quality is necessary, but not sufficient, for increasing information value.

[8] *This author refers to the concept of semantic information as "statistical information". It does not mean "statistical", in the sense of "data collected by statisticians" or similar.*

So knowing something "for its own sake" is not valuable. This is clearly a very narrow view of value and much of our information – especially entertainment, news and gossip – would not qualify as valuable by this definition[9]. However, information economists argue that information that does not result in a changed decision may still prove valuable if it reveals something to the decision-maker about their information sources (Lawrence 1999). So a rumour may not be immediately useful, but if it later proves correct, then the source may be seen as more credible and hence the decision-maker is more likely to act on future information from that source.

The elements of the information economic model (as used here) include:

- A set of states of nature, with a probability distribution over them.
- A set of outcomes, with pay-offs associated with each.
- A decision-maker, with a defined utility function (for risk-aversion and time-preferences).
- A set of options, from which the decision-maker can select.

If I can extend the semantic-level Ontological Model for IQ (meaning) to include these elements, then I have a natural extension for the pragmatic level (use) that lends itself to a quantitative *valuation*. The point of common contact is the state-based probabilistic view of nature and the distinction between semantic and pragmatic information quality.

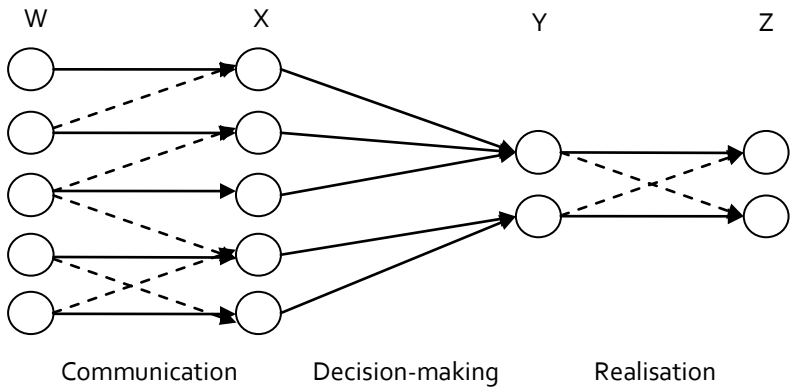

FIGURE 11 AUGMENTED ONTOLOGICAL MODEL

Here, I have extended the familiar Ontological Model to incorporate *decision-making*. It's represented here as a process that maps states of the external world (or its proxy, the IS) to an action. This function is the object of study in Decision Theory (and related disciplines like Game Theory), in a general sense. Theories from Management Science and Operations Research play a role in developing the particular decision functions used in practice. The precise means for how such functions are designed and implemented is not the concern of this study. It suffices to say that such functions exist, are used as a matter of course and can be specified.

As before, there is a state-space σ defining a set of states of interest. The external world state is described as a probability distribution, W, over σ while the IS is a probability distribution X over σ as well. These two random variables are related by the transition matrix T, representing the communication process.

[9] *Indeed, some branches of economics examine information value from the perspective of social signalling and conceive of information as a positional good, whereby being seen as someone "in the know" confers status.*

I now augment this with another state-space, π, describing a set of possible *actions*. Depending on the task, this set could comprise "yes" and "no" (at a minimum), or there may be a hundred actions, each indicating a price (from $1 to $100) that the decision-maker (or "DM") may bid at an auction. The only requirement is that the options are mutually exclusive – one and only one eventuates.

Two probability distributions are defined over this state-space. The first, $Y = [y_1, y_2, \ldots, y_n]$, is the action chosen by the DM. The second, $Z = [z_1, z_2, \ldots, z_n]$, has the same cardinality as Y. Its interpretation is, informally, the optimal action; that is, the action that the DM would have preferred, given full and perfect information. In the same way that W represents the "correct" state of the world, Z represents the "correct" action.

The combination of Y and Z – the realisation process – defines the possible outcomes from the decision task. The realisation matrix, R, enumerates the possible outcomes and assigns a probability to each occurrence, conditional on the decision taken. From the DM's point of view, Y and Z should be in constant agreement as this means the DM is making the "correct" decision each time. This case is described by the identity matrix.

$$\begin{array}{c} \\ Y \end{array}\begin{array}{c} Z \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ Y \end{array}\begin{array}{c} Z \\ \begin{pmatrix} 0.8 & 0.2 & 0 \\ 0.1 & 0.7 & 0.2 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

FIGURE 12 (A) PERFECT AND (B) IMPERFECT REALISATION

Note that the matrices in Figure 12 are expressed in terms of conditional probabilities $p(Z=z_i|Y=y_j)$ and in this form is a row-stochastic Markov matrix (Lawrence 1999). When expressed as joint probabilities $p(Z=z_i, Y=y_i)$, such matrices are referred to as "confusion tables" in the Decision Theory and Machine Learning literature. In any case, for a given $Y = [y_1, y_2, \ldots, y_n]$, I can derive the joint distribution from the conditional by multiplying the i^{th} row by y_i.

In these examples, the first realisation process is "perfect", in the sense that whenever y_1 or y_2 or y_3 is chosen, the optimal decision is z_1, z_2 or z_3, respectively. The second matrix describes an imperfect realisation process: sometimes, the DM makes sub-optimal choices. For example, when the DM chooses y_1, 20% of the time it eventuates that z_2 is the optimal choice. The DM would have been better off choosing y_2.

Analogously to T, the transition matrix that characterises the communication process at the semantic level, R is a transition matrix that characterises the realisation process. I can define measures on this matrix that evaluate how well the DM is performing. Indeed, that is precisely what well-known statistical measures such as χ^2 (the chi-square statistic) does, or ρ (Pearson's rho for correlation) for that matter (Neuman 2000). In addition, for the binary case, the usual Information Retrieval measures apply (precision and recall, also known as sensitivity and specificity in other contexts) as do more sophisticated approaches, like lift charts and ROC analysis, in Machine Learning and data mining (Hand 1997)

These intuitively appealing measures only apply in the binary case. A more general approach must allow for multiple action possibilities. In this situation, I can use Information Theory to quantitatively score classifier performance (Kononenko and Bratko 1991). The idea is to look at the amount of uncertainty about Z, and what Y tells us about Z, using their mutual information:

$$I(Y;Z) = H(Z) - H(Z|Y)$$

This can form the basis of a performance metric – the average information score – that can be used to compare classifier performance on the same decision task (for example, using different algorithms) and also *between* decision tasks, since it takes into account how "hard" different tasks are. For example, detecting pregnancy in a maternity hospital is a pretty trivial decision task: simply labelling all female patients "pregnant" will (presumably) get you to at least 95%. Detecting pregnancy in the general population, however, is not so simple. The use of information-theoretic measures takes into account these "prior probabilities".

While this does extend the Ontological Model into the pragmatic level in a way that allows quantification, it does not meet the goal of providing a valuation measure. For that, I must add one additional element: pay-offs.

The pay-offs are defined as changes to the DM's utility (or net satisfaction), expressed as equivalent cash amounts. Note the definition of the pay-offs as cash-equivalents of a change in utility, rather than as cash, to take into account the time preferences (discount rates) and attitudes to risk. However, it is not clear how to operationalise the "change in utility" measure, as it is an abstract and essentially introspective construct. This is where the inherently subjective nature of valuation comes in to play; it may be possible to derive these cash-equivalent amounts analytically or experimentally in some cases, but generally the quanta are a matter of judgement.

For the purposes of this model, the pay-offs are expressed as costs "relative to perfection" (that is, really penalties), in keeping with the established custom in the Machine Learning and Information Economics literature (Lawrence 1999). This is because using the case of perfect information results in zero cost whereas imperfect information results in a positive cost. It should be noted some research indicates that practitioners prefer to think of decisions in terms of costs and benefits (Chauchat et al. 2001; Drummond and Holte 2006). In practice, rather than zeroing the scale at perfect information, it may be more palatable to zero it at the current level of performance, so that changes can be assessed as costs or benefits, depending on direction. In any case, since the two methods are interchangeable, I stick with the cost-based system.

$$\Pi = \begin{pmatrix} 0 & 5 & 2 \\ 2 & 0 & 11 \\ 1 & 3 & 0 \end{pmatrix}$$

FIGURE 13 PAY-OFF MATRIX USING THE COST-BASED APPROACH. ALL UNITS ARE DOLLARS.

In this example, a "correct" decision attracts a \$0 penalty. However, deciding on the third option ($Y=y_3$) when the second was optimal ($Z=z_2$) results in a penalty of \$3 ($\Pi_{3,2}$). Inspecting the columns, we see that $Y=y_2$ is the most "risky" option – penalties for mistakes range up to \$11. By contrast, $Y=y_3$ ranges from 0 to 3. A rational DM would take this into account when choosing between y_2 and option y_3. Indeed, the advice from the machine learning community is to always use pay-offs to evaluate performance, where they're available and applicable (Hand 1997).

In order to compute the expected cost of imperfect information (or, equivalently, the expected value of perfect information), I invoke the Expected Utility Hypothesis as follows. For Y = [0.2, 0.5, 0.3], I derive the joint probability distribution from R, the realisation matrix, and do entry-wise multiplication with the pay-off matrix, Π:

$$V = \begin{pmatrix} 0.16 & 0.04 & 0 \\ 0.05 & 0.35 & 0.1 \\ 0 & 0 & 0.3 \end{pmatrix} \cdot \begin{pmatrix} 0 & 5 & 2 \\ 2 & 0 & 11 \\ 1 & 3 & 0 \end{pmatrix}$$

$$
\begin{aligned}
V = {} & (0.16 * 0) + (0.04 * 5) + (0 * 2) + \\
& (0.05 * 2) + (0.35 * 0) + (0.1 * 11) + \\
& (0 * 1) + (0 * 3) + (0.3 * 0) \\
= {} & 0.2 + 0.1 + 1.1 \\
= {} & 1.4
\end{aligned}
$$

Hence, in this case, the expected cost of the imperfect realisation process is $1.40. A rational DM would never spend more than this amount to improve their decision-making, since there is no way the cost could be recovered. In this sense, the expected value of perfect information represents a ceiling on the price a DM would pay for improvements.

I now consider the elements that go into making up this cost. From the definitions of the realisation process and the pay-offs, I note that the costs are accrued when the i^{th} action selected from Y is not optimal: Z = j ≠ i. I describe three sources, using the illustrative example of a mortgage approval process within a bank.

In this scenario, the semantic space σ, is very large and is the space of all possible situations applicants might be in, including age, income bracket, post code, employment status and so on. The pragmatic space π is the space of all possible actions the bank could take – y_1 is "accept application" and y_2 is "refuse application". This second state-space is much smaller than the first.

From the bank's perspective, W is a probability distribution, since they don't know the true situation of each applicant, but they do have information *a priori*. The quality of semantic information characterises how well, on average, the distribution X informs the bank about distribution W. When the external world state and the IS state differ, this is called an *error*.

Y is the action the bank takes for each applicant and Z is the optimal action, given hindsight, for that applicant. The bank uses a decision function, D, to map applicants to actions. For any applicant's situation (drawn from the semantic space σ), the bank will select an action from π. Naturally the bank will wish to ensure that its chosen action is as close to optimal as possible. When this does not occur (for example, accepting the application when the optimal action would have been to reject it), this is called a *mistake*. In contrast to an error, which is a wrong belief about the external world, a mistake is wrong action. A mistake attracts a penalty to the bank, as specified in the pay-off matrix.

Under this model, there are three sources of mistakes: the *quality of information*, the *decision function* and *residual domain uncertainty*. The first arises when an *error* results in a *mistake*. That is, a mis-characterisation of the EW by the IS (an incorrect mapping, in ontological terms) causes the decision-maker to select the wrong action. This is illustrated below.

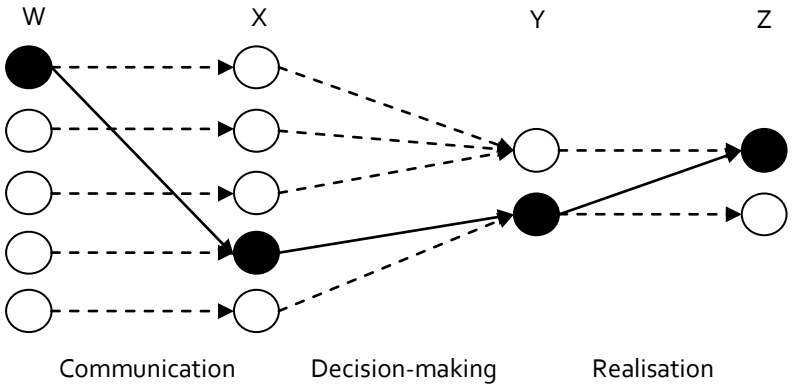

FIGURE 14 COSTLY INFORMATION QUALITY DEFECT

Here, an applicant should have been in the first state (w_1), but was mis-represented in the IS as being in the fourth state (garbled into x_4). As a result, the DM refused their application (mapped them into y_2). In fact, applicants in the first state should have been approved (z_1 was optimal). Not all garbling events in the communication process will result in a mistake. In this example, had the applicant been mis-represented as being in x_2 instead of x_1, the decision function would have correctly mapped them into y_1. In this instance, it would have been a semantic defect (error) but not a pragmatic defect (mistake).

The second source of mistakes is the decision function itself. Real decision-making processes will make mistakes even when presented with perfect information. This could be due to inherent limitations in the algorithms employed or defects with how it is implemented. Indeed, many researchers and practitioners involved in fields as diverse as Management Science, Operations Research and Computer Science focus entirely on improvements to decision-making where perfect information is often assumed.

The third and final source of mistakes is "residual domain uncertainty", which captures the notion that, despite all the information and algorithms in the world, some decision tasks are always subject to an unavoidable element of chance. This means there is a "ceiling" level of performance innate in a decision task, which cannot be bested by any amount of information quality improvements or algorithmic enhancements.

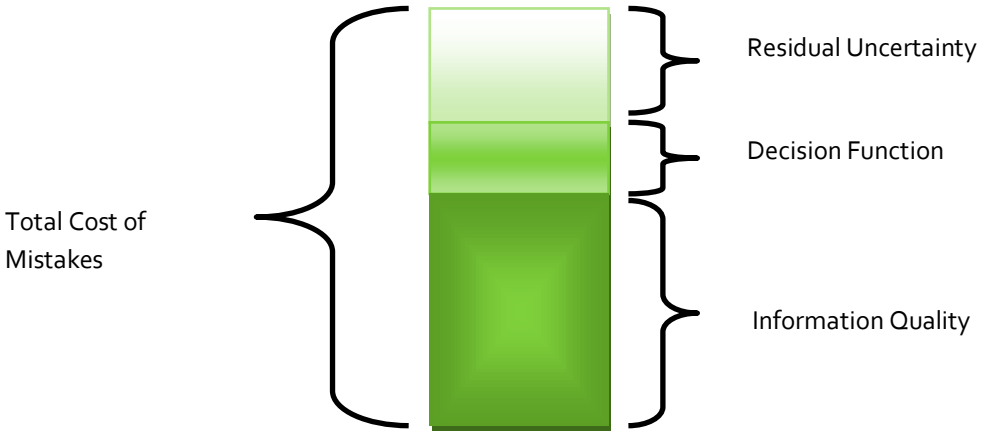

FIGURE 15 BREAKDOWN OF SOURCES OF COSTLY MISTAKES

While these sources of mistakes generate real costs that are borne by the organisation, I am here only interested in those arising from deficiencies with information quality. For example, spending money on fine-tuning the mortgage approval business rules – perhaps informed by benchmarking against comparable institutions – may be a worthwhile project. But this does not directly help practitioners formulate a business case for information quality improvement. As a consequence, I need to modify the model under development to only account for the cost of mistakes introduced by IQ, not all mistakes.

To do this, I introduce a new construct, Y^*, which substitutes for Z. Recall that Z is the "optimal action", perhaps taken with the wisdom of hindsight in the case of mortgage approvals. Given deficiencies in the business logic and the inherent vagaries of human behaviour, it is setting a very high bar. Instead, I define Y^* as the action that would have been taken had perfect information been used. Using a function notation D(·) to describe the decision-making process:

$Y = D(X)$ The actual action taken, using imperfect information.

$Y^* = D(W)$ The ideal action, using perfect information.

This more attainable comparator is used to compare the decision made with imperfect information (Y) with the decision made with perfect information (Y^*), using the same real-world decision function. Thus, it only measures shortcomings in IQ, not algorithms. If Y^* and Z are very different from each other, it suggests a serious problem with the decision function. (That is, the decision function consistently cannot produce the optimal answer even when presented with perfect information.)

This distinction between Z and Y^* also addresses the pathological situation of an error *improving* the decision, due to a deficiency in the decision function or the vagaries of human nature. Any complete economic analysis of error removal must include the costs introduced by new mistakes introduced when errors are removed. An information system that degrades in performance as errors are removed may seem entirely anomalous and unlikely. However, in situations where source systems are known to be compromised but fixing the problems there is not possible, it may be expected that the decision function is instead "tweaked" to address these issues. In such cases, where the "downstream" applications compensate for "upstream" data problems, fixing the source data may lead to a deterioration in decision performance. By using Y^* as the comparator, and not Z, these types of deficiencies are excluded from the analysis.

Recall that V was the expected pay-off using R (the mapping between Y and Z) and the pay-off matrix, Π. I now define R^* as the realisation process mapping Y and Y^*, so V^* is the product of R^* and Π. Whereas V measured the value of removing all *mistakes*, V^* measures the value of removing all *errors*, so $V \geq V^*$, with equality if and only if all mistakes are due to errors.

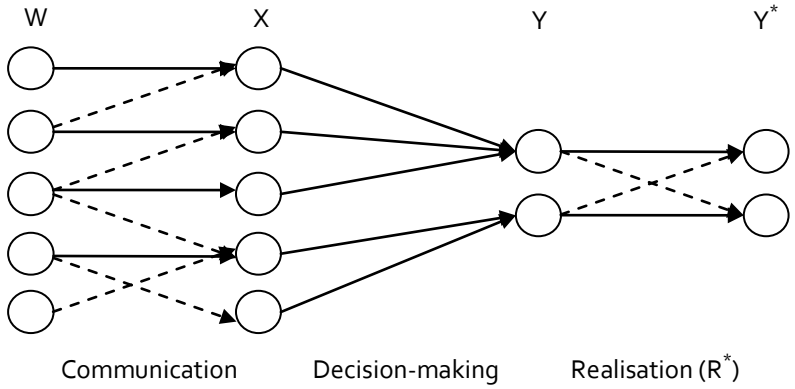

FIGURE 16 REVISED AUGMENTED ONTOLOGICAL MODEL

Of course, as discussed, this is not likely to be the case for real systems: mistakes will creep in even with perfect information. Similarly, since not all errors will lead to mistakes it is possible to have an imperfect IS that performs as well as one with perfect information.

Based on this discussion, I define the concept of *actionability*. An error is actionable if and only if it results in a different action being taken. In terms of the above model, an erroneous X (doesn't correspond to W) is actionable if and only if it results in a mistaken Y (doesn't correspond to Y^*). It follows that inactionable errors (or changes) must be worthless.

For deterministic decision functions, a change of state (through error or correction) either will always change a decision, or it never will. It is not probabilistic. However, I can generalise this concept of actionability to situations where the change applies to a class of states and hence may be characterised as probabilistic, in the sense that at the time of the change of state, there is uncertainty as to whether the decision will also change or not. A measure of this uncertainty about the impact of a change can be used to operationalise the general concept of *relevance*.

The concept of relevance is defined in a number of fields dealing with information and decisions, especially law and economics. In both cases, the specific test for relevance is whether or not it has the potential (or tendency) to induce changes in the probabilities. From a legal perspective, "relevance" is considered part of the rules of evidence. For example, section 55(1) of the *Uniform Evidence Act 1995* (Cth) defines relevance:

The evidence that is relevant in a proceeding is evidence that, if it were accepted, could rationally affect (directly or indirectly) the assessment of the probability of the existence of a fact in issue in the proceeding.

In economics, John Maynard Keynes formalised this notion. He explicitly defined a fact or piece of evidence as being *irrelevant* to an assertion if and only if the probability of x is the same with and without the evidence (Keynes 1923). More formally, for an existing set of knowledge k, new evidence e is irrelevant to assertion x if and only if $P(x|k) = P(x|k\&e)$.

In terms of the Semiotic Framework, relevance is identified as a pragmatic level criterion. Since improving the semantic quality of irrelevant information cannot – by definition – have any bearing or impact on decision-making, it is must be worthless. However, it may not be possible to know at the outset whether an improvement will change a decision. A probabilistic measure of relevance, as suggested above, captures the inherently statistical nature of such relations between information and decisions.

5.4 COMPONENTS

This section outlines the components of the framework, describing them conceptually and mathematically. This allows us to analyse customer information quality with a view to quantifying and valuing the impact of interventions designed to improve IQ.

Here, I restrict my analysis to situations involving data-driven customer processes: a large number of customers are considered in turn and are partitioned into a small number of subsets for differentiated treatment based on the application of business rules to their attributes. This could apply to common data-driven decision-making functions such as direct marketing, credit scoring, loan approvals, fraud detection, segmentation, churn prediction and other classification and prediction tasks.

Customer attributes include demographics (date of birth, gender, marital status, location), socio-economic details (education, employment, income, assets), product history (details of which products were purchased or used), contact history (inbound or outbound contacts) or third party "overlays" such as credit assessments, legal status (for example, bankruptcy or immigration) or other market transactions.

5.4.1 COMMUNICATION

The communication process models how well the external world is represented by the internal IS. In this case, I assume a customer database, comprised of C customers. Conceptually, each customer is represented by a row in a table (record) where each attributes is represented by a column. More formally, each customer has an external world individual state, denoted c_e, for the e^{th} customer. This state can be decomposed into a attributes, such that the customer semantic space, σ, is given by:

$$\sigma = A_1 \times A_2 \times A_3 \times \ldots \times A_a$$

While some attributes have a small range of possible values (for example, gender may be just {*male, female*}), others may have be very large. For continuous valued attributes like income, I assume "binning" (conversion into a discrete attribute through the use of appropriately sized intervals).

To give a sense of the dimensions involved, this framework anticipates the number of attributes (columns) to be in the order of ten to fifty, while the number of customers (rows) is in the thousands to hundreds of thousands.

The quality of the representation of the external world – the semantic information quality – can be quantified by asking "How much uncertainty is removed by observing the IS?" I can put this on a mathematical footing by defining a probability distribution, W, over σ that describes the external world. Similarly, I define another probability distribution, X, over σ that describes the IS. Applying Information Theory yields the previously described *fidelity measure*, ϕ:

$$\varphi = 1 - \frac{H(W|X)}{H(W)}$$

This expression – ranging from 0% to 100% - captures the amount of information *communicated* about the external world to the internal IS representation of it. It reaches 100% when H(W|X) = 0; that is, the uncertainty in the external world state given the IS state is zero (knowing X is always sufficient for knowing W). It reaches 0% when H(W|X) = H(W), which implies that knowing X always tells us nothing about W.

By extension, I can define fidelity at attribute level too, here for the e^{th} attribute:

$$\varphi_e = 1 - \frac{H(W_e|X_e)}{H(W_e)}$$

I use W_e (and X_e) to denote the value of W (and X) on attribute A_e. Note that in general these φ_e will *not* all add up to φ as there is "side information" amongst the attributes. For example, knowing something about a person's education will, on average, reduce my uncertainty about their income. For the case where all the attributes are perfectly statistically independent, the φ_e will add to φ.

The interpretation of the fidelity measures (the normalised mutual information score) is of the proportion of the uncertainty (in bits) reduced. In this sense, the measure takes into account how "hard" it is to "guess" the right answer and in doing so, makes comparisons between different attributes more meaningful than simple "accuracy" (percentage correct or P(W=i,X=i), for all i).

For example, consider the two attributes A_1 and A_2 and their respective channel matrices, C_1 and C_2, expressed as joint rather than conditional probabilities:

$$C_1 = \begin{bmatrix} 0.45 & 0.05 \\ 0.1 & 0.4 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 0.05 & 0.1 \\ 0.05 & 0.8 \end{bmatrix}$$

Note that both attributes are binary valued, but that while the first is balanced (a 55%-45% split) the second is quite unbalanced (a 10%-90% split). In more concrete terms, A_1 might be *gender* (with, say, 55% female) while A_2 might be a *deceased* flag (with 90% alive).

Both attributes are 85% correct; that is, 15% of customers are in error (the diagonal adds to 0.85). However, their fidelity measures are quite different, with $\varphi_1 = 40\%$ while $\varphi_2 = 12\%$. The interpretation is that it is more difficult (ie easier to be wrong) to represent *gender* than *deceased* status. In fact, given that 90% of customers in this group are alive, simply labelling all customers "not deceased" will achieve 90% correctness. Therefore, the *deceased* attribute is performing quite poorly, as reflected in its comparatively low fidelity.

From an information-theoretic perspective, fidelity measures the incremental amount of uncertainty removed, given the initial amount of uncertainty. These two attributes have different amounts of uncertainty in the first place (*gender* has 0.99 bits of uncertainty while *deceased* has only 0.47 bits), so the removal of uncertainty has a different impact on the score.

By taking into account the inherent "difficulty" (statistically-speaking) of capturing some attribute, I can make a better assessment of how well the information system is representing that attribute. The fidelity measure does this better than correctness or probability of error.

5.4.2 DECISION-MAKING

This framework is intended to be applied to a range of customer processes that rely on customer information. These data-driven decisions are made by considering each customer, one at a time, and applying business rules or logic to attributes of the customer to make a determination about how the organisation will treat that customer in future. Whether it's approving a loan, making a special sales offer or assigning them to a demographic marketing segment, each customer is mapped onto one action (treatment, offer, label etc) from a small set of possible actions.

Depending on the context, this process can be described as classification, segmentation, partitioning, prediction or selection. The function that performs this task could be implemented

formally, perhaps as a decision-tree (with a hierarchy of IF ... THEN ... ELSE logic) or regression model, or informally, as with a flow chart or customer interaction script.

The study and analysis of algorithms and techniques for making the decision is not within the scope of IQ improvements. Instead, for analytical purposes, I only require that such a function for mapping customers is *deterministic*. This means that the same action is produced for an identical input customer, each time. The decision function should not change (or "learn") over time, there is no discretion, judgement or randomness in the function and the order in which customers are presented should not matter. That is, if the list of customers is randomised and then re-presented to the decision function, each customer should be mapped to exactly the same action as in the first run.

While the determinism requirement excludes judgement-based customer processes[10], it still covers a large number of situations that rely on discretionless customer processes, such as found in direct marketing and credit management.

I define a probability distribution, Y^* over the set of possible actions, π; these are the actions produced by the decision function, D, if perfect information is presented to it. Another probability distribution over π, Y, is the actual decision produced by D if given the imperfect information in the information system. I express this as:

$$Y^* = D(W)$$
$$Y = D(X)$$

I can now ask "how much information is needed to make a decision, on average?" The answer is: "enough to remove the initial uncertainty in which action will be taken (indecision)". Suppose that a decision task requires assigning customers into two "buckets": the first will receive a special offer, the second will not. A decision-tree is built that uses customer attributes to make the determination. This function, D, assigns 30% of customers to receiving the offer, while 70% won't. The indecision in this task is given by:

$$H(Y) = -0.3 \log_2 0.3 - 0.7 \log_2 0.7$$
$$= 0.88\ bits$$

This means that, prior to the decision function being applied, there is 0.88 bits of uncertainty about what the decision will be. Afterwards, the uncertainty is 0 since, by definition, D is a deterministic function:

$$H(Y|X) = 0$$

So, the mutual information between Y and X is H(Y). This can be seen from the formula for mutual information:

$$I(X;Y) = H(Y) - H(Y|X)$$

Intuitively, there must be less uncertainty on average about what action to take than the there is about the state of the customer. In other words, there's less information in the decision than the description. Therefore, H(Y) ≤ H(X), since to suppose otherwise would entail creating information "out of thin air" violating the Data Processing Theorem (Cover and Thomas 2005).

[10] *Of course, this is not to say that judgement and discretion were not used in formulating the business rules and decision parameters in the first place.*

That is the case for the complete customer state, X. I can examine the situation where just one attribute is revealed and ask "How does knowing attribute X_e change our indecision?" If that knowledge has precisely no effect on the decision, then I can say (following Keynes) that that attribute is *irrelevant*. In general, I'd expect it to have some effect on the indecision, otherwise it would not be included in the decision function. I define the measure *influence* to describe the effect of the e^{th} attribute on the final decision:

$$I_e = 1 - \frac{H(Y|X_e)}{H(Y)}$$

It is the normalised mutual information between the decision Y and the e^{th} attribute, X. It ranges from 0% to 100%. The former occurs when $H(Y|X_e) = H(Y)$ (that is, telling us X_e does not change our assessment of the decision) while the latter occurs when $H(Y|X_e) = 0$ (that is, tell us X_e removes all uncertainty about the decision ie X_e completely "drives" the decision).

Consider the direct marketing example above. Initially, any given customer has a 30% chance of being selected for the special offer. Suppose that I am told that the customer is female and no other details. Given the way the decision function works, the odds now shift to a 35% chance of making the offer (and hence a 65% of not getting the offer). This suggests that *gender* has a modest bearing on the decision. If, instead, the *deceased* attribute showed that a customer was deceased, then the probability of making the offer drops from 30% to 0%. This suggests that deceased has a powerful effect on the decision – at least for those 10% of cases where the customer is marked "deceased". The influence measure formalises this idea.

By definition, the mutual information is symmetric:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X) \end{aligned}$$

So the influence measure can be defined in terms of uncertainty about the attribute value given the decision:

$$\begin{aligned} I_e &= 1 - \frac{H(Y|X_e)}{H(Y)} \\ &= 1 - \frac{H(X_e|Y)}{H(X_e)} \end{aligned}$$

This bi-directionality can be illustrated with the direct marketing example. Initially, we have that any given customer has a 10% chance of being deceased. If we find out that a particular customer made it onto the special offer list, our assessment of this unfortunate status changes to 0%. This means that knowing something about the decision tells us something about the customer attributes. In fact, the decision contains precisely as much information about the attribute as the attribute contains about the decision.

Either way, the individual influence scores for each attribute will not add to 100%; this is because of redundancy in "side information" between attributes. If all attributes were statistically independent, then the influence scores would add to unity.

Influence measures the relationship between customer attributes and decisions. As such, it is characterised by the decision-making function itself and not the relationship between the IS and the

external world. In other words, an influential attribute will be influential regardless of its semantic correctness.

It's also worth sounding a note of caution about attribute influence on decision outcomes. Similarly to correlation, influence does not imply causation. For example, a segmentation task may make extensive use of the *postcode* attribute, resulting in a high influence score. The *suburb* attribute would have a very similar score (since it is highly correlated to *postcode*), yet is never used by the decision function. Hence, any changes to *suburb* will not result in changes in the decision – they are not causally related.

The influence measure, defined here as the normalised mutual information between the decision and the attribute, quantifies the degree of relevance the attribute has on the decision.

5.4.3 IMPACT

The impact of customer IQ on customer processes lies in the decisions made. At the semantic level, IQ deficiencies result in misrepresentation of the external world (error ie $W \neq X$). Pragmatically, this is only a deficiency if this error results in a different decision than would have been made with perfect IQ (mistake ie $Y \neq Y^*$). Whether or not an error will become a mistake depends on how the decision function uses the information to arrive at a decision.

Formally, the goal is to get Y and Y^* as close to agreement as economic. The straightforward measures of agreement rely on comparing rates of false positives and false negatives: sensitivity/specificity, recall/precision, lift and ROC analysis, depending on the domain. As measures, they suffer from two drawbacks. Firstly, they only work in cases with binary "yes/no" ("approve/reject" or "offer/no offer") decisions. They do not scale well to situations involving more than two actions (eg "platinum/gold/standard/reject").

The second issue is that they do not address "prior probabilities". Akin to the discussion above regarding the inherent "difficulty" of stating different customer attributes, it is more difficult to do well in a situation where the possible decisions are split 50%-50% than one that is 99%-1%. With the former, there is more uncertainty (or, here, *indecision*). For example, a fraud detection application that simply reports "no fraud" in all cases will be 99% correct – if the underlying rate of fraud is 1%.

One response is to quantify how close Y^* and Y are by using an information-theoretic measure, the mutual information $I(Y^*;Y)$. The argument is that if the actual decisions made (Y) are a good predictor of the ones made under perfect information (Y^*) then I can be confident that the decision function is operating close to the ideal. This approach takes into account non-binary decisions and deals with the "prior probability" issue.

However, if information about the cost structures of mistakes is available, this should be used. Such information – expressed as the pay-off matrix Π – allows us to describe the impact of IQ on the process in financial terms. This has two advantages. Firstly, it allows us to compare IQ impact across different processes in a fair way. Secondly, it allows IQ impact to be assessed against a range of other costs borne by the organisation.

Determining the pay-off matrices for customer processes is a non-trivial task. Theoretically, the goal is to express the costs in terms of expected utility. This can be expressed in terms of the cash equivalent, where I assume a risk-neutral decision maker (and hence a constant marginal utility of wealth). Following accepted practice, the Net Present Value (NPV) is a reasonable proxy. As found in the practitioner interviews (Chapter 3), it is familiar and widely used in large organisations, including by IS managers. The discount rate should be chosen to comply with the organisation's financial

norms and standards. If necessary, more sophisticated variations could be employed. For example, the Weighted Average Cost of Capital (WACC) could be used to take into account the difference costs associated with funding capital with equity versus debt.

To build up the pay-off matrix for a customer process, a number of objective and subjective factors must be brought together. Details of the true cost of, say, mistakenly issuing a customer with a "gold" (premium) credit card when they should have been given a standard one will, in general, depend on objective factors including:

- the number of customers processed,
- the frequency with which the process is run,
- the time horizon for the process,
- the discount rate,
- fixed costs associated with operating the process.

However, the major factor is the subjective value placed on the mistake by the organisation. This cost includes lost revenue opportunities, additional financial and non-financial risks and damage to reputation and goodwill. In practical terms, this assessment is likely to be made by a sufficiently senior manager.

A properly constructed pay-off matrix for a customer process allows managers to understand the magnitude of the cost associated with each kind of mistake. The second part of the equation is the frequency with which these mistakes occur. This is captured by the realisation matrix R^*, which is the joint probability distribution between Y and Y^*.

Consider the example of a mortgage process for a financial services organisation. Suppose there are three possible decisions to make: *approve*, *partner* and *reject*. *Approve* means the mortgage is granted, *partner* means the applicant is referred to a partner organisation specialising in "low-doc" or "sub-prime" loans and *reject* is an outright rejection of the application.

After analysis of the objective and subjective costs, the organisation has the following pay-off matrix (expressed as total future expected costs per customer, discounted to current dollars):

$$\Pi = \begin{bmatrix} 0 & 1500 & 5000 \\ 500 & 0 & 1000 \\ 1500 & 500 & 0 \end{bmatrix}$$

Hence, approving an applicant that should have been rejected incurs the largest cost of $5000. By contrast, rejecting an applicant that should have been approved incurs a more modest cost of $1500, presumably in lost fees and other income.

The realisation of the process is given by the matrix R^*:

$$R^* = \begin{bmatrix} 0.2 & 0.07 & 0.01 \\ 0.06 & 0.3 & 0.04 \\ 0.03 & 0.04 & 0.25 \end{bmatrix}$$

This tells us that the organisation makes the right decision (or, at least, the one that would have been made with perfect information) 75% of the time (0.2+0.3+0.25). For the other 25% of customers, some mistakes will be made.

Using the Expected Utility criterion, the expected cost (per customer) is given by the scalar product of these two matrices:

$$V^* = 0.2(0) + 0.07(1500) + 0.01(5000) + 0.06(500) + 0.3(0) + 0.04(1000) + 0.03(1500) + 0.04(500) + 0.25(0)$$
$$= 0 + 105 + 50 + 30 + 0 + 40 + 45 + 20$$
$$= \$290$$

Note that in this example, the cost of mistakenly accepting applications that should have been referred to the partner constitutes the largest cost element ($105), more than double the cost of the next biggest mistake. This is despite having a moderate penalty ($1500) and reflects the relatively high frequency of its occurrence.

The interpretation of the amount of $290 is that it is the maximum amount a rational decision-maker would pay to acquire perfect information for a given customer. Multiplying this amount by the number of customers that go through a process (N) and the number of times each customer is processed (f) yields the *stake* of the process:

$$S = NfV^*$$
$$= NfR^* \cdot \Pi$$

In terms of correctly handling the time dimension, some adjustment should be made to the frequency to discount the time value of money. For example, for a marketing process with a static pool of 100,000 customers that is run six times a year over four years, the resulting 2.4 million "use-instances" should not all be given equal weight. Cash flows arising in the fourth year should be discounted using the organisation's internal discount rate. Further, a more detailed analysis should take into account the natural departure and arrival of customers over the four year period in question.

Stake is a measure of the total cost introduced into a customer process by using less-than-perfect customer information. As such, it provides an upper bound on the value of any improvements to customer information; a rational decision-maker would never pay more than the stake to improve a process since it would be impossible to ever recover more than the stake by just improving the quality of information.

5.4.4 INTERVENTIONS

The goal of the framework is not just to understand the costs imposed on an organisation by poor customer information quality. Rather, it is to appraise the financial implications of intervening in customer IQ with a view to improving it. Conceptually, the approach is to model these interventions as investments: an initial expenditure of resources followed by a (risky) financial return in the form of increased revenue or decreased costs.

Modelling IQ interventions in this way is useful for two reasons. Firstly, it allows for comparison between competing IQ initiatives, ensuring that the most valuable ones are selected. Secondly, it allows for the justification of IQ interventions with common organisational standards used by non-IQ initiatives so that such initiatives or projects are approved and funded.

An abstract approach to modelling IQ interventions is required, since IQ interventions can be radically different in mechanism (though not necessarily in effect). Typically, large organisations faced with customer IQ problems have a large range of options for tackling the problem, combining technical and managerial elements. Some examples include:

- change design or layout of customer forms,
- change business rules or definitions,
- change process flow,
- insert quality checks in processes,
- re-work ETL (extract, transform, load) scripts in databases,
- modify underlying data model to enforce integrity constraints,
- train staff involved in handling data,
- change management performance criteria or bonus/penalty structure,
- re-negotiate Service Level Agreements with data suppliers,
- audit, review and improve application code,
- implement specialised data quality software for consistency checking,
- rationalise the information systems in the organisation,
- employ extra staff or outsource manual data validation,
- purchase new data sources from data brokers (eg Dun and Bradstreet)m
- adopt new standards (eg XML) for handling data,
- introduce new management methodologies (eg Six Sigma) for quality assurance.

All of these activities are costly, time-consuming and risky. Plus, they will require significant support from a range of people within the organisation and other partners. The idea is to characterise these interventions in terms of their impact upon customer information quality, and hence, customer process outcomes. When these outcomes can be assigned a financial magnitude, I have the basis for comparing these very different candidate interventions on an objective basis.

Formally, I define an IQ intervention as a change in the IS state, X, into X′. In terms of the customer process model, the change is induced on one or more attributes of X. The decision function, D, is applied to the new IS state, X′, resulting in a new action, Y′. This is represented schematically as follows:

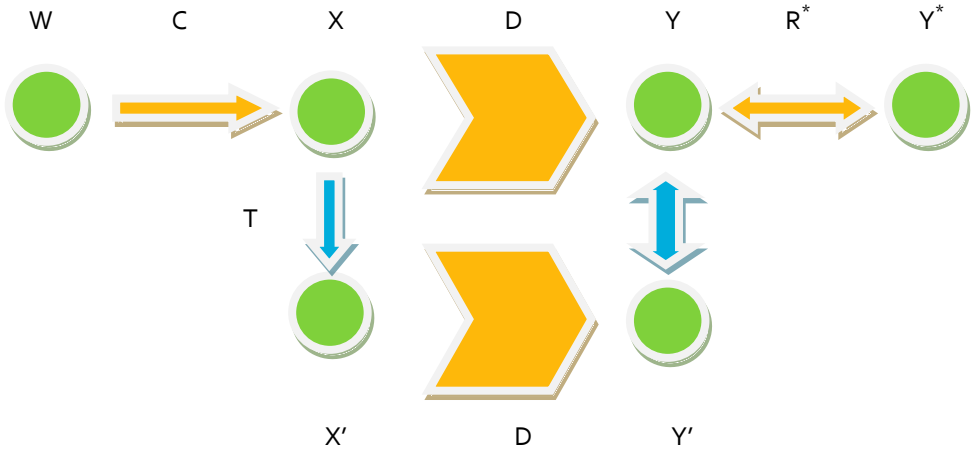

FIGURE 17 MODEL IF IQ INTERVENTION

The top half of this model proceeds as usual: the external world state, W, is communicated to the IS state, X, via process C. The decision function, D, is applied to the X resulting in the decision Y. The realisation process R^* relates this decision with the optimal decision, Y^*.

The bottom half introduces the intervention process, T. This process maps the IS state, X, into a new state X′. The intent is to achieve an IS state that is a better representation of the external world, W. When the regular decision function D is re-applied to this new state, a new decision, Y′, may result. If X′ is a better representation of W than X, then Y′ will be a better decision than Y.

In terms of comparing different interventions, it is natural to prefer the ones that will have a bigger impact. For any particular intervention, it could have a number of possible effects on an individual customer attribute:

- **No effect.** This means the intervention "agrees" with IS ie the new value is the same as the old value. Hence, no change in the decision and no change in value.
- **Inactionable correction.** A change of the IS state at the semantic level, but no change in the decision and hence value.
- **Actionable correction.** A change of the IS state at the semantic level and the pragmatic (decision) level, corresponding to a change in value.
- **Valuable correction.** An "actionable correction" resulting in a *positive* change in value.

This leaves open the possibility of an actionable correction with negative value – a situation that arises when an intervention actually makes a process perform worse. This may still be warranted, if overall the intervention has a net positive impact. This situation is analogous to a public inoculation programme, where the downside of some individuals' allergic reactions is offset by the community benefit to eliminating the disease.

However, in general, it is not possible to anticipate whether a particular change of state will be valuable or not, or, indeed, be actionable or not. This must be done by running the decision function over the changed state and seeing if it produces a different decision.

In this way, it is possible to value a particular intervention process, T, by examining its effect on the decision outcomes. An intervention that results in precisely the same set of decisions being made (ie $Y=Y^*$) is economically useless even if it corrected errors in the original communication process. This is the same distinction between semantic quality improvement and pragmatic quality improvement that motivated the earlier definition of actionability:

An Information Quality intervention is actionable if and only if the change in IS state results in a changed decision.

Clearly, any intervention that does not change the IS state (ie X=X') is not actionable and inactionable interventions are, by definition, worthless. From a design perspective, the goal is to efficiently identify interventions that are most likely to have an impact on decision-making, especially high-value decisions.

The value of an IQ intervention, T, is given by the difference in costs between the baseline position (no intervention) and the situation with the intervention. Using the Net Present Value model for the stake discussed above, I have the Yield:

$$V_T = S - S'$$

Where S is the baseline stake (cost of imperfect information) and S' is the stake under the intervention. So, for intervention T on the p^{th} process, I have:

$$\begin{aligned} V_{T,p} &= NfV_p^* - NfV_p^{*\prime} \\ &= R_p^* \cdot Nf\Pi_p - R_p^{*\prime} \cdot Nf\Pi_p \\ &= (R_p^* - R_p^{*\prime}) \cdot Nf\Pi_p \end{aligned}$$

In words, the value of any intervention for a particular process is the difference in the contingency matrices multiplied by a suitably-scaled pay-off matrix. However, the shared nature of information

resources within most organisations means that a change in a source information system, such as a data warehouse or master data repository, will impact across a number of "downstream" processes. The total value is simply the value of the intervention summed over each process within the scope of the analysis:

$$V_T = \sum_p (R_p^* - R_p^{*\prime}) \cdot Nf\Pi_p$$

This metric is the key "performance measure" for evaluating a proposed intervention, as it quantifies the financial impact of a proposed intervention across the processes of interest. The economically optimal intervention is the subset of candidate interventions that maximises this quantity. However, the intervention design task involves more than a simple brute-force search of all possible interventions. Estimating the model parameters is itself likely to be expensive and prone to error. Changes to the source information system will have different impacts on different processes, which themselves vary in their worth to the organisation. A rational approach to designing and evaluating information quality intervention requires not only that the final intervention is itself "good", but that the process that led to it is reasonably efficient and transparent.

A "map" describing at the outset the stages of design, estimation and evaluation of competing options will help give confidence in the final result. A value-led approach will prevent wasting time and resources on investigating futile, inconsequential or insignificant interventions, as well as offering guidance on how to modify or extend interventions to increase their value. Further, a transparent and objective method of appraising interventions may go some way to assuaging concerns about special-interests in joint-funding of projects.

5.5 USAGE

This section outlines the sequence of steps to design and evaluate an Information Quality intervention. It proceeds by undertaking a value-driven analysis of both the opportunities for improvements (technical capacity) and areas of greatest need (business requirements) using the mathematical metrics defined on the constructs described above.

The scenario targeted for adoption of the method has the following elements:

- a single information source (such as a data warehouse, operational datastore or similar),
- comprising a set of customer records, one for each customer,
- with each record having a number of attributes, including demographic and transactional data,
- used by a number of customer decision processes to partition, segment, classify, predict, allocate or label on a per-customer basis,
- where multiple candidate information quality improvement interventions are to be evaluated.

Note that the "single information source" does not have to be a single physical table or even database; a network of inter-linked information systems acting in concert to produce an abstract view of the customer suffices.

An example scenario could be a financial services firm where a customer database, augmented by demographic data from an external supplier, is used periodically by a set of marketing, campaign management, customer relationship management applications and credit scoring processes. The enterprise may be considering the relative merits of training contact centre staff on data entry, purchasing a data cleansing tool or integrating customer data from a subsidiary insurance business.

The goals of the method are:

- **Effectiveness**. The initiatives recommended by the method should be provably near-optimal in terms of value creation.
- **Efficiency**. The method should produce an analysis using a minimal amount of resources, including time and expertise.
- **Feasibility**. The computing requirements, availability of data, degree of theoretical understanding and disruption to IS operations should be within acceptable limits.
- **Transparency**. The constructs, metrics and steps employed should be intelligible and reasonable to IS professionals, and perceived as being unbiased and aligned to organisational-wide interests.

Rather than evaluating proposed interventions – that is, asking "What can we fix?" – the method proceeds by asking "What needs to be fixed?". Only then do we ask "What is the best way to fix it?" This approach mitigates developing good interventions for IQ problems of comparatively little economic significance.

The method has two phases. First, it starts with a wide scope of possible problem areas and narrows it through successive iterations of data collection and analysis using the performance metrics. (See Figure 18 below.). Second, candidate interventions are evaluated in terms of costs and benefits, providing an assessment of their value in terms consistent with the organisation's requirements for formal decision-making. The same metrics can be used to track and review the implementation phase of interventions.

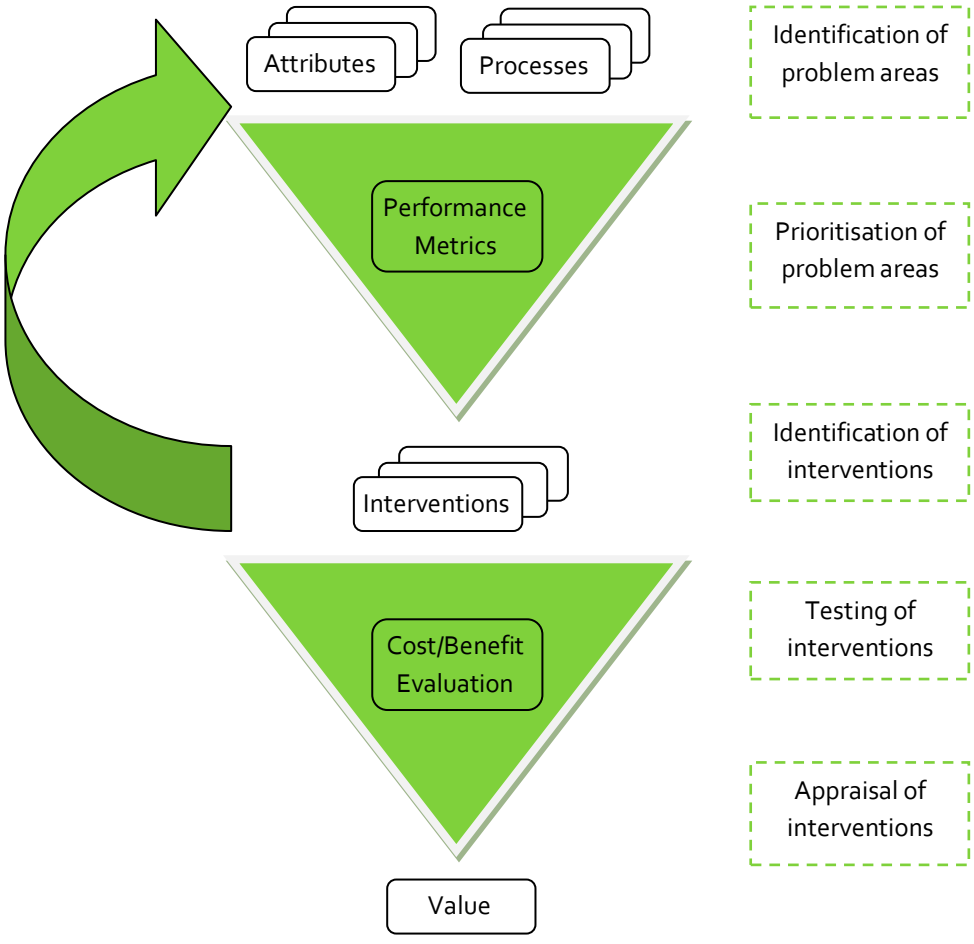

FIGURE 18 OVERVIEW OF METHOD

5.5.1 ORGANISATIONAL PROCESSES

A consequence of taking an organisation-wide view of customer IQ improvement is that a dollar saved from Marketing is worth the same as a dollar saved from Sales. As such, all organisational processes that rely on customer information to create value should be included within the scope. For many organisations, this is a potentially large number of processes spanning numerous information sources and organisational processes. Rather than a full audit of all processes, the Stake metric can be used to prioritise those processes that are likely to have business significance.

$$S = NfR^* \cdot \Pi$$

The factors are:

- **N**. The number of customers that are processed. This can be expressed as a percentage of the customer base for comparative purposes.
- **f**. The frequency with which the process is run. This can be expressed as an annual rate or, if the analysis has a fixed time horizon (such as four years), then the rate over that interval.
- **R^*.** The realisation matrix. Depending on the nature of the project, these probabilities may be expressed as business performance measures like lift or default rates.
- **Π.** They pay-off matrix. The dollar-amounts associated with success of failure of a decision-process may not be readily available on a per customer basis. In such cases, a suitable business owner may be able to nominate approximate amounts based on an understanding of the business cost structures and internal accounting model.

At this point, the organisational processes are ranked according to Stake. Therefore, the absolute dollar amounts and probabilities are not so important. Estimating these factors is likely to be difficult, so the analyst should proceed by trying to eliminate processes as soon as possible. For example, it is a waste of effort to conduct interviews with senior managers to ascertain particular values of Π for some process only to find out that it applies to a fraction of 1% of the customer base.

Another approach is to estimate a range (minimum value to maximum value) for these parameters. The product of the minimum values gives a lower bound, the product of the maximum values gives an upper bound and the product of the mean values gives a mean estimate. The upper bound on a process's Stake can be used to eliminate a process from further consideration. The lower bound can be used as a conservative measure for inclusion of a process on the list.

The output is a list of the most valuable organisational processes that rely on customer information, ranked by their Stake. (This metric can be interpreted as the amount of value lost due to poor IQ.)

5.5.2 DECISION-MAKING FUNCTIONS

The organisational processes under analysis rely on different customer attributes to different degrees and not all attributes are used in all processes. This step identifies the attributes that matter most using the Influence metric.

$$I_e = 1 - \frac{H(Y|X_e)}{H(Y)}$$

The components here are:

- **Y**. The probability distribution over the decision.
- **X_e.** The probability distribution over the e^{th} attribute.
- **H(·).** The entropy (uncertainty) function.

The Influence of a particular attribute on a particular decision-making function is computed without regard to the inner workings of the function. It can be described solely by inspection of the joint frequency of occurrence between inputs (IS value) and outputs (process decision). To take a tiny illustrative example, consider the Influence of a binary-valued attribute (gender) on a binary-valued decision (marketing offer):

$P(X_{gender}, Y)$	*Yes*	*No*
Male	0.25	0.1
Female	0.05	0.6

TABLE 11 EXAMPLE OF ATTRIBUTE INFLUENCE ON A DECISION

In this example, the customer base is 35% male and 65% female and the marketing offer is made to 30% of customers (*Yes*=30%). Using the Influence formula, I compute:

$$I_{gender} = 1 - \frac{H(Y|X_{gender})}{H(Y)}$$
$$= \frac{I(Y; X_{gender})}{H(X_{gender})}$$
$$= \frac{0.325}{0.934}$$
$$= 34.8\%$$

So in this case, the gender attributes has around a 35% bearing on the decision. This indicates how much uncertainty about the decision is removed when the gender is known. For a randomly selected customer, there is a 30% chance that the decision function will classify them as receiving an offer. However, upon finding out the customer is female, this drops to a 7% chance. (Had they been male, the probability would have increased to 83%.) In this sense, the attribute is considered quite influential.

As this calculation does not rely on having "true" external world values or "correct" decisions, it is readily available and cheap to perform using existing data. The only requirement is for a suitable query language (such as SQL or even XQuery) or a very simple spreadsheet.

Using this metric, the most influential attributes for the high-stake processes can be identified and ranked. To calculate the aggregate effect of an attribute upon a set of organisational processes, I can adopt the following heuristic:

The Importance *M* of attribute *a*, is the product of its Influence and Stake over the set of processes P of interest:

$$M_a = \sum_{p \in P} S_p I_a$$

Since the percentage values of the Influence do not add to unity, weighting the Stakes in this way results in a figure of merit that is useful only as a guide and should not be taken as a real financial measure. However, it does go some way to providing a sense of the financial importance of each attribute aggregated across the set of processes of interest.

The output of this step is a list of attributes that are most influential on decision-making in high-value processes, ranked by Importance.

5.5.3 Information System Representation

Using the short-list of important attributes, the analyst proceeds to examine the opportunities for improvement. The first place to look is the areas with the poorest information quality: the assumption is that the greatest scope for improvement (and hence financial gain) is those attributes that perform the worst.

As discussed, the raw "error rate" is not a good measure for improvement, owing to the issue of prior probabilities. Instead, the *Fidelity* metric (on a per-attribute basis) is a fairer way to compare attributes with one another:

$$\varphi_e = 1 - \frac{H(W_e|X_e)}{H(W_e)}$$

The Fidelity metric is a "gap measure" in the sense that it provides a normalised description of how far the reality falls short of the ideal. As such, it gives a sense of the comparative "improvability" of each attribute under examination.

Naturally, sourcing the true "external world" value of a customer attribute is expensive and difficult – if it were not this entire exercise would not be required! – so collecting these values should be done sparingly. In a practical sense, this is achieved by:

- **Identifying a suitable source.** Depending on the attribute, this could be done through finding an authoritative source (such as an applicable government registry) or direct customer confirmation. In some instances, it may be sufficient to find another trusted source system or industry benchmark source.
- **Sampling the customer base.** Since computation of the Fidelity metric requires only a probability distribution, only a sub-set of customers need to be audited. The question "How many customers do I need in my sample?" is answered by "Enough to be confident that the attributes are ranked in correct order." That is to say, it may not be necessary to estimate the metric for each attribute to a high degree of confidence as long as there is confidence in their rank[11].

Based on the trusted source and sample, the Fidelity metric for each attribute is computed and a new attribute ranking is produced. This new ranking can be used to eliminate attributes that, while important in the sense outlined above, are already of high-quality. This means the short-list of important attributes is further reduced to just those where significant improvement is both warranted and feasible.

5.5.4 Information Quality Interventions

The next step is to examine candidate IQ improvement interventions. These may have been proposed in advance of this analysis or new ones may have been proposed on the basis of the earlier analysis into organisational processes, decision-making functions and the representational effectiveness of the IS.

The short-list of attributes is the starting point. Interventions that address these are more likely to be (comparatively) valuable than ones that address attributes not on the list, since these attributes are

[11] *There is a well-developed body of knowledge around the practicalities of statistical sampling, especially sample size and statistical significance of rankings can be determined with Spearman's ρ or Kendall's τ (Neuman, 2000).*

both important and improvable. The value of a particular intervention, T, can be computed using the Yield formula:

$$V_T = \sum_p (R_p^* - R_p^{*\prime}) \cdot Nf\Pi_p$$

This requires computing two Realisation matrices across the set of processes of interest. The first relates the prior (actual) decisions Y to the decisions with perfect information, Y^*. The second relates the *revised* decisions Y′ (after the intervention) to Y^*. In general, this can only be achieved by applying the intervention (that is, correcting the customer data) and "re-running" it through the same decision processes and comparing the rate at which decisions change. This is likely to be expensive, time-consuming and possibly disruptive to operations since it uses operational resources like storage space, network bandwidth and CPU cycles.

Similarly to computing Fidelity, only estimates of overall probabilities are required so sampling will help reduce the cost of the exercise. However, before this is undertaken across all candidate interventions, some may be eliminated beforehand. Recall that for any intervention on a particular customer attribute, there will be a proportion of instances in which the corrected value "disagrees" with the original value.

Mathematically, I define this proportion as:

$$\tau = \Pr\left(X_e \neq X'_e\right)$$

(This metric is called *Traction*, since it characterises the degree to which an intervention actually changes the status quo.)

Not all of these interventions will be actionable, that is, result in a changed decision. Further, not all actionable changes will have a positive impact on value. In general, we might expect a trade-off between the proportion of customer records that are changed and whether or not the change is beneficial. An intervention with a large Traction may be termed aggressive, while an intervention that focuses on ensuring all changes are beneficial might be described as cautious.

I can estimate the Traction for an intervention without "re-running" the customer records through the decision function. Therefore, it is a comparatively cheap metric, given that a sample of the intervention is available. Taking a conservative approach, candidate interventions with low Traction can be eliminated if they have a low value even when assumed to be maximally cautious ie every change results in the maximum positive impact. This can be calculated by picking the maximum from each processes' pay-off matrix and multiplying it by τ. An example illustrates:

Suppose the *Region* attribute is under review. This attribute, X_{region}, has four states: *north*, *south*, *east* and *west*. The intervention, T, involves replacing values in *Region* with those from a database held by a recently-acquired subsidiary business with (presumably) better geo-coding, X'_{region}.

$$X_{region} = [\,0.15 \;\; 0.1 \;\; 0.15 \;\; 0.6\,]$$
$$X'_{region} = [\,0.15 \;\; 0.11 \;\; 0.25 \;\; 0.49\,]$$

I can compare the two by sampling the joint probability mass functions:

$$T_{region} = \begin{bmatrix} 0.14 & 0 & 0 & 0.10 \\ 0 & 0.08 & 0.02 & 0 \\ 0 & 0 & 0.13 & 0.02 \\ 0.01 & 0.03 & 0.14 & 0.46 \end{bmatrix}$$

The Traction for this intervention is given by:

$$\begin{aligned} \tau_{region} &= \Pr(X_{region} \neq X'_{region}) \\ &= 1 - (0.14 + 0.08 + \ 0.13 + 0.46) \\ &= 0.19 \end{aligned}$$

This means that 19% of customer records will be changed. Some of those changes will be actionable, some won't. Some of those actionable changes will have a positive impact on value (ie improve the decision) while the remainder will have a negative impact.

Suppose I consider an alternative intervention, T′, on the same attribute, via an internal consistency check with those customers with a current street address:

$$\begin{aligned} X_{region} &= [\ 0.15\ \ 0.1\ \ 0.15\ \ 0.6\ ] \\ X''_{region} &= [\ 0.16\ \ 0.14\ \ 0.15\ \ 0.45\ ] \end{aligned}$$

I can compare the two by sampling the joint probability mass functions:

$$T'_{region} = \begin{bmatrix} 0.15 & 0 & 0 & 0 \\ 0 & 0.09 & 0 & 0.01 \\ 0 & 0 & 0.15 & 0 \\ 0.01 & 0.01 & 0 & 0.58 \end{bmatrix}$$

The Traction for this second intervention is given by:

$$\begin{aligned} \tau'_{region} &= \Pr(X_{region} \neq X''_{region}) \\ &= 1 - (0.15 + 0.09 + \ 0.15 + 0.58) \\ &= 0.03 \end{aligned}$$

So the first intervention has a Traction of 19% while the second only 3%. While this doesn't necessarily mean the former is preferred, it does suggest that the latter could be dropped from the short-list for further (expensive) evaluation on the grounds that, even if it were perfect, it could not have more than a 3% impact on any given process. Of course, if further evaluation revealed that the first intervention was pathological (that is, introduced more errors than it removed), then this one could be pursued again.

The short-listed candidate interventions can now be examined in more detail. Again, a sampling approach is used to estimate the expected benefit of the intervention. To implement this, a sample of "corrected" customer records is fed into the decision-making function and the outputs (decisions) compared with original outputs. The differences in decisions are scaled by the pay-off matrix for each process and aggregated. The total expected value of each intervention (yield) is then computed:

$$V_T = \sum_p \left(R_p^* - R_p^{*\prime}\right) \cdot Nf\Pi_p$$

The interventions can then be ranked by their expected value. At this point, proposed interventions can be combined, disaggregated or more closely targeted. For example, fixing date-of-birth and postcode may both show significant benefits, enough to justify implementation. However, when combined, they may yield even higher returns than singly through decision synergy. Alternatively, the yield may improve if the intervention is targeted to the top 25% of customers, rather than applied across the entire customer-base.

Formerly-eliminated proposals can be revived if the short-listed interventions show a lower-than-expected benefit. If this occurs, then the "re-run" and subsequent analysis can be repeated in order to increase the value realised by the entire set of interventions.

5.6 Conclusion

This chapter presents a framework for valuing improvements to the quality of customer information. It comprises a mathematical model, grounded in semiotic and economic theory and used to derive performance measures, and a method for systematically analysing the value opportunities in customer IQ improvements.

The framework responds to practitioner requirements to build robust, testable business cases to support improvements in IQ based on cash-flows. These financial measures are necessary for IS practitioners and business managers to communicate the value of such initiatives and influence existing organisational resourcing processes.

The target users of the framework are analysts within the organisation engaged in a value-led, technology-agnostic analysis exercise. The method uses iterative elimination of proposals to focus on high-value opportunities and seeks to minimise wasted effort on irrelevant, ineffective or low-stakes interventions. It also takes into account the costs of acquiring information about the performance of different aspects of the model.

The output is a set of interventions which optimises the overall expected financial yield from the organisation's customer processes. This does not capture the intangible and the elusive "soft" benefits (improvements to morale, forecasting and planning, reputation etc), so it is going to be a lower-bound on the value of the interventions. However, it is a hard lower-bound that is more acceptable to financial controllers, enabling IQ projects to compete with a range of IS and non-IS investments.

The key constructs, metrics and steps in the method are outlined below. The analysis relies on the use of a set of candidate interventions, proposed by different business and IS stakeholders, being successively refined and eliminated.

Step	Construct	Metrics	Description	Resources
1	Organisational Processes	Stake (Realisation matrix, R)	An audit of organisational processes that rely on customer information is undertaken.	Internal documentation relating to process costs and performance.
2	Organisational Process	Stake (Process pay-offs, Π)	For each, the decision outcomes are identified and their pay-offs and penalties estimated.	Value estimates from process owners.
3	Decision-Making Function	Influence and Importance	For each process, the Influence of each attribute is computed. The aggregate Importance is then derived.	Transaction history of processes, including outcomes.
4	Information System	Fidelity	A sampling approach is taken to understand how well the IS represents the external world.	An authoritative information source (or surrogate) for attributes of interest.
5	Quality Interventions	Traction	Sampling of interventions used to gauge magnitude of change on the database.	Intervention applied to representative subset of records.
6	Quality Interventions	Yield	Promising interventions are "re-run" through process to estimate net financial benefit.	"Revised" records processed and compared with original.

TABLE 12 OUTLINE OF METHOD FOR VALUATION

From a project management perspective, the scope of the analysis is determined principally by two factors:

- The number of organisational processes of interest.
- The number of proposed quality interventions of interest.

Other factors – the number of attributes in the database and the state of existing knowledge – are not directly controllable but are determined by existing conditions. The time taken to prepare the analysis depends on the initial scope and the quality of the outputs, that is, the level of financial rigour and detail. This level is determined by the intended use: a formal cost/benefit analysis for very large projects will be judged to a higher standard than a smaller, informal analysis.

Much of the analysis can be re-used later. For example, the Influence metrics will remain constant as long as the underlying business logic and database definitions don't change too much. Similarly, Fidelity and Stake will change slowly with respect to business conditions and, once estimated, may only need to be updated periodically to retain their usefulness.

Chapter 6

Simulations

SIMULATIONS

6.1 SUMMARY

In Chapter 5, a theoretical framework for Customer Information Quality interventions was developed through a conceptual study. This framework seeks to describe quantitatively the effect of improving Information Quality (IQ) deficiencies on certain information systems (automated, data-driven customer processes in large organisations) and relate data attributes within these systems to wider organisational goals (business value). The intent is to provide analysts with some theoretically-grounded constructs, measures and steps to support them in evaluating investments in interventions to improve IQ. To evaluate and refine the theoretical development of the framework, simulations and mathematical analyses are used to investigate the relationship between improvements to IQ deficiencies and business value in this context.

Following a Critical Realist research approach, the investigation proceeds by triggering, under controlled conditions, the underlying mechanisms found in the ambient environment. The goal is to understand the operation of these mechanisms through the use of CMO patterns ("context-mechanism-outcome"). This understanding can then inform decision-making by practitioners in allocating resources to improve information quality (IQ).

The experimental design requires creating conditions that will activate (and inhibit) the mechanisms within the confines of the study, just as would happen in practice. To achieve this and ensure external validity, real-world customer datasets are sourced and decision models are developed and deployed using the same tools and procedures as encountered in practice.

The experiments employ a series of computer simulations of the customer processes to test the impacts of synthetic "noise" (information quality deficiency) upon the processes' performance. These results and subsequent mathematical analyses are used to validate the metrics developed in the framework (from Chapter 5) that helps analysts design, evaluate and prioritise investments in IQ interventions.

The key findings are that:

- the effects of the "garbling" noise process on customer data can be analysed mathematically with a high degree of confidence.
- the information-theoretic entropy metrics (derived from theory) are useful and practicable for selecting and prioritising IQ interventions.
- these metrics can be translated readily into business impacts, expressed in terms of cash flows.

Based on the internal validity (robust and careful experimental design and execution) and external validity (re-creation of ambient conditions), the case for the generalisability of the experimental results is made.

The rest of this chapter is structured as follows. Section 2 examines the philosophical basis for this empirical work, linking it back to the research design (Design Science) and research philosophy (Critical Realism). Sections 3 and 4 introduce the scenarios under examination (including the

datasets, algorithms and "noise" processes) as well as the explaining the practicalities and technical details of how the simulations were undertaken. Section 5 uses results from the simulations to argue that the theoretical framework developed in Chapter 5 (including its metrics) can be operationalised and used in practical situations. Specifically, Section 5.1 shows how the pattern of observed outcomes from the "noise" process is well-described by its mathematical characterisation. Section 5.2 demonstrates that the proposed Influence metric can be used as a cheaper, more practicable proxy for assessing the actionability of an attribute in a particular organisational process. Section 5.3 models the effects of interventions and "noise" on the organisation's costs and benefits. Finally, Section 6 looks at how these results can be packaged into a method for analysts to apply to specific situations.

6.2 Philosophical Basis

This section relates how concepts from Chapter 2 (Research Method and Design) apply to the design and conduct of a series of experiments into the effect of IQ deficiencies on the operation of customer decision models. Specifically, the applicability of Bhaskhar's Critical Realism for this task is discussed as well as the criteria for internal and external validity of the experiments (Mingers 2003).

In undertaking a Critical Realist (CR) experiment, it is important to recognise that the real world under study is ontologically differentiated and stratified (Bhaskar 1975; Bhaskar 1979) into three overlapping domains: *the real*, *the actual* and *the empirical*. The world exists independently of our experiences (that is, an ontologically realist position is adopted) while our access to and knowledge of it filtered through socially-constructed categories and concepts (an epistemologically interpretivist stance).

This view is appropriate when considering customer attributes and events, which may have their roots in the natural world but are defined through social mechanisms such as the law. For example, gender is a complex genetic, biological and social phenomenon. In the context of customer information systems, the role of chromosomes is irrelevant; what counts is the social construction, whether that is by legal definition (eg birth certificate) or self-definition (eg asking the customer). Similar remarks could be made for date of birth and marital status.

This matters because assessing how well a system describes the attributes, statuses and events associated with customers depends on the host organisation's existing shared concepts and categories. By way of illustration, consider marital status. Determining whether or not the list of possibilities is complete or that a given customer is mapped correctly will always be subjective and grounded in a particular culture or legal setting. Marriage is not a natural event or attribute and there is no objective determination of it. In CR terms, marital status is firmly in the transitive dimension.

The role of CR in these experiments extends to the nature of the claims to knowledge arising from them. The systems under analysis – data-driven customer decision-making processes – are systems that describe and predict aspects of real-world customers and their behaviour. Whether it is mortgage approvals or a direct mail campaign, these systems are created for (and assessed against) their ability to inform action based on likely future behaviours like credit defaults, responses to a marketing message and so on.

It is not a goal of this research to explain why people default on credit card payments or sign up to subscription offers, nor is it a goal to build better models of such behaviour. The "generative mechanisms" that trigger and inhibit such complex social behaviours lie, in ontological terms, in the realm of *the real*. Untangling these deep patterns of causality is outside the scope of this research. Instead, the domain of *the actual* is the focus, where we find events. In a business context, these

events include applying for a loan, changing one's name, making a purchase, signing up for a service and so on. We don't have access to these events (we can't perceive them directly) but instead our knowledge of them comes to us through our sense-perceptions of the "traces" they leave in *the empirical*: coloured images flashed on a computer screen, audio reproductions of human speech on a telephone in a call-centre, text printed on a receipt.

The databases and decision functions developed and deployed in the customer decision processes are themselves transitive objects embedded in the three domains. The "generative mechanisms" operating in *the real* are the underlying laws of nature that govern the operation of electrical and mechanical machinery. The patterns of causality operating at this level are highly controlled to give rise to the intended events in *the actual*: changes in system state and operation. We access the occurrence of these events through *the empirical*: perceptions of the graphics, texts and sounds (on screen or on paper) through which the state variables are revealed to us.

These system events (from *the actual* domain) are designed to reflect, or correspond to, the "real-world" customer events (also from *the actual* domain). Empirically, they may be entirely distinct. A decision model that predicts a customer will purchase a product may express this graphically on a computer screen. This looks nothing like an individual making a purchase at a cash register in a department store. However, the former corresponds to the latter (although they are not the same event).

Similarly, in the domain of *the real* the customer and system mechanisms are distinct. The underlying causal patterns that give rise to the purchase decision event are complex and grounded in social psychology, economics and cognitive science. The underlying causal patterns that gave rise to the model's prediction event are grounded in electronics and computer engineering, constrained by software code that implements a mathematical function. That the latter can predict the former is due to the ability of the model to mimic (to an extent) these complex psycho-social causal structures. The pattern of customers in a certain postcode being more likely to make a certain purchase is detected, extracted and then implemented by the model. This mimicking mechanism – a decision tree, neural net, regression function or similar – is entirely distinct from the customer's and bears no resemblance. From the perspective of the organisation that developed and deployed the decision model, it is only to be assessed against how well it predicts customer events, that is, how it performs in the domain of *the actual*. Superficial appearances in *the empirical* domain or deep understanding of causality in *the real* domain only matter to the extent that they impact upon events in *the actual*.

In terms of these experiments in information quality, the customers' "generative mechanisms" (in the domain of *the real*) that give rise to their behaviours (events in the domain of *the actual*) are not relevant. What is important is how the "generative mechanisms" within the customer information systems give rise to the systems' events (predictions of behaviours) and how the operation of these systems is perturbed by IQ deficiencies.

In short, these experiments concern the events arising from algorithmic models of customers, not customer behaviour. The extent to which these models reflect customers is not relevant for this research.

The experimental logic is to re-create in the laboratory these conditions (mechanisms, events and perceptions from their respective domains) and, in a controlled fashion, introduce IQ deficiencies. The impact of these deficiencies is manifested in the domain of *the actual*: the decision function ("generative mechanism") may give rise to different events when different (deficient) customer events (as encoded in the data) are presented.

Under certain conditions, the systems' "generative mechanisms" (the decision functions) remain unchanged and it becomes possible to ascribe causality to the changes in data. Specifically, if the underlying decision functions are, formally speaking, deterministic then any changes to observed events are attributed to changes in the input data alone and we can attempt to establish causality. This is only possible in closed systems where there is no learning or other changes taking place, such as with human decision-making.

Since these IQ deficiencies can only impact upon the operation of the customer information system at the level of *the actual*, their generative mechanism doesn't matter for these experiments. Consider the example of a mortgage approval system. If the "deceased flag" for a customer is changed from "alive" to "deceased", then it may impact upon the mortgage approval decision (it is generally unlawful to grant credit to a deceased person, as well as commercially unwise). Such a change ("noise[12]") may have a plethora of causes: perhaps a mis-keying at the point of data entry, a failed database replication process or a malicious fraud by an employee. Understanding the root-cause (aetiology) may be important for detecting the change and ensuring similar events do not occur subsequently. However, the *effect* of this change-event under the conditions of that particular mortgage approval system will be the same regardless of what caused it.

So understanding the "down-stream" effects of IQ deficiencies does not require understanding their "up-stream" causes. This is not to say that root-cause analysis is not important; it's just to say that it is not necessary for the purpose at hand.

These experiments are concerned with the effects of noise-events in customer information systems upon the prediction-events arising from the algorithmic models of customer behaviour. The external validity of the experiments rests on how well the laboratory re-creates the ambient conditions found in practice, and the extent to which the systems' "generative mechanisms" triggered (or inhibited) during the experiments mimic those found in practice. As argued above, it does not depend on how well the algorithms predict real customer behaviour, nor how well the IQ deficiencies match those found in ambient conditions.

From a CR perspective, the internal validity of the experiments is determined by whether the systems' "generative mechanisms" (that is, the operation of electro-mechanical machines constrained by software that implements a statistical customer decision model) give rise to changed prediction-events in the presence of induced IQ deficiencies manifested as noise-events. More simply, the internal validity depends on the how well I can exclude other potential explanations for changes in the prediction-events, such as programming errors or hardware failure.

The next section explains how the experiments were designed to meet these criteria for internal and external validity.

6.3 SCENARIOS

In order to ensure the external validity of the experiments, it is necessary to re-create the conditions and "generative mechanisms" in the laboratory as they operate in the ambient environment. This involves using technology, models, algorithms, statistics and data as employed in the target organisations, defined in Chapter 5 (Conceptual Study).

[12] *The term "noise" is used here to describe unwelcome perturbations to an information-bearing signal, as used in statistics, physics and engineering.*

It's important to note these contrived laboratory conditions are not a "simulation[13]" of the ambient environment: it's a reconstruction. Therefore, it's important that the pertinent elements that comprise the "generative mechanisms" are taken from those ambient environments. This means using real datasets, real tools and real customer models. The noise introduced, however, is synthetic and thus constitutes a simulation.

The IQ deficiencies are studied at the level of events (*the actual*) not mechanisms (*the real*). Hence, it not required to re-create the same root-causes or "generative mechanisms" in the laboratory as in the ambient environment. That is, the noise is deliberate and contrived rather than occurring as a result of, for example, mis-keying by data entry clerks, disk failure or poorly-specified business rules.

As a consequence, the noise in the experiments will not reflect ambient conditions in prevalence or distribution. What is important with this element is that the noise process introduces IQ deficiencies in a controlled fashion, allowing deductions to be drawn from the observed effects. In practical terms, a well-defined, replicable noise-adding procedure is required.

6.3.1 DATASETS

The requirement that the customer datasets be real constrained the candidate sets to those made publicly available for research purposes. The principle catalogue of such datasets is the UCI Machine Learning Repository (Asuncion and Newman 2007), which holds approximately 173 datasets. These datasets are donated by researchers and research sponsors (organisations) and are suitably anonymised and prepared for analysis. The typical use of these datasets is in the design, development and improvement of data mining and related statistical algorithms.

Of the 173 datasets, approximately one third each are categorical (or nominal) data, numerical (or ratio data) and mixed. Most are drawn from the domains of life sciences (48), physical sciences (28) and computer science/engineering (26). Five datasets explicitly relate to business and 14 to social sciences and hence are likely to be construed as customer data.

In selecting the datasets that represent the ambient environment, the criteria were:

- The selection should provide a representative class of decision tasks (classification, segmentation and prediction) using customer data.
- Each dataset should contain both nominal and numerical data types.
- The customer attributes should be generic (applicable across domains).
- There should be sufficient attributes (columns) and instances (rows) to build realistic models.
- The size and complexity of the datasets should not present practical or resourcing issues for the experiments.

Based on a review of the available options and these criteria, three datasets were selected for analysis, as described below.

[13] *One may argue that statistical models created by organisations are, in a fashion, simulations of customer behaviour. In this case, these experiments are a re-construction of a simulation.*

6.3.1.1 ADULT

This dataset was derived from US Census data (1994) and was donated in 1996. It contains 48,842 instances (customer records) with the following 15 attributes (columns):

Attribute	Type	Values
Age	Numerical	
Workclass	Nominal	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlgwt	Numerical	
Education	Nominal	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	Numerical	
marital-status	Nominal	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Nominal	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship	Nominal	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	Nominal	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Nominal	Female, Male
capital-gain	Numerical	
capital-loss	Numerical	
hours-per-week	Numerical	
native-country	Nominal	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
income	Nominal	<=50K, >50K

TABLE 13 ADULT DATASET

The last attribute – income – is the target variable, or class, for the scenario; that is, the task is to predict (or classify) whether a given customer's income is over or under $50,000 per annum. In a practical context, such a task may be performed in order to support the targeting of marketing messages (for example, branding or making an offer).

Note that for reasons of practicality, the dataset was sampled from almost 50,000 instances down to 10,000 (20% random sampling). This significantly improved the time and computing resources required during the experiments without impacting upon the validity of results. This is discussed further below.

6.3.1.2 CRX

This dataset concerns credit applications in an Australian financial service provider (identity is confidential). It was supplied by Ross Quinlan (University of Sydney) in 1986. The 16 attributes have been de-identified so that the semantics of the labels is not recoverable. There are 690 customer instances in this dataset.

Attribute	Type	Values
A1	Nominal	b, a
A2	Numeric	
A3	Numeric	
A4	Nominal	u, y, l, t
A5	Nominal	g, p, gg
A6	Nominal	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff
A7	Nominal	v, h, bb, j, n, z, dd, ff, o
A8	Numeric	
A9	Nominal	t, f
A10	Nominal	t, f
A11	Numeric	
A12	Nominal	t, f
A13	Nominal	g, p, s
A14	Numeric	
A15	Numeric	
A16	Nominal	+, -

TABLE 14 CRX DATASET

The target variable, A16, presumably relates to subsequent customer credit worthiness (+) or defaulting behaviour (-), though this is conjecture.

6.3.1.3 GERMAN

This dataset relates another customer credit task, this time using a wider range of customer attributes (21) to determine a personal loan in Germany. It was donated by Hans Hoffman (University of Hamburg) in 1994. It contains 1000 customer records, as detailed below.

Attribute	Type	Values
Account Status	Nominal	A11, A12, A13, A14
Duration	Numerical	
Credit History	Nominal	A30, A31, A32, A33, A34
Purpose	Nominal	A40, A41, A42, A43, A44, A45, A46, A47, A48, A49, A410
Credit Amount	Numerical	
Savings	Nominal	A61, A62, A63, A64, A65
Employment	Nominal	A71, A72, A73, A74, A75
Installment Ratio	Numerical	
Marital Status and Sex	Nominal	A91, A92, A93, A94, A95
Guarantors	Nominal	A101, A102, A103
Residence	Numerical	
Property	Nominal	A121, A122, A123, A124
Age (years)	Numerical	
Other credit	Nominal	A141, A142, A143
Housing	Nominal	A151, A152, A153
Existing Credits	Numerical	
Job	Nominal	A171, A172, A173, A174
Dependents	Numerical	
Telephone	Nominal	A191,, A192
Foreign Worker	Nominal	A201, A202
Creditworthiness	Nominal	Good, Bad

TABLE 15 GERMAN DATASET

There are two identified limitations with the selected datasets. First is the nature of the decision task. They are all binary-valued rather than a more complex segmentation task, such as assigning each customer to one of, say, 50 segments or recommending a product to a customer based on related purchases (collaborative filtering). Secondly, the decision distributions in all three dataset are quite balanced (ranging from a 30-70 split to a 50-50 split). Some applications, like fraud detection, are heavily unbalanced (with perhaps <1% of customers being fraudulent).

These three datasets meet the explicit criteria. The decision tasks relate to three typical customer decision processes: segmentation (perhaps to support a direct marketing campaign), credit card approval and the awarding of personal loans. All datasets contain a mix of nominal and numerical data and commonly-used customer attributes such as sex, age, education, credit history, work type, income and family situation are represented. The number of attributes ranges from 15 to 21, while the number of instances ranges from 690 to 10,000. These datasets have been used widely by data mining and related researchers for many years to develop and test models, and so are likely to have been deemed sufficiently representative of what happens in practice by this research community.

6.3.2 DECISION FUNCTIONS

The goal in selecting the statistical algorithms used to produce the decision functions is to reproduce in the laboratory the same generative mechanisms found in the ambient environment. This means choosing a subset of candidate functions that is likely to be representative of the modelling tools and techniques used in practice. Thus algorithms were selected not on the basis of their performance *per se*, but on their wider adoption by customer modelling practitioners. Hence, functions were sought

that have been known for a long time, are well understood by researchers, taught at universities to students, implemented in many software packages and feature in textbooks.

Based on these criteria, the following five decision functions (algorithms) were selected. These descriptions are illustrative, and more detail about the parameter selection is provided subsequently. The specifics of how these decision functions operate are not important for the present purpose, but details can be found in most data mining texts (Han and Kamber 2006).

- **ID3** (ID3-numerical). This algorithm is the modified version of the original ID3 rule induction algorithm, designed to deal with numerical as well as nominal data. ID3 was one of the first decision-tree algorithms to be developed and remains an important algorithm for building decision trees.
- **AD** (Alternating Decision Tree). Another decision tree algorithm, this one is more modern (1999) and employs the machine learning technique of *boosting*.
- **NB** (Naïve Bayes Tree). This decision tree induction algorithm uses the Naïve Bayes algorithm at the leaves. That is, it assumes statistical independence in the input attributes.
- **BNet** (Bayes Net). This algorithm employs a Bayesian Network (a model of interconnected nodes, weighted by their conditional probabilities) to construct the decision function.
- **LMT** (Logistic Model Tree). In this algorithm, linear logistic regression is embedded within the tree induction process.

Two other algorithms were considered for inclusion owing to their popularity. Quinlan's C4.5 algorithm (Quinlan 1993) is a successor to the ID3 that uses a different splitting criterion (information gain ratio instead of information gain), supports nominal data and handles missing values. Since the implementation of ID3 used here addresses all these issues[14], C4.5 was excluded. The second algorithm was the CHAID ("chi squared automatic interaction detection"). This algorithm uses the familiar χ^2 statistic as the splitting criterion during tree induction. However, when evaluated under test conditions it behaved the same as ID3.

The major limitation with this set of decision functions is that issues of practicality means certain classes of more esoteric functions (such as neural networks, genetic algorithms and support vector machines) are omitted, as is the more mundane approach of manually-derived IF-THEN type decision rules used in simpler, smaller-scale situations. However, there is no theoretical or practical reason to suggest that these alternatives would behave markedly differently.

A survey of 203 data mining and analytics practitioners in March, 2007 suggests that the use of these esoteric algorithms are not widespread, with neural networks reportedly used by 17.2% of respondents within the previous 12 months, support vector machines by 15.8% and genetic algorithms by 11.3% (Piatetsky-Shapiro 2007b). By contrast, decision trees were used by 62.6%.

The set of machine learning rule induction algorithms selected here are non-trivial, widely used and based on different theoretical approaches (Information Theory, Bayesian statistics and linear logistic functions). They are well understood by researchers, supported by most statistical and data mining software packages and are practicable to implement.

6.3.3 NOISE PROCESS

Since these experiments are concerned with the *effects* of IQ deficiencies and not the *causes*, a "noise process" is required that can introduce errors to the datasets in a controlled fashion. That is, the noise

[14] *Discussed further below in Section 4 (Experimental Process).*

process is not required to reflect the "generative mechanisms" for noise in ambient conditions but rather one that induces noise-events in the domain of *the actual*.

In order to allow other researchers to verify these results, the noise process should be simple, practicable (in terms of programming and execution effort), analysable (closed-form algebraic solutions), repeatable, generic (applies to all data types) and explicable (without requiring deep mathematical knowledge). Failure to meet any of these criteria would undermine the replicability of the study.

The noise process selected is termed garbling[15] (Lawrence 1999) and it is applied to the dataset on a per-attribute (column) basis. It has a single parameter, g, that ranges from 0 (no garbling) to 1 (maximal garbling). In essence, it involves swapping field values in the dataset, as follows:

For a given dataset attribute (column) and garbling rate, *g*:

1. For the ith customer, C_i, pick a random number, p, on the interval (0,1].
2. If $p \leq g$ then garble this value ...
   2.1. Select another customer, C_j, at random from the dataset.
   2.2. Swap the ith and jth customers' values for the given attribute.
3. Move to the (i+1)th customer and repeat step 1 until all records processed.

Thus, when g=0 none of the records are garbled and when g=1 all of them will be. In this way, a controlled amount of noise is introduced to any attribute in each of the datasets. (Please see Appendix 1 for the actual source code, in JavaScript, that implements this algorithm.)

It should be noted that not all records will be *altered* when they are garbled. Consider an example when Customer Record #58 has the "gender" field (presently "male") garbled. Suppose Customer Record #115 is randomly selected to swap with #58. Suppose #115 also has a "gender" field of "male", so that when the values are swapped they are both unaltered. This is analysed in detail in Section 6.5.1.

This noise process has some desirable properties. Firstly, it is quite intuitive and easy to visualise and implement. Secondly, it preserves the prior probability distributions over the dataset. That is, if the breakdown of frequencies of field-values in a given attribute is [0.15, 0.35, 0.2, 0.3] beforehand, then this will not change after garbling. As a consequence, it will not change the distributional statistics like mode or mean (for numeric data types). Lastly, it handles numeric data types without assuming an underlying statistical model. For example, it does not require assumptions of linearity or normality and since it only ever uses existing field-values it will not inadvertently generate "illegal" values (such as negative, fractional or out-of-range values).

In an information-theoretic sense, garbling is a "worst-case" noise event. That is, all information about the true external world value is completely and irrevocably lost: if a customer's attribute has been garbled then there is absolutely no clue as to what the original value might have been. An observer is no better informed about the external world value after looking at the record than before. By contrast, another noise process like Gaussian perturbations (ie adding a random offset) retains some clue as to the original value.

In terms of the Ontological Model of IQ (Wand and Wang 1996), this worst-case situation occurs when ambiguity is maximised, perhaps as a result of a design failure. For example, the system is in a

[15] *The term "garbling" is adapted from Blackwell's seminal work in The Theory of Experiments*

meaningless state (ie one which does not map to an external world state) or there is an incomplete representation (the external-world value cannot be expressed in the system).

In a practical sense, this kind of "disastrous" error would arise in situations where:

- a field-value has been deleted or is missing,
- an indexing problem meant an update was applied to the wrong customer record,
- two or more customers have been inadvertently "fused" into the same record,
- the field has been encoded in such a way as to be meaningless to the user or application,
- an external-world value is not available, so an incorrect one is used instead "at random".

It would not apply to situations where the IQ deficiency retains some information about the original value. For example, a problem of currency in customer addressing might arise when a customer changes residence. In such situations, the correct external world value is likely to be correlated to some degree with the "stale" value. Another example might be the use of subjective labels (such as eye colour) where one might expect some correlation between incorrect and correct values (eg "brown" is more likely to be mis-mapped to "hazel" than "blue"). Lastly, issues around precision in hierarchical data (a form of ambiguity) would also not be reflected by this process. For example, mis-describing a customer as residing in "United States" rather than "San Francisco, California, USA" would not arise from garbling.

6.4 Experimental Process

This section outlines the sequence of steps undertaken to implement the series of experiments. The goal is to explain how the internal and external validity of the study was maintained, to place the outcomes into context and allow repetition of the study to verify outcomes.

6.4.1 Technical Environment

The technical environment for the experiments were contrived to reproduce the ambient conditions found in practice. As outlined in Section 3, the scenarios (including the datasets, decision tasks and algorithms) were selected against criteria designed to realise this reproduction. The implementation platform for the experiments was constructed in keeping with this goal, and comprised the following technical elements:

- standard low-end desktop PC (ie 2GHz processor, 512MB RAM, 120GB HDD, networking and peripherals),
- windows XP SP2 (operating system),
- RapidMiner 4.1 (data mining workbench),
- WEKA 3.410 (machine learning algorithm library),
- Microsoft Excel (data analysis spreadsheet),
- GNU command line tools (batched data analysis),
- wessa.net (online statistics package).

The bulk of the model building, experimental implementation and data collection were undertaken with the RapidMiner tool. This is the leading open source data mining and predictive analytics workbench. Formerly known as YALE ("Yet Another Learning Environment"), it is developed by the University of Dortmund, Germany, since 2001. It is a full-featured tool for building, testing and analysing models, incorporating a very large number of learning algorithms with a graphical user interface for setting up experiments.

WEKA ("Waikato Environment for Knowledge Analysis") is a similar open source workbench, developed by New Zealand's University of Waikato since 1997. WEKA's library of over 100 learning functions is available for use within the RapidMiner environment and, owing to its more comprehensive selection and code documentation, was used in this instance.

A survey of 534 data mining and analytics practitioners in May 2007 found that RapidMiner was ranked second, used by 19.3% of respondents (Piatetsky-Shapiro 2007a). The most used was the commercial product, SPSS Clementine, at 21.7%. WEKA had a 9.0% share. While web-based surveys are open to "gaming" by vendors with a commercial interest – as acknowledged by the researchers – this does provide support for the assertion that the laboratory conditions in this experiment recreate those found in ambient environments.

It must be emphasised that both commercial and open source tools are used to build, validate and analyse exactly the same decision models as they implement a roughly overlapping set of learning algorithms. While they differ in their interfaces and have some variation in their capabilities, the resulting models are identical and as "generative mechanisms", invoke the same events in the domain of *the actual*.

6.4.2 CREATING MODELS

The first step is to create the decision functions (models) from each of the three datasets (ADULT, CRX and GERMAN) using each of the five learning algorithms (ID3, AD, NBTree, BNet, LMT). This results in 15 decision models.

As explained in Section 3, the datasets contain a set of attributes and a target variable, or class, which is the "correct" decision or classification, as assessed by the domain experts who provided the datasets. The purpose of the learning algorithms is to build models that successfully predict or classify this target value. The attribute values are taken to be "correct" but there are some missing values. The ADULT dataset has 1378 missing values (0.98%), CRX has 67 (0.65%) while GERMAN has none. RapidMiner's built-in missing value imputation function was used to substitute the missing values with the mode (for nominal data) or mean (for numerical data).

Building the model consists of presenting the data in CSV (comma separated value) format to the RapidMiner tool and applying the specified learning algorithm to construct the model. In most cases, the learning algorithm has a number of parameters that are available for tuning the performance. Rather than employing sophisticated meta-learning schemes (whereby another learner is used to tune the parameters of the original model), modifications were made by hand, using the performance criterion of "accuracy"[16]. To ensure the models weren't "over-trained", automatic validation was employed where the algorithm was tested against a "hold out set" (subset of data unseen by the learning algorithm).

The resulting 15 models are considered, for the purpose of these experiments, to be those constructed with "perfect information". In each case, the model was exported as a set of rules (or weights), encoded in XML for subsequent re-use.

To illustrate, below is the decision model created by the ID3 algorithm with the ADULT dataset is reproduced in a graphical form. In general, the resulting models are far too complex to be visualised in this way (especially the Bayesian Networks and Logistic Modelling Trees). However, this image does give some insight into the general form that these tree-type decision models take: a series of

[16] *In this software package, it simply means "percentage correctly classified". Since we are dealing with binary decision problems, it is adequate and entropy-based measures are not required.*

nodes that encode conditional logic (IF-THEN rules) being traversed sequentially before arriving at a "leaf node" or final prediction or classification, in this ">50K" or "<=50K".

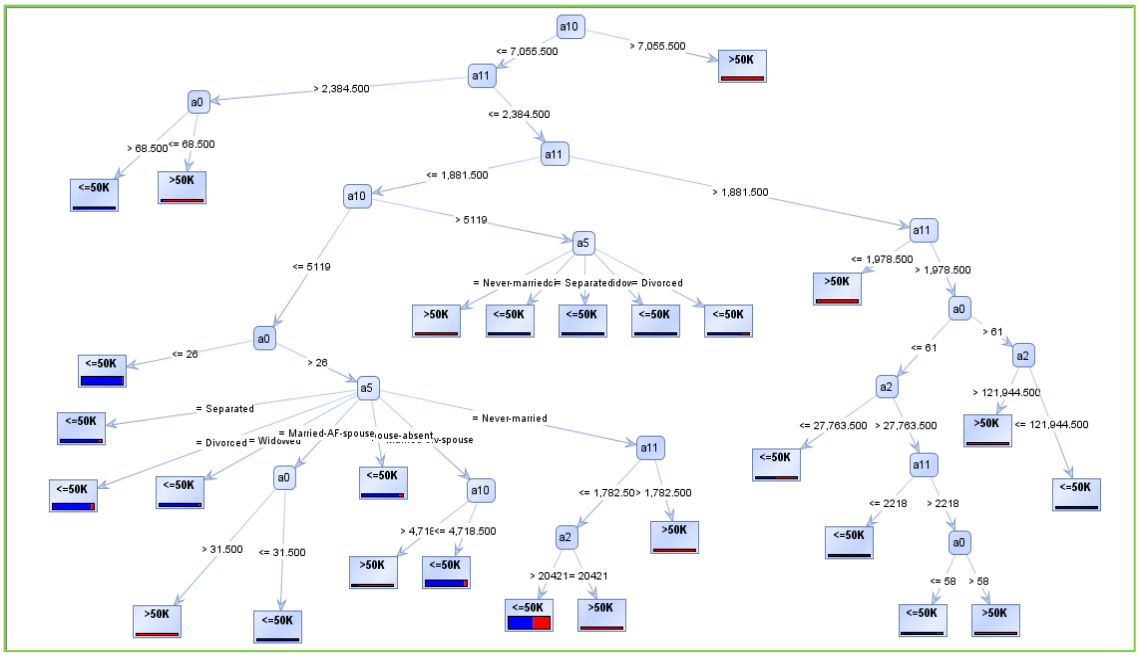

FIGURE 19 ID3 DECISION TREE FOR ADULT DATASET

The resulting models' performances are detailed below. Rather than just provide the classification "accuracy" rates (in bold), these are reported as "mistake rates" broken down into Type I mistakes (false positives) and Type II mistakes (false negatives), where positive in this case refers to the majority class. (Given that the classes are in general quite well-balanced, the ordering is somewhat arbitrary.) This extra information is used in subsequent cost-based analysis, since different mistake types attract different costs.

Model Mistake Rates (Type I, Type II)	**ADULT**	**CRX**	**GERMAN**	***Averages***
ID3	**18.3%** (17.6%, 0.67%)	**14.1%** (11.0%, 3.04%)	**27.2%** (16.6%, 10.6%)	*19.9%*
AD	**14.5%** (9.91%, 4.63%)	**12.8%** (5.65%, 7.10%)	**14.6%** (14.7%, 9.90%)	*14.0%*
NBtree	**13.0%** (7.99%, 4.98%)	**5.80%** (2.46%, 3.33%)	**18.3%** (11.9%, 6.40%)	*12.4%*
BNet	**16.7%** (4.80%, 11.9%)	**11.7%** (3.48%, 8.26%)	**22.9%** (13.3%, 9.60%)	*17.1%*
LMT	**13.1%** (9.11%, 3.97%)	**3.19%** (1.45%, 1.74%)	**17.0%** (11.2%, 5.80%)	*11.1%*
Averages	*15.1%*	*9.52%*	*20.0%*	*14.9%*

TABLE 16 - DECISION MODEL PERFORMANCE BY ALGORITHM AND DATASET

The internal validity of these experiments hinges on using the tools correctly in applying the nominated algorithms to the datasets and producing the intended models. To support this, model performance results were sourced from the peer-reviewed machine learning literature and compared with these experimental results. As it was not possible to find studies that produced results for every algorithm on every dataset, a representative sample across the datasets and algorithms was chosen. Note also that, as discussed in Section 3b above, C4.5 is a substitute for the ID3-numerical (with

information gain ratio as the splitting criterion, as used in this study). Since studies on ID3-numerical weren't found, C4.5 is used for comparing these results with other studies.

Firstly, when Kohavi introduced the NBtree algorithm he compared the new algorithm against Quinlan's C4.5 using a number of datasets, including ADULT and GERMAN (Kohavi 1996). Summarising, he found that C4.5 on ADULT (at 10,000 instances) had a mistake rate of 16%[17] and the NBtree algorithm improved that by 2% (to 14%). Sumner, Frank and (2005) reported mistakes rates using LMT on ADULT at 14.39% and using LMT on GERMAN at 24.73%. Ratanamahatana and Gunopulos (2003) reported a mistake rate of 26% on GERMAN with C4.5. For the CRX dataset Liu, Hsu and Yiming (1998) report 15.1% mistake rate using C4.5 and 27.7% with the same algorithm on GERMAN. Cheng and Greiner Cheng (1999) found a mistake rate of 14.5% for BNet on ADULT. For the AD algorithm, Freund and Mason (1999) found a mistake rate of 15.8% on CRX. For the same dataset and algorithm, Holmes, Pfarhringer et al. (2002) had a result of 15.1%.

Of course, the numbers reported in the literature do not exactly align with those found here. Slight differences can be accounted for by factors such as the handling of missing values (some studies simply drop those instances; here, the values were instead imputed) or the setting of particular tuning parameters (these are not reported in the literature so reproducing them is not possible). The largest discrepancy was for LMT on GERMAN (17.0% here compared with 25.0% in one study). This algorithm also has the largest number of tuning options, increasing the chances of divergence and the possibility of model over-fitting.

The overall closeness of the results reported in the literature with those reproduced here give support to the claim of internal validity: the events induced in these experiments (reported as performance metrics) result from the triggering of the intended underlying generative mechanisms and not "stray" effects under laboratory conditions. This gives assurance that the technical environment, selection of tools, handling and processing of datasets, application of algorithms and computing of performance metrics was conducted correctly, ensuring that, for example, the wrong dataset wasn't accidentally used or there is a software problem with the implementation of an algorithm.

6.4.3 Data Preparation

The next step is to prepare the data for analysis. A simple web page was developed to provide an interface to custom JavaScript code used in the data preparation. This form consisted of an input for the dataset and testing and tuning parameters to control the process of introducing noise.

Firstly, as outlined above, two of the datasets had some missing values (ADULT and CRX), denoted by a "?" character. For nominal data, they were substituted with the mode value. For numerical data, the mean was used. This is known as imputation. Secondly, noise was introduced using the garbling algorithm (Section 3.3). As discussed, this involved iterating through each attribute, one at a time, and swapping data values according to a threshold parameter, g. For each attribute in each dataset, ten levels of garbling at even increments were applied (g=0.1, 0.2, …, 1.0). Finally, the resulting "garbled" datasets were written out in a text-based file format, ARFF, used in a number of analytic tools.

This data preparation step resulted in ten "garbled" datasets for each of the 49 attributes (that is, 14 from ADULT, 15 from CRX and 20 from GERMAN) for a total of 490 datasets.

[17] *This is the same as the reported results for ID3 in the notes attached to the dataset.*

6.4.4 EXECUTION

This phase involves applying each of the five derived decision functions to the 490 garbled datasets, sequentially, and determining how the resulting decisions differ from the original. It is realised by using the RapidMiner tool in "batch mode" (that is, invoked from the command line in a shell script, rather than using the Graphical User Interface).

It's worth emphasising that the decision functions are developed (trained) using "perfect information" – no garbling, no missing values. Even so, the mistake rate (misclassifications) is around 15%, reflecting the general difficulty of building such decision functions. This study is concerned with the incremental effect of noisy data on realistic scenarios[18], not the performance of the decision functions themselves. So, in each case, the baseline for evaluation is not the "correct decision" (supplied with the dataset) as used in the development (training) phase. Rather, the baseline is the set of decisions (or segments or predictions) generated by the decision function on the "clean" data (g=0). To generate this baseline data, the models were run against the three "clean" datasets for each of the five decision functions, for a total of 15 runs.

For each of the five decision functions, all 490 garbled datasets are presented to the RapidMiner tool, along with the baseline decisions. For each instance (customer), RapidMiner uses the decision function supplied to compute the decision. This decision is compared with the baseline and, if it differs, reported as a misclassification. Since all scenarios involved a binary (two-valued) decision problem, the mistakes were arbitrarily labelled Type I (false positive) and Type II (false negative).

In total, 2450 runs were made: five decision functions tested on the 49 attributes (from three datasets), each with 10 levels of garbling. The estimated computing time for this series of experiments is 35 hours, done in overnight batches.

6.4.5 DERIVED MEASURES

The last step is computing a key entropy statistic, the Information Gain, for each attribute to be used in subsequent analyses. This is done within the RapidMiner environment and using the built-in Information Gain function. For the case of numerical attributes, the automatic "binning" function (minimum entropy discretisation) was used to create nominal attribute values and the numerical values mapped into these.

A series of shell scripts using command line GNU tools (grep, sed and sort) pull the disparate metrics into a single summary file with a line for each experiment comprising:

- decision function identifier
- dataset identifier
- attribute identifier
- garble rate
- garble events
- error rate
- mistakes (total, Type I and Type II)

This summary file was loaded into a spreadsheet (Excel) for further analysis, as outlined below.

[18] *The pathological case of an error in the data actually improving the decision (ie an error correcting a mistake) is discussed in Section 6.*

6.5 RESULTS AND DERIVATIONS

The key results are provided in tabular form, grouped by experimental conditions (dataset and algorithm). Supporting derivations and analyses are provided below alongside the experimental results, to show how they support the development of the metrics used during the design and appraisal of IQ interventions.

6.5.1 EFFECTS OF NOISE ON ERRORS

The first metric, called gamma (γ) relates the garbling mechanism with actual errors. Recall that an *error* refers to a difference in the attribute value between the external world and the system's representation. For example, a male customer mis-recorded as female is an error. Suppose a correct attribute value (male) is garbled, that is, swapped with an adjacent record. There is a chance that the adjacent record will also be male, in which case, the garbling event will not introduce an error. The probability of this fortuitous circumstance arising does not depend on the garbling process itself or the rate of garbling (g), but on the intrinsic distribution of values on that attribute.

Attribute	ID3	AD	NB	BNet	LMT	Average
a0	98%	98%	98%	98%	97%	**98%**
a1	43%	43%	43%	43%	43%	**43%**
a2	100%	100%	100%	100%	100%	**100%**
a3	81%	81%	80%	81%	81%	**81%**
a4	81%	81%	81%	81%	81%	**81%**
a5	66%	67%	66%	67%	67%	**67%**
a6	89%	89%	88%	89%	89%	**89%**
a7	74%	73%	73%	72%	74%	**73%**
a8	26%	26%	25%	26%	25%	**26%**
a9	44%	44%	44%	44%	44%	**44%**
a10	16%	16%	16%	16%	16%	**16%**
a11	9%	9%	9%	9%	9%	**9%**
a12	76%	76%	76%	77%	75%	**76%**
a13	15%	15%	15%	15%	16%	**15%**
c0	41%	46%	41%	44%	44%	**43%**
c1	100%	101%	101%	101%	99%	**100%**
c2	97%	101%	99%	100%	101%	**100%**
c3	37%	37%	35%	36%	36%	**36%**
c4	38%	37%	36%	37%	37%	**37%**
c5	88%	91%	89%	92%	93%	**91%**
c6	59%	58%	61%	61%	60%	**60%**
c7	97%	99%	95%	96%	97%	**97%**
c8	49%	50%	49%	49%	50%	**49%**
c9	50%	52%	52%	52%	50%	**51%**
c10	69%	68%	65%	67%	66%	**67%**
c11	49%	50%	48%	49%	49%	**49%**
c12	17%	17%	17%	18%	16%	**17%**
c13	94%	97%	94%	93%	94%	**94%**
c14	81%	83%	79%	80%	81%	**81%**
g0	70%	69%	70%	69%	69%	**69%**
g1	91%	89%	87%	89%	87%	**89%**
g2	62%	62%	62%	62%	63%	**62%**
g3	82%	82%	82%	82%	81%	**82%**
g4	101%	101%	100%	100%	101%	**101%**
g5	61%	59%	58%	59%	59%	**59%**

g6	76%	75%	75%	77%	77%	**76%**
g7	66%	68%	69%	69%	67%	**68%**
g8	60%	59%	58%	61%	58%	**59%**
g9	18%	16%	18%	17%	18%	**17%**
g10	70%	69%	71%	69%	69%	**69%**
g11	73%	73%	73%	73%	70%	**73%**
g12	98%	98%	95%	97%	97%	**97%**
g13	31%	32%	31%	31%	31%	**31%**
g14	45%	46%	46%	45%	45%	**46%**
g15	47%	49%	49%	49%	48%	**48%**
g16	52%	54%	54%	53%	54%	**54%**
g17	27%	25%	25%	25%	25%	**25%**
g18	49%	48%	48%	49%	48%	**49%**
g19	7%	7%	7%	8%	7%	**7%**

TABLE 17 GAMMA BY ATTRIBUTE AND DECISION FUNCTION

The attributes are labelled so that the first letter (a, c or g) corresponds to the dataset (ADULT, CRX and GERMAN respectively). The following number indicates the attribute identifier within the dataset. Note that, by definition the valid range for γ is 0% to 100%, but that some values here are in excess (101%). This is because, for a given garbling rate g, the actual number of garbles performed has a small variance around it.

So, to illustrate, the attribute a9 ("sex", in the ADULT dataset) has a γ of 44%. This means that swapping a customer's sex value with another chosen at random has a 56% chance of leaving the value unchanged.

The derivation for γ is as follows. Firstly, I model the garbling process as a simple 1st order Markov chain, with a square symmetric transition probability matrix where the marginal distribution follows the prior probability mass function. For example, attribute a9 from above has prior probability distribution of $A = [0.67\ 0.33]^T$. The garbling process for this attribute can be modelled with the following matrix, T_{a9}:

$$T_{a9} = \begin{bmatrix} 0.67 \\ 0.33 \end{bmatrix} [\ 0.67\ 0.33\ ]$$

$$= \begin{bmatrix} 0.45 & 0.22 \\ 0.22 & 0.11 \end{bmatrix}$$

The interpretation is that 45% of the time that a record is garbled, it will start out as "male" and stay "male" (no error), 11% of the time it will start out as "female" and remain "female" (no error) and the remaining 44% of the time (22% + 22%) the value will change (an error). In this sense, γ is a measure of the inherent susceptibility of the attribute to error in the presence of garbling.

In general, for a given attribute's probability distribution over N values $[w_1, w_2, \ldots, w_N]$, the value of γ (probability of garbling leading to an error) is computed by summing the off-diagonal values:

$$\gamma = 1 - \sum_i w_i^2$$

This metric effectively measures the inverse of the "concentration" of values for an attribute. An attribute that is "evenly spread" (eg. A = [0.24 0.25 0.26 0.25]) will have a high γ value, approach 1. The most "evenly spread" distribution is the uniform distribution, when each of the N values is equal to 1/N:

$$y_{max} = 1 - \sum_i \frac{1}{N^2}$$

$$= 1 - \frac{N}{N^2}$$

$$= 1 - \frac{1}{N}$$

For example, attribute a_2 has γ =0.999 because it has a large number of possible values, each with approximately 2% probability of occurring.

By contrast, an attribute that is highly concentrated, A = [0.001 0.001 0.001 0.997] will have a lower γ approaching 0. The extreme case is when an attribute follows the Kronecker's delta distribution of [1 0 0 0 … 0]. In this case, $\gamma_{min} = 0$.

So γ is an intrinsic property of an attribute, constant regardless of the underlying garbling rate, g, or indeed how the attribute is used in decision functions. I can now analyse how γ and g combine to produce the observed errors under the garbling process described here.

I begin by defining ε as the error rate of an attribute, g is the garbling parameter and γ is as above. Recall that the garbling process works by sweeping through each customer record and for each record, with probability g, will swap that record's value with another. By way of terminology, I say the original record is the source and the randomly selected second record is the target. As shown above, the chance that this swap will result in a changed value (ie error) is γ.

However, the proportion of records that is garbled is not simply g, the garbling parameter. The reason is that a given customer record might be selected for swapping during the sweep (that is, as the source, with probability g) but it may also be selected as the target in another swap. Whether selected as a source or a target, there is still a probability γ that the swap will result in an error.

$$\varepsilon = \gamma R_S + \gamma R_T$$

Here, R_S is the rate at which records are the source in a swap and R_T is the rate at which they are the target. Clearly, R_S is g, the garbling parameter. A value of 0 implies no record is selected for swapping (hence $R_S = 0$) and a value of 1 implies all records are selected (R_S=1). To calculate R_T, the probability of a record being selected as a target, consider the simple case of a thousand records and g=0.1. In this scenario, I would expect 100 swaps (1000*0.1). This means there are 100 sources and 100 targets. Each record has a 1/1000 chance of being selected at random, and it undergoes this risk 100 times. If we think of this as a Bernoulli trial, then R_T is a binomial random variable with probability of being selected 1/N over Ng trials:

$$R_T \sim B(Ng, \frac{1}{N})$$

In general, for n trials and probability of p, the distribution K of the count of occurrences has the probability mass function (pmf) of:

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Owing to the memorylessness property of the garbling process, it does not matter if a record is swapped once, twice or ten times: it is subject to the same probability of error. Hence, I am only interested in whether a record is not swapped (K = 0) or swapped (K > 0). In this case, I have n=Ng and p=1/N and K=0:

$$\begin{aligned}\Pr(K = 0) &= \left(1 - \frac{1}{N}\right)^{Ng} \\ &= \left(1 - \frac{g}{Ng}\right)^{Ng}\end{aligned}$$

At this point, I introduce the well-known limiting approximation:

$$\lim_{a \to \infty} \left(1 - \frac{\lambda}{a}\right)^{a} = e^{-\lambda}$$

So that the probability of a record never being selected is:

$$\Pr(K = 0) = e^{-g}$$

And hence the probability of a record being selected more than zero times is:

$$\Pr(K > 0) = 1 - e^{-g}$$

This last quantity is the estimate of the probability that a given customer record is selected "at random" as a target in a swap at least once (ie K > 0). (In effect, the Poisson distribution is used as an approximation to the binomial, with parameter λ = np = Ng/N = g and K=0.)

Going back to my formula for ε:

$$\varepsilon = \gamma \, \mathrm{R_S} + \gamma \, \mathrm{R_T}$$

I have the probability of being selected as a source, $\mathrm{R_S}$ = g. However, if the record is not selected as a source (with probability 1-g), then there is still a chance it will be selected as a target ie Pr(K > 0):

$$\mathrm{R_T} = (1 - \mathrm{g})(1 - e^{-g})$$

Substituting back into the original yields:

$$\varepsilon = \gamma \, \mathrm{g} + \gamma \, (1 - \mathrm{g})(1 - e^{-g})$$

This formula gives the probability that a record is in error for a given extrinsic garble rate, g, and intrinsic attribute statistic of γ.

As a function of g, the error rate ε varies from 0 (when g=0) to a maximum of γ (when g=1). In the example shown below, γ=0.7. The effect of varying γ is to simply the scale the graph linearly.

FIGURE 20 ERROR RATE (E) VS GARBLING RATE (G)

In order to establish that this formula correctly describes the behaviour of the garbling process, the predicted error rates are compared with those observed during the experiments. Since there were 2450 experimental runs, it is not practical to display all of them here. Some examples of the comparison between predicted and observed are provided in this table, followed by a statistical analysis of all experiments.

	a_0		c_0		g_0	
g	*predicted*	*experiment*	*predicted*	*experiment*	*predicted*	*experiment*
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.1	0.1815	0.1838	0.0780	0.0849	0.1287	0.1337
0.2	0.3374	0.3372	0.1449	0.1528	0.2392	0.2262
0.3	0.4707	0.4733	0.2022	0.2133	0.3338	0.3387
0.4	0.5845	0.5833	0.2511	0.2435	0.4145	0.4300
0.5	0.6813	0.6804	0.2926	0.3026	0.4831	0.4797
0.6	0.7631	0.7688	0.3278	0.3470	0.5411	0.5530
0.7	0.8321	0.8338	0.3574	0.3614	0.5901	0.5918
0.8	0.8899	0.8882	0.3823	0.3919	0.6310	0.6300
0.9	0.9380	0.9380	0.4029	0.4229	0.6652	0.6643
1.0	0.9778	0.9786	0.4200	0.4301	0.6934	0.6977

TABLE 18 PREDICTED AND OBSERVED ERROR RATES FOR THREE ATTRIBUTES, A_0, C_0 AND G_0

As expected, there is a close agreement between the predicted number of errors and the number of error events actually observed. Note that the last row (where g=1.0), the error rate ε reaches its maximum of γ, the garble parameter. In order to establish the validity of the previous analysis and resulting formula, all 49 attributes are considered. The analysis hinges on the use of the limit approximation above (as a tends to infinity). In this situation, a = Ng, suggesting that the approximation is weakest when the number of customer records (N) is small or when the garbling rate, g, is close to zero. This can be seen above, where the discrepancy between the predicted and

experimental value is greatest at g=0.1 and for attribute co (the fewest records, at 690). Below is a comparison between the predicted result and experimental result, averaged, for each attribute.

Attribute	Predicted ε (average)	Observed ε (average)	Average Difference	Maximum Difference	Root Mean Square	Correlation Coefficient
a0	0.6656	0.6658	0.0001	0.0057	0.0023	1.0000
a1	0.2933	0.2930	0.0003	0.0047	0.0020	0.9999
a2	0.6671	0.6804	0.0132	0.0199	0.0144	1.0000
a3	0.5512	0.5527	0.0015	0.0049	0.0027	0.9999
a4	0.5506	0.5520	0.0014	0.0038	0.0023	0.9999
a5	0.4523	0.4532	0.0009	0.0000	0.0000	0.9999
a6	0.6049	0.6078	0.0029	0.0073	0.0030	0.9999
a7	0.5002	0.4995	0.0007	0.0062	0.0028	0.9999
a8	0.1729	0.1732	0.0003	0.0024	0.0014	0.9999
a9	0.3022	0.3012	0.0010	0.0045	0.0029	0.9999
a10	0.1065	0.1060	0.0005	0.0009	0.0005	0.9999
a11	0.0626	0.0625	0.0001	0.0012	0.0007	0.9996
a12	0.5178	0.5163	0.0015	0.0055	0.0029	0.9999
a13	0.1043	0.1048	0.0004	0.0019	0.0009	0.9997
c0	0.2859	0.2910	0.0051	0.0200	0.0117	0.9981
c1	0.6668	0.6916	0.0248	0.0470	0.0284	0.9996
c2	0.6646	0.6877	0.0230	0.0287	0.0207	0.9996
c3	0.2487	0.2741	0.0254	0.0436	0.0271	0.9963
c4	0.2487	0.2697	0.0211	0.0360	0.0268	0.9991
c5	0.6100	0.6404	0.0305	0.0514	0.0213	0.9991
c6	0.4056	0.4275	0.0219	0.0289	0.0203	0.9989
c7	0.6560	0.6651	0.0090	0.0364	0.0155	0.9991
c8	0.3396	0.3522	0.0125	0.0250	0.0134	0.9975
c9	0.3332	0.3522	0.0189	0.0339	0.0210	0.9976
c10	0.4434	0.4591	0.0157	0.0306	0.0172	0.9992
c11	0.3380	0.3617	0.0238	0.0227	0.0158	0.9989
c12	0.1175	0.1342	0.0167	0.0218	0.0177	0.9975
c13	0.6355	0.6622	0.0267	0.0446	0.0256	0.9994
c14	0.5487	0.5688	0.0202	0.0430	0.0190	0.9989
g0	0.4720	0.4688	0.0032	0.0155	0.0080	0.9991
g1	0.6092	0.5994	0.0098	0.0093	0.0058	0.9997
g2	0.4231	0.4264	0.0033	0.0147	0.0088	0.9986
g3	0.5519	0.5594	0.0075	0.0156	0.0099	0.9995
g4	0.6671	0.6845	0.0173	0.0281	0.0173	0.9997
g5	0.3989	0.4061	0.0072	0.0111	0.0070	0.9996
g6	0.5156	0.5158	0.0002	0.0135	0.0078	0.9994
g7	0.4608	0.4591	0.0018	0.0099	0.0060	0.9995
g8	0.4034	0.4011	0.0023	0.0105	0.0060	0.9994
g9	0.1177	0.1178	0.0001	0.0049	0.0026	0.9986
g10	0.4734	0.4659	0.0076	0.0077	0.0041	0.9996
g11	0.4988	0.4967	0.0021	0.0074	0.0045	0.9997
g12	0.6595	0.6574	0.0021	0.0101	0.0050	0.9998
g13	0.2150	0.2134	0.0016	0.0087	0.0053	0.9985
g14	0.3049	0.3097	0.0048	0.0113	0.0068	0.9989
g15	0.3319	0.3317	0.0002	0.0088	0.0050	0.9993
g16	0.3681	0.3654	0.0027	0.0093	0.0041	0.9998
g17	0.1783	0.1722	0.0062	0.0075	0.0042	0.9995

g18	0.3278	0.3313	0.0035	0.0139	0.0066	0.9987
g19	0.0485	0.0488	0.0003	0.0034	0.0017	0.9961

TABLE 19 COMPARING EXPECTED AND PREDICTED ERROR RATES

Here, the average error rates across all levels of g are shown in the second and third columns. The fourth column is the difference between these two figures. In absolute terms, the difference in averages ranges from 0.01% (a11) to 2.02% (c14). However, comparing averages doesn't tell the whole story. To understand what's happening between the predicted and observed values at each level of g, I can use the RMS (root mean square) difference measure. Commonly used for such comparisons, this involves squaring the difference, taking the mean of those values and then taking the square root. This is a better measure, since it takes into account differences at the smaller values of g rather than rolling them up as averages. Again, there are small differences found between the predicted and observed (ranging from 0.0000 up to 0.0284 with a mean of 0.0095), suggesting a close fit between the two. The last column shows the pairwise Pearson correlation coefficient for each attribution, indicating a very strong correlation between predicted and observed values. (The correlations were all highly significant to at least 10^{-9}).

Lastly, to check the "worst case", the maximum difference is reported for each attribute. The biggest gap (0.0514) in all 490 cases occurs for attribute c5 (at g=0.2, specifically), where the predicted value is 0.3091 and the observed value is 0.3606. Note that, as suggested by the limiting approximation, this occurs for the dataset with the fewest customer records and a small value of g.

This comparison of the predicted and observed error rates shows that, on average, the formula derived from mathematical analysis is a very close approximation, with an expected RMS discrepancy less than 1% and an expected correlation of 0.9992. Furthermore, the "worst case" check provides confidence that the experimental procedure was conducted correctly.

6.5.1.1 RELATIONSHIP TO FIDELITY

This section has shown that the *error rate* associated with an attribute subject to garbling noise can be estimated mathematically using a simple formula. This formula, derived above from first principles, relies on two quantities: g, which is the garbling parameter of the noise process and γ which is an intrinsic statistical property of the attribute.

The theoretical framework developed in Chapter 5 proposed the use of the *fidelity metric, φ, t*o quantify the effect of noise on an attribute. This role is replaced by g and γ in the experiments, since g can be continuously varied (controlled) to produce the desired level of errors. Their relationship with the more general φ metric is illustrated through Fano's Inequality (Cover and Thomas 2005), which links the error rate with the equivocation for a noisy channel. Recall the definition of φ, for the external world state W and IS state X:

$$\varphi = 1 - \frac{H(W|X)}{H(W)}$$

We see that the fidelity improves as the equivocation, H(W|X), decreases. When H(W|X)=0 the fidelity is maximised at 100%. Fano's Inequality bounds the equivocation for a given error rate ∈:

$$H(W|X) \leq H(\epsilon) + \epsilon \log N - 1$$

Where N is the number of states in W and H(∈) is the binary entropy function of ∈:

$$H(\epsilon) = -\epsilon \log \epsilon - (1-\epsilon) \log 1-\epsilon$$

As ϵ is a function solely of g and γ and H(W) and log(N-1) are constant, an increase in φ must result from a decrease in either g or γ. For a given attribute, γ is fixed, so changes in g yield an opposite change in φ. In this way, we can see that the garbling noise process constrains a particular model on fidelity so we can describe it as a non-linear function of g, $\varphi(g)$. (Note that this represents a lower limit of the fidelity, as φ is constrained by the inequality.)

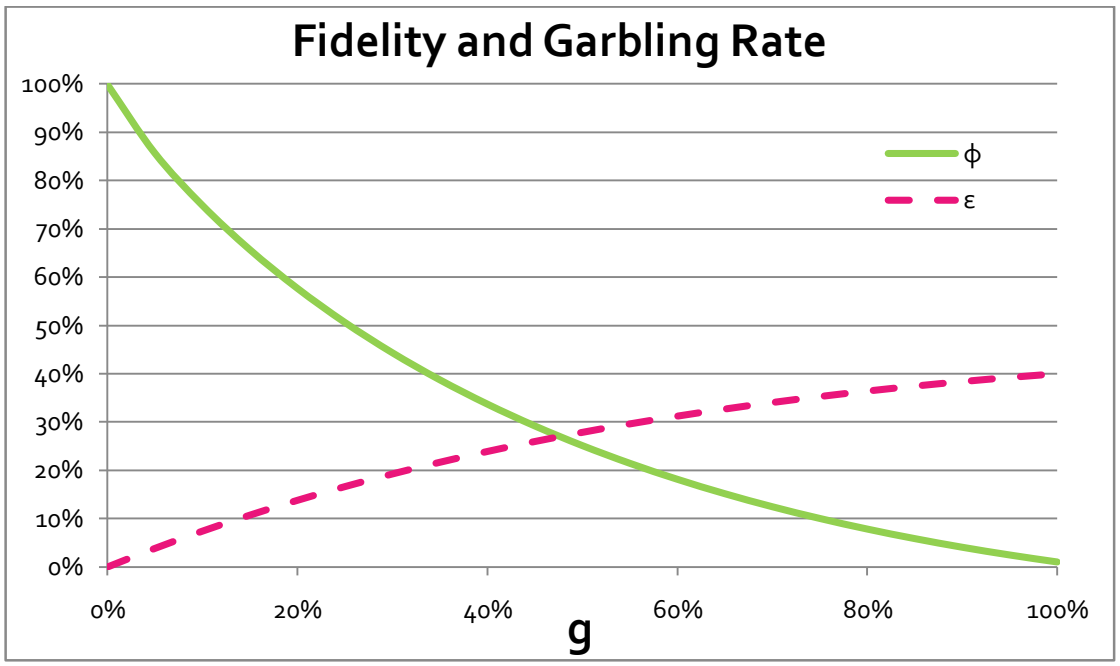

FIGURE 21 EFFECT OF GARBLING RATE ON FIDELTY

Above, Figure 21 shows the effect of varying the garbling rate, g, from 0% to 100%. The broken line shows ϵ reaching its maximum at 40% (as the value of γ in this example is 0.4). The unbroken line shows $\varphi(g)$ falling from 100% (when g=0%) to 2% (when g=100%). The value for H(W) is 3 and the value for N = 33. This illustrates that fidelity decreases non-linearly as the garbling rate, g, increases and the error rate, ϵ, increases.

In general, I can expect different kinds of noise processes to impact on fidelity in different ways. While the garbling noise process used here is amenable to this kind of closed-form algebraic analysis, other types may require a numerical estimation approach.

However, I can always compute the γ for a given attribute and estimate the error rate, ε, by direct observation. In such cases, I can use the formula to derive an "effective garbling rate", g_{eff}. For example, suppose there are two attributes, W_1 and W_2 with observed error rates of ε_1=0.05 and ε_2=0.1, respectively. Further, their gamma levels are measured at γ_1=0.40 and γ_2=0.7. Their garbling rates can be read from the above charts as follows. For W_1 I use Figure 21 (where γ=0.40) and see that ε=0.05 corresponds to g_{eff} =0.06. For W_2 I use Figure 20 (where γ=0.7) and read off a value of g_{eff}=0.07 for ε_2=0.1. This illustrates a situation where the underlying effective garbling rates are almost the same, but the error rate is twice as bad for the second attribute since its γ value is so much larger.

The interpretation of the effective garbling rate is that it quantifies the number of customer records impacted by a quality deficiency, as distinct from the ones actually in error. When detecting and correcting impacted records is expensive, understanding the likelihood that an individual impacted record will translate into an error is useful for comparing competing attributes.

6.5.2 Effects on Mistakes

The next metric to define is dubbed alpha (α), which describes the *actionability* of an attribute. This is the probability that an error on that attribute will result in a mistake. Recall that a mistake is a misclassification or "incorrect decision" when compared the relevant baseline decision set. This metric is in the range of 0 to 1, where 0 means that no changes to that attribute will change the decision whereas 1 means every single change in attribute value will change the decision.

Attribute	ID3	AD	NB	BNet	LMT	Average
a_0	0%	4%	3%	8%	5%	**4%**
a_1	0%	0%	4%	4%	4%	**2%**
a_2	0%	0%	0%	0%	1%	**0%**
a_3	0%	0%	7%	7%	2%	**3%**
a_4	0%	16%	7%	6%	13%	**9%**
a_5	1%	16%	18%	15%	16%	**13%**
a_6	0%	4%	5%	7%	7%	**5%**
a_7	0%	0%	5%	14%	4%	**5%**
a_8	0%	0%	2%	5%	2%	**2%**
a_9	0%	0%	1%	8%	1%	**2%**
a_{10}	54%	39%	27%	14%	35%	**34%**
a_{11}	44%	17%	16%	16%	21%	**23%**
a_{12}	0%	2%	3%	8%	5%	**3%**
a_{13}	0%	0%	5%	6%	7%	**4%**
c_0	0%	0%	2%	0%	0%	**0%**
c_1	0%	0%	1%	1%	4%	**1%**
c_2	1%	1%	7%	2%	4%	**3%**
c_3	1%	6%	7%	3%	5%	**4%**
c_4	0%	0%	2%	2%	0%	**1%**
c_5	0%	1%	2%	4%	5%	**3%**
c_6	0%	0%	2%	4%	4%	**2%**
c_7	0%	6%	1%	4%	2%	**3%**
c_8	98%	59%	64%	21%	54%	**59%**
c_9	0%	14%	4%	8%	7%	**7%**
c_{10}	0%	0%	2%	8%	10%	**4%**
c_{11}	0%	0%	7%	0%	0%	**1%**
c_{12}	0%	0%	14%	4%	5%	**5%**
c_{13}	0%	3%	2%	5%	10%	**4%**
c_{14}	0%	4%	6%	4%	4%	**4%**
g_0	44%	23%	18%	21%	22%	**26%**
g_1	4%	17%	9%	8%	12%	**10%**
g_2	20%	16%	11%	14%	13%	**15%**
g_3	0%	13%	8%	9%	9%	**8%**
g_4	2%	0%	4%	5%	12%	**5%**
g_5	5%	19%	28%	12%	10%	**15%**
g_6	0%	0%	6%	7%	6%	**4%**
g_7	1%	0%	0%	0%	5%	**1%**
g_8	0%	0%	6%	5%	9%	**4%**
g_9	0%	0%	8%	10%	9%	**6%**
g_{10}	1%	0%	0%	0%	0%	**0%**
g_{11}	2%	3%	6%	8%	7%	**5%**
g_{12}	0%	0%	0%	0%	10%	**2%**
g_{13}	0%	1%	8%	10%	12%	**6%**
g_{14}	0%	0%	11%	11%	7%	**6%**

g_{15}	0%	0%	1%	0%	5%	**1%**
g_{16}	0%	0%	3%	3%	1%	**1%**
g_{17}	0%	0%	0%	0%	1%	**0%**
g_{18}	0%	0%	1%	3%	7%	**2%**
g_{19}	0%	0%	7%	15%	10%	**6%**

TABLE 20 ALPHA BY ATTRIBUTE AND DECISION FUNCTION

Values for α range from 0 (errors lead to no mistakes) to a maximum of 98% (for c_8 using ID_3). This means that an error on attribute c_8 will, in 98% of cases, result in a changed decision. Upon inspection of the α values for other attributes in the CRX dataset using ID_3, I can see they are nearly all zero. This indicates that ID_3, in this case, nearly entirely relies on c_8 to make its decision. Since c_8 is a binary valued attribute, it is not surprising that almost any change in the attribute will change the decision.

Other algorithms are not so heavily reliant on just one attribute; Naïve Bayes (NB) and the Logistic Model Tree (LMT), for example, draw on more attributes as can be seen by their higher α values across the range of attributes. Despite these variations, there is broad agreement between the algorithms about which attributes have highest values of α. In general, each dataset has one or two dominating attributes and a few irrelevant (or inconsequential) attributes, regardless of the specific algorithm used. It is to be expected that there would be few irrelevant attributes included in the dataset: people would not go to the expense of sourcing, storing and analysing attributes that had no bearing on the decision task at hand. In this way, only candidate attributes with a reasonable prospect of being helpful find their way into the datasets.

The similarity between observed α values for attributes across different decision functions is not a coincidence. There are underlying patterns in the data that are discovered and exploited by the algorithms that generate these decision functions. There are patterns in the data because these data reflect real-world socio-demographic and socio-economic phenomena.

So the ultimate source of these patterns lies in the external social world: high-income people, for instance, tend to be older or more highly-educated or live in certain post codes. At the level of *the real*, there are generative mechanisms being triggered resulting in observable customer events (*the actual*). These events are encoded as customer data in the system. The algorithms then operate on these data to produce rules (decision functions) that replicate the effect of the generative mechanism in the external social world. However, the generative mechanism for the system bears no resemblance to the external social world, as it is an artefact composed of silicon, software and formulae operating according to the laws of natural science.

As long as the system's generative mechanism (hardware and software) operates correctly, any sufficiently "good" learning algorithm will detect these patterns and derive rules that can use them. In other words, the capacity of a system to detect and implement the underlying patterns (thereby replicating events in the external social world) is constrained by the properties of the patterns themselves, not the mechanisms of the system.

The theoretical framework developed in Chapter 5 proposes a quantitative measure of the relation between error events and mistakes for an attribute used in a decision task: *influence*. Recall that this entropy-based measure was defined as the normalised mutual information between the attribute, X and the decision, Y:

$$Influence = \frac{I(X;Y)}{H(Y)} = \frac{H(Y) - H(Y|X)}{H(Y)}$$

$$= 1 - \frac{H(Y|X)}{H(Y)}$$

Informally, influence can be described in terms of changes to decision uncertainty: before a decision is made, there is a particular amount of uncertainty about the final decision, given by H(Y). Afterwards, there is o uncertainty (a particular decision is definitively selected). However, in between, suppose just one attribute has its value revealed. In that case *some* uncertainty about the final decision is removed. The exact amount depends on which value of the attribute arises, but it can be averaged over all possible values. Hence, each attribute will have its own influence score on each particular decision task.

This quantity has a convenient and intuitive interpretation in the context of machine learning, predictive analytics and data mining: *information gain ratio (*Kononenko and Bratko 1991). The *information gain* is the incremental amount of uncertainty about the classification Y removed upon finding that an attribute X takes a particular value, X=x. Formally, it is the Kullback-Leibler divergence of the posterior distribution p(Y|X=x) and the prior distribution p(Y):

$$IG = D_{KL}(p(Y|X = x)||p(Y))$$

If I take the expected value over all possible values of x, I have:

$$\begin{aligned} \boldsymbol{E}[IG] &= \sum_{x} D_{KL}(p(Y|X = x)||p(Y)) \\ &= D_{KL}(p(X,Y)||p(X)p(Y)) \\ &= I(Y;X) \end{aligned}$$

This quantity is used frequently in data mining and machine learning to select subsets of attributes Yao (Yao et al. 1999) and performance evaluation (Kononenko and Bratko 1991). In practice, the related quantity of the *information gain ratio* is used, where the information gain is divided by the intrinsic amount of information in the attribute ie H(X). This is done to prevent very high gain scores being assigned to an attribute that takes on a large number of values. For example, compared with gender, a customer's credit card number will uniquely identify them (and hence tell you precisely what the decision will be). However, a credit card has approximately 16 * $\log_2 10$ bits (53 bits) whereas gender has approximately 1 bit. Information gain ratio will scale accordingly.

$$IGR = \frac{IG}{H(X)}$$

This discussion provides the motivation for examining how the observed actionability of each attribute, as arising during the experiments, aligns with the entropy-based measure of influence. In general, I would expect influential attributes to be prone to actionable errors. Conversely, a low influence score should have low actionability (so that no errors lead to mistakes).

Before looking into this, it's worth considering the importance of finding such a relationship. From a theoretical perspective, it would mean that the underlying generative mechanisms (in the domain of *the real*) that give rise to customer behaviours and events (at *the actual*) are being replicated, in some sense, within the information system. That is, the algorithms that construct the decision functions are picking up on and exploiting these persistent patterns while the system itself is operating correctly and implementing these functions. The degree of agreement between the two measures indicates the success of the system in "mirroring" what is happening in the external world.

At a practical level, if the influence score is an adequate substitute or proxy for actionability, then the question of how to design and appraise IQ improvement interventions becomes more tractable. The whole effort of generating some noise process, applying it to each attribute, comparing the output to the benchmark and then re-running it multiple times is avoided. In general, I might expect such experiments in real-world organisational settings to be time-consuming, fraught with error and disruptive to normal operations.

Perhaps more significantly, the experimental approach requires an existing decision function to be in place and (repeatedly) accessible. Using the influence score, only the ideal decision values are required, meaning that the analysis could proceed before a system exists. This could be very useful during situations such as project planning. Further, it could be used when access to the decision function is not possible, as when it is proprietary or subject to other legal constraints. This is further explored during Chapter 7.

The analysis proceeds by examining the information gain for each attribute across the three decision tasks (for a total of 49 attributes) and five decision functions. The seventh column shows the average information gain for the five decision functions. The last column, Z, shows the "true information gain". This is calculated by using the correct external world decision (ie training data) rather than the outputs of any decision function.

Attribute	ID3	AD	NB	BNet	LMT	Average	Z
a_0	0.0184	0.0667	0.0770	0.1307	0.0737	0.0733	0.0784
a_1	0.0073	0.0225	0.0326	0.0451	0.0314	0.0278	0.0186
a_2	0.0003	0.0005	0.0008	0.0013	0.0006	0.0007	0.0004
a_3	0.0349	0.1956	0.1905	0.1446	0.1683	0.1468	0.0884
a_4	0.0178	0.1810	0.1439	0.0833	0.1367	0.1126	0.0415
a_5	0.0271	0.1664	0.2058	0.3359	0.1664	0.1803	0.1599
a_6	0.0204	0.1160	0.1206	0.1337	0.1189	0.1019	0.0683
a_7	0.0288	0.1674	0.2086	0.3612	0.1699	0.1872	0.1693
a_8	0.0012	0.0082	0.0115	0.0214	0.0074	0.0099	0.0088
a_9	0.0056	0.0350	0.0390	0.1045	0.0318	0.0432	0.0333
a_{10}	0.1779	0.1206	0.0847	0.0565	0.1010	0.1081	0.0813
a_{11}	0.0639	0.0262	0.0205	0.0138	0.0234	0.0296	0.0209
a_{12}	0.0137	0.0392	0.0533	0.0916	0.0516	0.0499	0.0428
a_{13}	0.0040	0.0091	0.0133	0.0151	0.0139	0.0111	0.0102
c_0	0.0003	0.0017	0.0026	0.0047	0.0003	0.0019	0.0004
c_1	0.0287	0.0358	0.0257	0.0283	0.0204	0.0278	0.0211
c_2	0.0412	0.0472	0.0525	0.0641	0.0462	0.0502	0.0394
c_3	0.0177	0.0311	0.0376	0.0378	0.0305	0.0309	0.0296
c_4	0.0177	0.0311	0.0376	0.0378	0.0305	0.0309	0.0296
c_5	0.0813	0.0944	0.1079	0.1442	0.1128	0.1081	0.1092
c_6	0.0558	0.0513	0.0492	0.0702	0.0460	0.0545	0.0502
c_7	0.1103	0.1771	0.1170	0.1428	0.1101	0.1314	0.1100
c_8	0.9583	0.6159	0.5133	0.4979	0.4404	0.6052	0.4257
c_9	0.1428	0.2615	0.1998	0.3151	0.1785	0.2195	0.1563
c_{10}	0.1959	0.2729	0.2049	0.3207	0.2023	0.2393	0.2423
c_{11}	0.0057	0.0036	0.0021	0.0023	0.0003	0.0028	0.0007
c_{12}	0.0180	0.0264	0.0150	0.0395	0.0125	0.0223	0.0100
c_{13}	0.0099	0.0293	0.0149	0.0204	0.0156	0.0180	0.2909
c_{14}	0.0004	0.1198	0.1203	0.1439	0.1084	0.0985	0.1102
g_0	0.3803	0.2227	0.1367	0.2068	0.1748	0.2243	0.0947
g_1	0.0072	0.0964	0.0494	0.0842	0.0369	0.0548	0.0140

g_2	0.1155	0.0944	0.0840	0.1052	0.0758	0.0950	0.0436
g_3	0.0184	0.0485	0.0454	0.0691	0.0307	0.0424	0.0249
g_4	0.0206	0.0257	0.0338	0.0648	0.0264	0.0343	0.0187
g_5	0.0224	0.0866	0.0525	0.0577	0.0471	0.0532	0.0281
g_6	0.0063	0.0083	0.0229	0.0386	0.0119	0.0176	0.0131
g_7	0.0000	0.0026	0.0001	0.0000	0.0000	0.0005	0.0030
g_8	0.0018	0.0034	0.0209	0.0182	0.0187	0.0126	0.0068
g_9	0.0010	0.0042	0.0060	0.0105	0.0058	0.0055	0.0048
g_{10}	0.0008	0.0000	0.0000	0.0063	0.0001	0.0015	0.0000
g_{11}	0.0108	0.0195	0.0720	0.0943	0.0230	0.0439	0.0170
g_{12}	0.0043	0.0062	0.0022	0.0048	0.0187	0.0072	0.0107
g_{13}	0.0014	0.0050	0.0132	0.0195	0.0090	0.0096	0.0089
g_{14}	0.0104	0.0070	0.0515	0.0737	0.0204	0.0326	0.0128
g_{15}	0.0039	0.0008	0.0020	0.0005	0.0003	0.0015	0.0015
g_{16}	0.0023	0.0052	0.0171	0.0259	0.0019	0.0105	0.0013
g_{17}	0.0007	0.0002	0.0014	0.0002	0.0000	0.0005	0.0000
g_{18}	0.0008	0.0000	0.0013	0.0015	0.0062	0.0020	0.0010
g_{19}	0.0010	0.0013	0.0050	0.0083	0.0050	0.0041	0.0058

TABLE 21 INFORMATION GAINS BY ATTRIBUTE AND DECISION FUNCTION

As expected, there is broad agreement between the different decision functions as to how much information can be extracted from each attribute. The gain ranges from effectively zero (eg a_2) through to 0.95 (eg c_8 with ID3). The attributes with high gains also show some differences in how the decision functions are able to exploit information: a_3, c_8 and g_0, for example, show considerable variation from the average.

When comparing the "true information gain" with the average for the five decision functions, there is also broad agreement, with a Pearson correlation co-efficient, ρ=0.8483 (highly significant to at least 10^{-14}). This suggests that, by and large, the decision functions are effective at detecting and using all the available or "latent information" in each attribute. However, some notable exceptions are a_4, c_{13} and g_0. It may be that other algorithms for building decision functions could better tap into this information.

Now I can examine how well the information gain works as a proxy or substitute for actionability, α. To do this, the Pearson correlation co-efficient, ρ, is used to gauge how closely they are in step. All results are significant at <0.01 unless otherwise reported.

ρ	ID3	AD	NB	BNet	LMT	Average	Z
a	0.8898	0.3212	0.2983	0.5233	0.1799	0.2247	0.2006
c	0.9704	0.8938	0.8202	0.9191	0.8773	0.9057	0.7439
g	0.9847	0.9215	0.6707	0.7510	0.8080	0.9266	0.9320
ALL	0.8698	0.7367	0.6724	0.5817	0.6414	0.7459	0.5678

TABLE 22 CORRELATION BETWEEN INFORMATION GAIN AND ACTIONABILITY, BY DATASET AND DECISION FUNCTION

Note that here the average column contains the correlations between the average information gain and the average actionability (where the averaging is done over all five decision functions). It is not the average of the correlations. As before, Z refers to the "true information gain" when using the correct decisions in the dataset as the benchmark. The rows describe the datasets (Adult, CRX and German, respectively) and ALL describes the correlation when all 49 attributes are considered collectively.

The correlation coefficients (ρ) range from 0.18 to to 0.98, averaging around 0.70. This constitutes a moderate to strong positive correlation, but there is significant variability. Using information gain instead of actionability in the case of the Adult dataset would, regardless of decision function, result in prioritising different attributes. In the German dataset, the deterioration would be much less pronounced.

Based on the widespread use of the *information gain ratio* (IGR) in practice, this measure was evaluated in an identical fashion, to see if it would make a better proxy. As explained above, it is computed by dividing the information gain by the amount of information in the attribute, H(X). The following table was obtained:

Attribute	*H(X)*	ID3	AD	NB	BNet	LMT	Average	Z
a_0	*5.556*	0.33%	1.20%	1.39%	2.35%	1.33%	1.32%	1.41%
a_1	*1.392*	0.53%	1.61%	2.34%	3.24%	2.25%	1.99%	1.34%
a_2	*5.644*	0.01%	0.01%	0.01%	0.02%	0.01%	0.01%	0.01%
a_3	*2.928*	1.19%	6.68%	6.51%	4.94%	5.75%	5.01%	3.02%
a_4	*2.867*	0.62%	6.31%	5.02%	2.91%	4.77%	3.93%	1.45%
a_5	*1.852*	1.46%	8.98%	11.11%	18.14%	8.99%	9.74%	8.63%
a_6	*3.360*	0.61%	3.45%	3.59%	3.98%	3.54%	3.03%	2.03%
a_7	*2.161*	1.33%	7.75%	9.65%	16.71%	7.86%	8.66%	7.83%
a_8	*0.783*	0.15%	1.04%	1.47%	2.73%	0.94%	1.27%	1.12%
a_9	*0.918*	0.61%	3.81%	4.25%	11.39%	3.47%	4.71%	3.63%
a_{10}	*0.685*	25.99%	17.62%	12.37%	8.25%	14.75%	15.80%	11.88%
a_{11}	*0.480*	13.31%	5.45%	4.28%	2.87%	4.88%	6.16%	4.36%
a_{12}	*3.269*	0.42%	1.20%	1.63%	2.80%	1.58%	1.53%	1.31%
a_{13}	*0.761*	0.52%	1.19%	1.74%	1.98%	1.83%	1.45%	1.33%
c_0	*0.881*	0.04%	0.19%	0.29%	0.54%	0.03%	0.22%	0.05%
c_1	*5.627*	0.51%	0.64%	0.46%	0.50%	0.36%	0.49%	0.38%
c_2	*5.505*	0.75%	0.86%	0.95%	1.16%	0.84%	0.91%	0.72%
c_3	*0.816*	2.17%	3.81%	4.60%	4.63%	3.73%	3.79%	3.63%
c_4	*0.816*	2.17%	3.81%	4.60%	4.63%	3.73%	3.79%	3.63%
c_5	*3.496*	2.32%	2.70%	3.09%	4.12%	3.23%	3.09%	3.12%
c_6	*1.789*	3.12%	2.87%	2.75%	3.92%	2.57%	3.05%	2.80%
c_7	*5.133*	2.15%	3.45%	2.28%	2.78%	2.14%	2.56%	2.14%
c_8	*0.998*	95.98%	61.69%	51.41%	49.86%	44.11%	60.61%	42.64%
c_9	*0.985*	14.50%	26.55%	20.28%	32.00%	18.12%	22.29%	15.87%
c_{10}	*2.527*	7.75%	10.80%	8.11%	12.69%	8.00%	9.47%	9.59%
c_{11}	*0.995*	0.57%	0.36%	0.21%	0.23%	0.03%	0.28%	0.07%
c_{12}	*0.501*	3.60%	5.28%	2.99%	7.89%	2.49%	4.45%	2.00%
c_{13}	*4.744*	0.21%	0.62%	0.31%	0.43%	0.33%	0.38%	6.13%
c_{14}	*3.911*	0.01%	3.06%	3.08%	3.68%	2.77%	2.52%	2.82%
g_0	*1.802*	21.10%	12.36%	7.59%	11.48%	9.70%	12.44%	5.26%
g_1	*3.726*	0.19%	2.59%	1.32%	2.26%	0.99%	1.47%	0.38%
g_2	*1.712*	6.75%	5.51%	4.91%	6.14%	4.43%	5.55%	2.55%
g_3	*2.667*	0.69%	1.82%	1.70%	2.59%	1.15%	1.59%	0.93%
g_4	*5.643*	0.36%	0.46%	0.60%	1.15%	0.47%	0.61%	0.33%
g_5	*1.688*	1.33%	5.13%	3.11%	3.42%	2.79%	3.15%	1.67%
g_6	*2.155*	0.29%	0.38%	1.06%	1.79%	0.55%	0.82%	0.61%
g_7	*1.809*	0.00%	0.14%	0.00%	0.00%	0.00%	0.03%	0.16%
g_8	*1.532*	0.12%	0.22%	1.36%	1.19%	1.22%	0.82%	0.44%
g_9	*0.538*	0.18%	0.78%	1.11%	1.95%	1.07%	1.02%	0.89%
g_{10}	*1.842*	0.04%	0.00%	0.00%	0.34%	0.01%	0.08%	0.00%

g_{11}	*1.948*	0.55%	1.00%	3.70%	4.84%	1.18%	2.25%	0.87%
g_{12}	*5.226*	0.08%	0.12%	0.04%	0.09%	0.36%	0.14%	0.20%
g_{13}	*0.845*	0.17%	0.59%	1.56%	2.31%	1.06%	1.14%	1.05%
g_{14}	*1.139*	0.91%	0.61%	4.52%	6.47%	1.80%	2.86%	1.12%
g_{15}	*1.135*	0.35%	0.07%	0.18%	0.04%	0.03%	0.13%	0.13%
g_{16}	*1.413*	0.16%	0.37%	1.21%	1.83%	0.13%	0.74%	0.09%
g_{17}	*0.622*	0.11%	0.04%	0.23%	0.04%	0.01%	0.08%	0.00%
g_{18}	*0.973*	0.08%	0.00%	0.13%	0.15%	0.63%	0.20%	0.10%
g_{19}	*0.228*	0.44%	0.57%	2.19%	3.62%	2.21%	1.81%	2.55%

TABLE 23 INFORMATION GAIN RATIO BY ATTRIBUTE AND DECISION FUNCTION

As suggested by its name, the information gain ratio is expressed as a percentage. A value of 100% for a particular attribute implies that that attribute wholly governs the operation of the decision function. In other words, once the value of that one attribute is known, there is no longer any uncertainty about the decision. Since each row in this table is simply the values of the last table divided through by a constant H(X) (in italics, second column), there is the same broad agreement between decision functions and with the "true information gain ratio", labelled Z.

From these data, the correlation table was produced, showing the direction and degree of agreement between the information gain ratio and the actionability for each attribute:

ρ	ID3	AD	NB	BNet	LMT	Average	Z
a	0.9704	0.8379	0.7885	0.6558	0.7802	0.8181	0.7558
c	0.9875	0.9620	0.9040	0.9268	0.9291	0.9576	0.9506
g	0.9856	0.8712	0.7149	0.8843	0.7906	0.9274	0.8992
ALL	0.9138	0.8298	0.8194	0.5871	0.8049	0.8524	0.8110

TABLE 24 CORRELATION BETWEEN INFORMATION GAIN RATIO AND ACTIONABILITY, BY DATASET AND DECISION FUNCTION

All results are statistically significant at <0.001. The correlation coefficients range from 0.59 (looking at all 49 attributes at once when using the BNet decision function) up to 0.99 (using ID3 on the GERMAN dataset). This is a much more robust range than for information gain. Of the 28 cells in the table (each corresponding to a different slice of the data, for evaluation purposes), 12 have a correlation coefficient >0.90 while only three have a value <0.75.

This analysis indicates that the information gain ratio is a better substitute or proxy for actionability than the un-normalised information gain. The interpretation is that errors in attributes with a high information gain ratio are more likely to result in mistakes than attributes with a low information gain ratio. The reason is that an error in an attribute with a large amount of entropy H(X) (such as a continuous-valued attribute) is likely to affect only a small proportion of cases and hence is less likely to be exploited by an algorithm when building the decision model.

When ranking attributes, the use of the IGR is particularly effective at screening out irrelevant (low gain) attributes and prioritising high-gain ones. The table below compares the IGR rank (from first to last) of each attribute when sorted by actionability, α.

Rank	IGR				IG		
	a	**c**	**g**		**a**	**c**	**g**
1	1	1	1		5	1	1
2	4	2	3		10	3	4
3	2	4	2		2	12	2
4	7	5	8		4	9	3
5	8	3	7		6	2	6
6	3	13	6		1	13	15
7	12	10	9		7	6	12
8	11	11	4		12	8	8
9	10	9	10		8	4	14
10	5	7	5		3	5	5
11	9	8	14		11	7	7
12	6	14	11		9	14	10
13	13	12	12		13	11	9
14	14	6	16		14	10	13
15		15	15			15	16
16			13				11
17			20				19
18			17				17
19			18				20
20			19				18
Spearman ρ	*0.71*	*0.71*	*0.92*		*0.60*	*0.53*	*0.82*

TABLE 25 RANKINGS COMPARISON

If we look at the top and bottom ranked attributes for each dataset, we see that IGR correctly identifies the most and least actionable attributes for ADULT and CRX. For GERMAN, IGR picks up the most actionable and places the second-least actionable last. In this sense, IGR out-performs the un-normalised information gain, IG.

The last row shows the Spearman rank correlation coefficient for the three datasets. This is a non-parametric statistic that measures how closely the attributes are ranked when using actionability as compared with IGR (or IG). All results are significant at <0.05. These results indicate a strong relationship, and that in each case, IGR outperforms IG. Note that in raw terms, this statistic is a little misleading in this context, since it "penalises" a mis-ranking at the top end the same as a mis-ranking in the middle. That is, mixing up the 7^{th}- and 11^{th}-ranked attribute is penalised as heavily as mixing the 1^{st}-and 5^{th}-ranked, even though in a business context the second situation is likely to be worse. Worse still, rankings are highly sensitive to slight variations; the 11^{th} through 15^{th} ranked attributes may differ by as little as 1%, distorting the effect of mis-ranking.

A better way to visualise the performance of substitution of the IGR for actionability is the "percent cumulative actionability capture" graph, below. The idea is that for a given scenario, there is a total amount of actionability available for "capture" (obtained by summing the actionability scores, α, for each attribute). Obviously, using the actionability scores will give the best results, ie selecting the attributes in the sequence that will yield the greatest actionability.

But actionability isn't evenly distributed amongst the attributes: The top one or two attributes contribute a disproportionately large amount, while the bottom few contribute a small proportion. Selecting the top, say, three attributes may yield 50% of the total available actionability. Selecting 100% of the attributes will capture 100% of the actionability.

When I rank the attributes by actionability, the plot of the percent cumulative actionability as a function of the number of attributes selected represents the "best case" ie directly using actionablity scores to select attributes. If I repeat the exercise but this time ranking the attributes by a proxy measure (in this case, IGR), I get another curve. In general, using IGR instead of α will result in a slightly different order of selection of attributes (as shown in Table 13). The area between these two curves represents the "lost value" in using the proxy measure in lieu of actionability. The "worst case" – corresponding to just picking attributes at random – would be a straight line at a 45° angle.

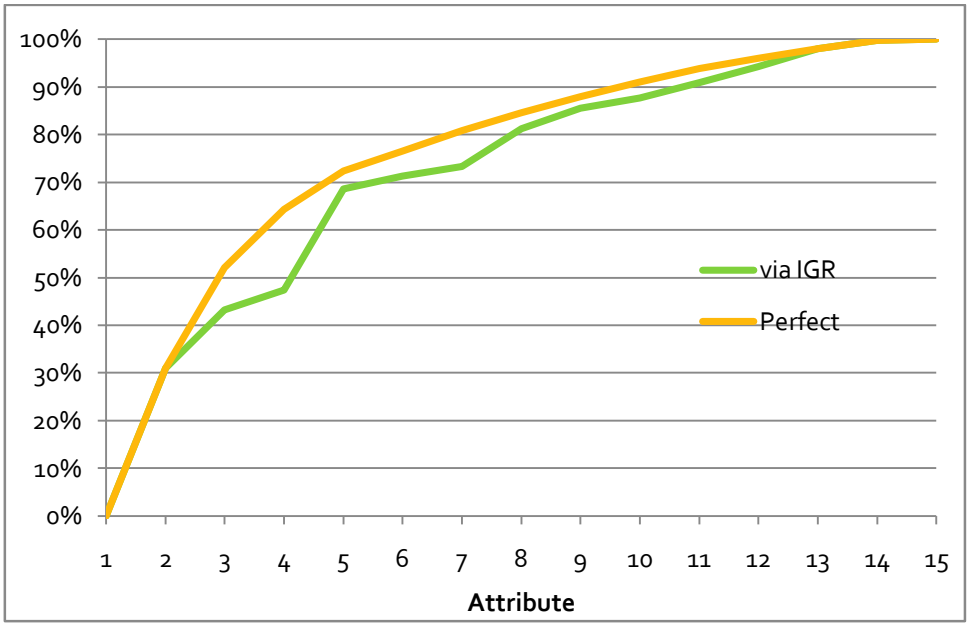

FIGURE 22 PERCENT CUMULATIVE ACTIONABILITY FOR ADULT DATASET

Compared with the other datasets (below) there is a larger gap between the perfect case (using α scores) and using IGR. The gap reaches a maximum at the 4th-ranked attribute (a 16% point discrepancy) and the average gap across the dataset is 3.7% points.

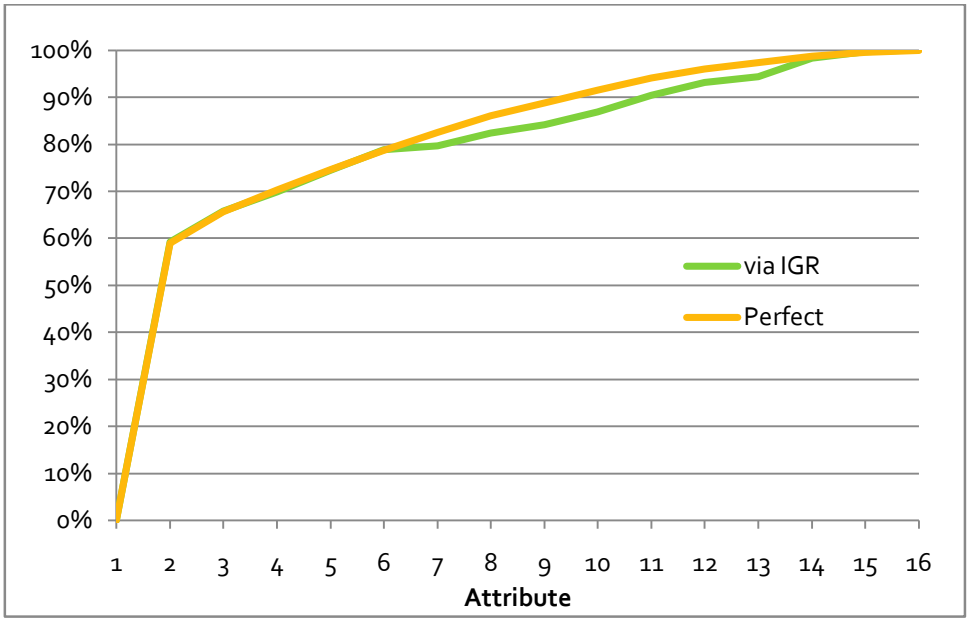

FIGURE 23 PERCENT CUMULATIVE ACTIONABILITY FOR CRX DATASET

For the CRX dataset, ranking the attributes by IGR instead of α results in little loss of actionability until the 6th-ranked attribute. The gap reaches a maximum at the 9th-ranked attribute (4.6% points) and the average shortfall across all attributes is 1.7% points.

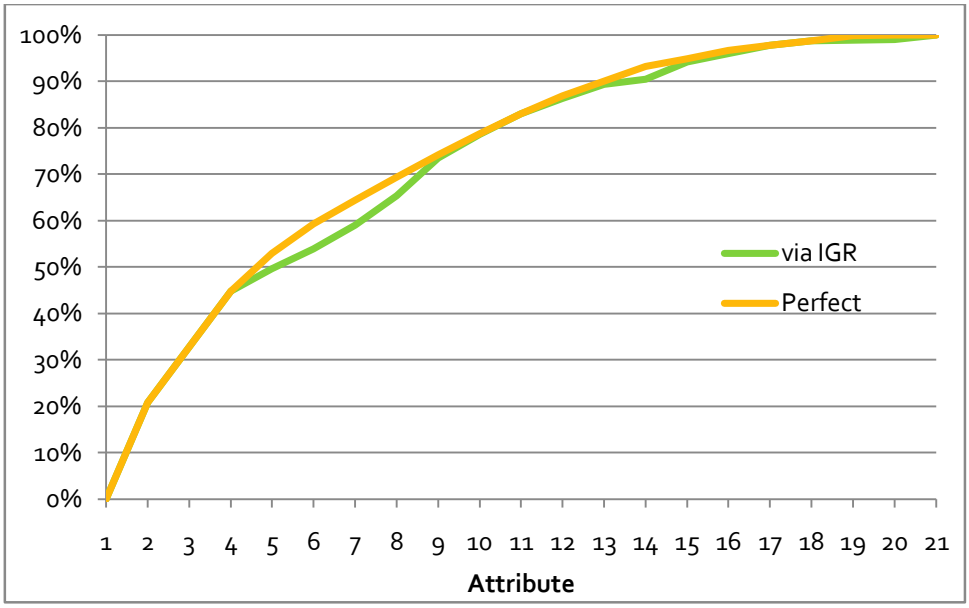

FIGURE 24 PERCENT CUMULATIVE ACTIONABILITY FOR GERMAN DATASET

For the GERMAN dataset, using IGR instead of α results in actionabiity capture that tracks closely with the best case. Here, the maximum gap is at the 7th-ranked attribute (5.4% points) and the average gap is 1.2% points.

Lastly, all 49 attributes from the three datasets are combined to show how IGR works as a proxy for α across a number of scenarios. This situation performs worse than when considered individually, with the biggest gap opening up at the 8th-ranked attribute (11% points) and an average loss of 4.9% points across all 49 attributes.

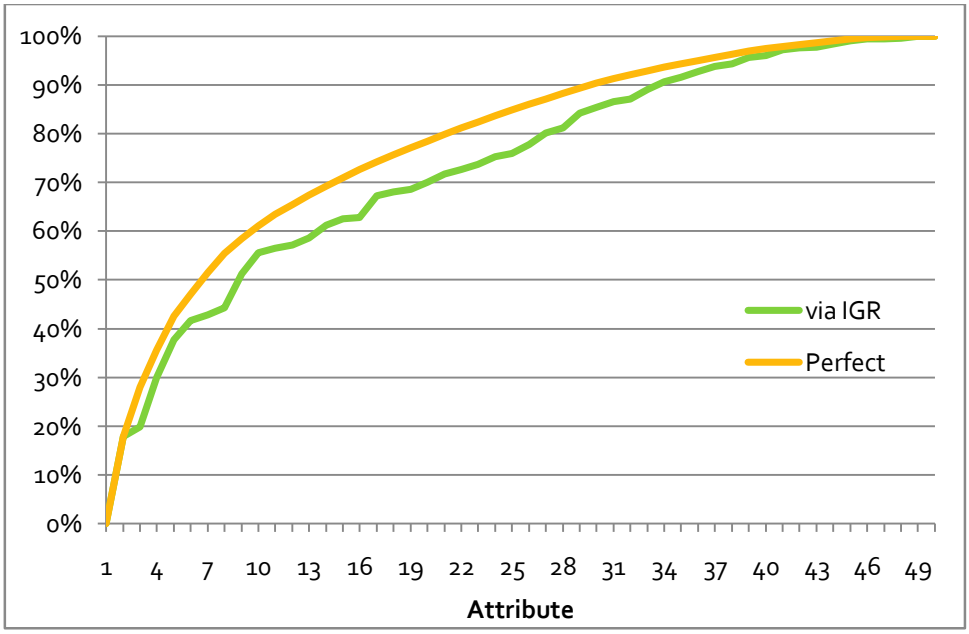

FIGURE 25 PERCENT CUMULATIVE ACTIONABILITY FOR ALL DATASETS

This examination of the performance of IGR scores as a substitute for α scores when selecting attributes indicates that information gain ratio is a robust predictor of actionability. IGR is useful for estimating which attributes are the most likely to have errors translate into mistakes. This, in turn, is useful for prioritising attributes for information quality interventions.

6.5.3 Effects on Interventions

As examined in Chapter 5, a very wide range of organisational and technological changes could be implemented to improve information quality. This spans re-training for data entry employees, re-configuration of the layout and functionality of packaged enterprise systems, enhanced quality assurance and checking for key business processes, use of alternate external information sources, re-negotiation of incentive structures for senior managers, reviews of source code, improvements to information and communication technology infrastructure, re-engineering of data models or the deployment of specialised matching and "data cleanup" tools.

To quantify the performance of a particular intervention, the framework outlined in Chapter 5 proposed a measure, τ, for traction. This was defined as:

$$\tau = \Pr\left(X_e \neq X'_e\right)$$

Here, X_e refers to the original value of the e^{th} attribute while X'_e is the value of the same attribute after the intervention. Mathematically, this takes the same form as the error rate, ε. For errors though, the original value (X_e) is compared with the true external world value (W_e).

In terms of the model, the primary effect of these disparate IQ interventions is to reduce the effective garbling rate, g_{eff} to a new, lower value. Fewer items being garbled result in fewer errors (as mediated by the garbling parameter, γ). Fewer errors result in fewer mistakes (as mediated by the attribute's actionability score, α). Fewer mistakes mean lower costs. The value of a proposed IQ intervention is the change in the costs to the process, minus the costs of the intervention itself. That is, the value of the intervention is the benefit minus the cost, where the benefit is the expected drop in the cost of mistakes.

To simplify discussion for the time being, suppose the expected per-customer cost of a mistake for a given process is given by M. I can appraise (or value) a particular proposed IQ intervention, on a per-customer basis[19], as benefits minus costs:

$$\begin{aligned} V_c &= B_c - C_c \\ &= M(\mu_1 - \mu_2) - C \end{aligned}$$

Where μ_1 is the original mistakes rate and μ_2 is the expected mistakes rate after the intervention. C is the cost of the intervention itself.

For a particular attribute, the rate of mistakes is the error rate multiplied by the actionability or α:

$$\mu = \varepsilon\alpha$$

By substituting in the formula for error rate, ε, in terms of garbling rate, g, and garbling parameter, γ, I obtain the following expression:

[19] *These results are aggregated into a total value (and total cost) basis in the subsequent section.*

$$\mu = [\gamma\, g\, +\, \gamma\,(1-g)(1-e^{-g})]\alpha$$
$$= [\,g\, +\, (1-g)(1-e^{-g})]\gamma\alpha$$

Since γ and α are constant for a particular attribute in a given process, regardless of garbling rate, the mistake rate is a function solely of g. Let g_1 be the initial garbling rate and g_2 be the post-intervention garbling rate. Substituting the mistake rates back into the benefit in the value equation yields:

$$B_c = \gamma\alpha M([g_1\, +\, (1-g_1)(1-e^{-g_1})] -\, [g_2\, +\, (1-g_2)(1-e^{-g_2})])$$

An alternative way to characterise interventions is to introduce a *correction factor*, χ. This takes a value from 0% to 100%, with 0% implying no change to the underlying garbling rate ($g_1 = g_2$) while 100% implies all garbling events are removed (so that g_2=0). In general, $g_2 = (1-\chi)g_1$. Under this modelling assumption, the traction, τ, is related to the correction factor, χ, by:

$$\tau = \Pr(X_e \neq X'_e)$$
$$= g_1\chi$$

This is because the IS state before and after the intervention only differ on those customer records where a garbling has been removed (corrected). Rather than using the traction directly, the explicit use of the garbling rate and correction factor will emphasise the underlying noise model used here.

So, when expressed in this way, the benefit of an intervention is given by:

$$B_c = \gamma\alpha M\left([g\, +\, (1-g)(1-e^{-g})] -\, \left[(1-\chi)g\, +\, \left(1-\left((1-\chi)g\right)\right)\left(1-e^{-(1-\chi)g}\right)\right]\right)$$
$$= \gamma\alpha M\left(g\chi\, +\, (1-g)(1-e^{-g}) -\, (1-g\, +g\chi\,)\left(1-e^{g(\chi-1)}\right)\right)$$
$$\approx 2\gamma\alpha M g\chi e^{-g}$$

For small values of g (<0.1), this can be further approximated as:

$$B_c \approx 2\gamma\alpha M g\chi(1-g)$$

Modelling the per-customer costs of proposed IQ interventions is heavily dependent on the specific conditions. For instance, a one-off re-design of a key enterprise system might entail a single fixed cost at the start. Other interventions (such as staff training) might be fixed but recurring. In general, there might be *fixed* costs (such as project overheads) and *variable costs* (that depend on the number of customer records). The variable costs might be a function of the number of customer records *tested* and the number of customer records that are updated (*edited*), whether or not the edit was correct. Assuming the intervention tests all the records and only edits the ones in error, the cost function is:

$$C_c = \kappa_F + \kappa_T\, +\, \varepsilon\kappa_E$$
$$\approx \kappa_F + \kappa_T\, +\, 2g\gamma\kappa_E$$

(Here, κ_F is the per-customer fixed-cost component, calculated by dividing the fixed costs by the number of customers.) A more sophisticated analysis is possible if the garbled records are identifiable. For example, suppose during a recent organisational take-over the target organisation's customer records were garbled during the database merge. In this case, only that proportion, g, needs to be tested:

$$C_c \approx \kappa_F + g\kappa_T + 2g\gamma\kappa_E$$

Substituting the simpler approximated cost and benefit equations into the per-customer value equation gives:

$$\begin{aligned} V_c &= B_c - C_c \\ &\approx 2\gamma\alpha M g\chi e^{-g} - \kappa_F - \kappa_T - 2g\gamma\kappa_E \end{aligned}$$

To put this into an investment perspective, the value (and hence benefits and costs) must be scaled up from a per-customer (and per-decision instance) basis to an aggregated form, across the customer and a period of time (investment window).

As outlined in Chapter 5, the Stake metric (expected cost of mistakes for a process) captures this. Here, the parameter M represents the expected cost of a single mistake. Not all of an organisation's customers are subject to all the processes; a bank, for instance, may expect just a few percent of its customers to apply for a mortgage in any year. Multiplying M by the number of customers undergoing a process expresses the value in absolute amounts. For comparison purposes, it is sufficient to express this as proportion of the entire customer base, β, rather than a raw count.

The third factor affecting a process's Stake is the expected number of times it is executed in the investment window. Mortgage applications might be quite rare, whereas some direct marketing operations may be conducted monthly. Most simply, this is the annual frequency, f, multiplied by the number of years, n. (To properly account for the time-value of money, a suitable discount rate must be used, in accordance with standard management accounting practice. This is addressed below.) The Stake for a given process (without discounting) is given by:

$$S = M\beta fn$$

Similarly, the costs (κ_F, κ_T and κ_E) can be scaled by βfn to reflect their recurrence. If the intervention attracts only a single cost in the first period (such as a one-off data cleansing or matching initiative) rather than ongoing costs, this scaling would not be required.

So the total value of an intervention on a particular attribute, a, for a particular process, p, is given by[20]:

$$V_{a,p} \approx \beta fn[2\gamma\alpha_p M_p g\chi e^{-g} - \kappa_F - \kappa_T - 2g\gamma\kappa_E]$$

Note that this is essentially a factor model: the benefit part of the equation is the product of eight factors. Should any of these factors become zero, the value is zero. What's more, the value varies linearly as any one factor changes (except for the garble rate, g, which is of a product-log form). This means that if you double, say, the correction factor χ while keeping everything else constant, the value will also double. Conversely, if any factor (such as γ) halves then the resulting value will halve.

To capture the total value of an intervention, the benefit arising from that intervention must be aggregated across all the processes, P_a, that use the attribute under question while the costs are only incurred once. This gives the total aggregated value of an intervention on attribute a:

[20] *The α and M parameters vary for each process, whereas all other parameters are a function only of the attribute.*

$$V_a \approx \beta f n[2\gamma g\chi e^{-g} \sum_{p \in P_a} \alpha_p M_p - \kappa_F - \kappa_T - 2g\gamma\kappa_E]$$

Where an annual discount rate of d needs to be applied, the discounted total aggregated value is:

$$V_r = \sum_{i=0}^{n-1} \frac{V_a}{n}(1-d)^i$$

When d=0 (ie the discount rate is zero), $V_r = V_a$.

This section reported and interpreted the results of the experimental process. This started by developing a generic garbling procedure to simulate the effect of noise on data values, accompanied by a statistical model of the resulting errors. The model was shown to fit closely with the pattern of events observed in the experiments. Drawing on the theoretical framework from Chapter 5, a proxy measure for predicting how errors translate into mistakes was obtained. This proxy – the *information gain ratio* – was shown to be useful for prioritising attributes by *actionability*. A measure for characterising proposed interventions were then developed, which in turn led to a benefit model and a cost model. This cost/benefit model was then expressed in discounted cash flow terms.

6.6 Application to Method

This section shows how the measures and formulae derived above can be employed by analysts designing and implementing IQ interventions. There are two broad uses for these constructs within an organisation. Firstly, they can focus analysts on the key processes, attributes and interventions that offer the greatest prospect for delivering improvements. Secondly, they can help appraise objectively competing proposals or initiatives.

When designing IQ interventions, is important to note the space of possible solutions is extremely large. There might be dozens or even scores of customer decision processes and scores – possibly hundreds – of customer attributes to consider, each subject to multiple sources of noise. Lastly, with different stakeholders and interests, there could be a plethora of competing and overlapping proposed interventions for rectifying information quality problems.

It is also important to understand that considerable costs are involved in undertaking the kind of quantitative analysis employed here. For example, to estimate properly the α measure would require the same approach as Section 6.5.2: repeatedly introducing errors into data, feeding it into the decision process, and checking for resulting mistakes. Undertaking such activity on a "live" production process would be costly, risky, time-consuming and prone to failure. Repeating this for the all the attributes as used in all the processes would be a formidable task. In contrast, the organisation's preferred discount rate for project investments, d, is likely to be mandated by a central finance function (or equivalent) and so is readily obtained.

Rather than estimating or computing all the measures for all possibilities, the ability to focus attention on likely candidates is potentially very valuable. The following table describes the measures, ranked from easiest to most difficult to obtain. The top half are external to any IQ initiative while the bottom half would be derived as part of the IQ initiative. Based on these assumptions, a cost-effective method for designing and appraising IQ interventions is developed.

Symbol	Name	Scope	Source	Definition
n	Period	Organisation	Project sponsor	The number of years for the investment.
d	Discount rate	Organisation	Finance	Set using financial practices taking into account project risks, the cost of capital etc.
f	Frequency	Process	Business owner	The expected number of times per year the process is executed.
β	Base	Process	Business owner	The proportion of the organisation's customer base that is subject to the process on each execution.
M	Mistake Instance Cost	Process	Business owner	The expected cost of making a mistake for one customer in one instance.
γ	Garble Parameter	Attribute	IQ project	The probability that a garbled data value will change.
IGR	Information Gain Ratio	Attribute	IQ project	A measure of an attribute's influence on a decision making process.
ε	Error Rate	Attribute	IQ project	The proportion of data values in error.
g	Garble Rate	Attribute	IQ project	A measure of the prevalence of garbling resulting from a noise process.
χ	Correction Factor	Intervention	IQ project	The net proportion of garbles corrected by an intervention.
κ	Cost Factors	Intervention	IQ project	The costs of an intervention: fixed (κ_F), testing (κ_T) and editing (κ_E).
α	Actionability	Process, Attribute	IQ project	The rate at which errors translate into mistakes.

TABLE 26 VALUE FACTORS FOR ANALYSIS OF IQ INTERVENTION

Below is the sequence of steps to investigate, design and appraise IQ interventions. The inputs are the organisation's set of customer decision processes that use a shared set of customer attributes. The output is an estimate of the Net Present Value (NPV) of candidate IQ interventions. The approach is to focus on the key processes, attributes and interventions that realise the largest economic returns, whilst minimising the amount of time and cost spent on obtaining the above measures.

1) Stake

Goal: *Select the most-valuable processes within scope*. Define values of d and n appropriate for the organisation. Identify the key customer decision processes. Use β and f to gauge "high traffic" processes. Use M to estimate high impact decisions. Multiplying these factors gives S, the stake. Use S to rank the processes and select a suitable number for further analysis.

2) Influence

Goal: *Select the most-important attributes*. For each of the top processes, use IGR (Influence) to select top attributes. This requires getting a sample of inputs and outputs for the decision functions and performing the entropy calculation. It does not require any manipulation of the systems themselves. For each attribute, sum the product of its Influence and Stake over each process to get an aggregated view of importance. Use this importance measure to select the top attributes.

3) Fidelity

Goal: *Select the most-improvable attributes*. For each of the top attributes, measure its γ value. This involves estimating the probability of each data value occurring and summing the squares. Observe the ∈ values (error rates between external-world and system representation) and, using the formula

in Section 6.5.1, estimate the garble rate, g. For the highly garbled attributes, estimate α values for the high-stake processes. This requires introducing noise to the attributes and seeing how it translates into mistakes. Use these to populate the Actionability Matrix (see Table 27 below).

4) Traction

Goal: *Select the most-effective interventions*. For each of the top attributes, estimate χ for various intervention proposals. To do this, analysts must either undertake the intervention on a sample and measure the drop in the effective garbling rate or draw on past experience in similar circumstances. In doing this, the cost factors (various κ) can be estimated. Use the values to estimate the costs and benefits of the candidate interventions with the NPV formula.

$$V_a \approx \beta f n[2\gamma g\chi e^{-g} \sum_{p \in P_a} \alpha_p M_p - \kappa_F - \kappa_T - 2g\gamma\kappa_E]$$

These models can be used in business cases to inform organisational decision-making about IQ investments.

Note that when the discount rate is applied, the NPV calculation requires using V_r:

$$V_r = \sum_{i=0}^{n-1} \frac{V_a}{n}(1-d)^i$$

Not all organisations use NPV directly. This expression can be re-cast as Internal Rate of Return (IRR) by setting V_r=0 and solving for d, or for payback period by solving for n. Alternatively, if Return on Investment (ROI) is required, then V_a is expressed as a ratio of benefit/cost instead of the difference (benefit-cost) before discounting.

While these steps are presented in a linear fashion, it is envisaged that an analyst would move up and down the steps as they search for high-value solutions, backtracking when coming to a "dead end". For example, an attribute selected at Step 2 (ie with high Influence) may have very low error rates (and hence low garble rates) and so afford little opportunity for improvement, regardless of how high the α and χ values may be. In this case, the analyst would go back to Step 2 and select the next-highest attribute. Similarly, a problematic and important attribute may simply not have any feasible interventions with a χ over 5%, in which case any further efforts will be fruitless and Step 3 is repeated.

To help the organisation keep track of the measures during the evaluation exercise, the following "Actionability Matrix" is proposed. This table of values is constantly updated throughout the project and it is important to note that it is not intended to be fully populated. In fact, determining all the values in the table indicates that the selection process has gone awry.

Suppose the organisation has a set of customer processes, P_1, P_2, ..., P_7 that use (some of) the customer attributes A_1, A_2, ..., A_{11}. Each cell, $\alpha_{a,p}$ records the actionability for the a^{th} attribute and p^{th} process. The last column records the error rate (∈) for the attribute, while the last row records the "annual stake[21]" ($\beta f M_p$) for the process. The rows and columns are arranged so that the attributes with the highest error rates are at the top and the processes with the highest stake are on the left.

[21] *That is, the stake as defined above, but divided by the number of years in the investment window, since this will be the same for all interventions.*

α	P_1	P_2	P_3	P_4	P_5	P_6	P_7	*Error Rate (∈)*
A_1	0.12	0.00	0.11		0.17			***55%***
A_2		0.05		0.15				***35%***
A_3	0.05	0.42	0.07	0.11				***15%***
A_4	0.22	0.11	0.61	0.03	0.07			***12%***
A_5	0.13			0.09				***10%***
A_6		0.07		0.55				***8%***
A_7								***8%***
A_8								***5%***
A_9								***3%***
A_{10}								***3%***
A_{11}								***0%***
Stake (S)	***$128***	***$114***	***$75***	***$43***	***$40***	***$24***	***$21***	

TABLE 27 ILLUSTRATION OF AN ACTIONABILITY MATRIX

Cells with high values of α are highlighted, drawing attention to the strong prospects for economic returns. To gauge the amount of cash lost each year on a certain process due to IQ deficiencies with a particular attribute, the cell value (α) is multiplied by the marginals (S and ∈). For example, the annual loss due to attribute A_6 on process $P_4$is 0.55*0.08*43 = $1.89. For A_3 in P_2 it is 0.42*0.15*114 = $7.18. The top "value leaker" is A_1 in P_1, with 0.12*0.55*128 = $8.48.

Note that the values of α are determined by how the decision function for a process uses the attributes. As such, they will not change unless the underlying decision function changes, meaning they will persist for some time. This means that as new processes are added to the organisation, the existing α values will not have to be updated, ensuring that organisational knowledge of how information is used can accumulate.

Throughout this method, it is assumed that the α values are obtainable (although expensive) and that IGR is used as a cheaper proxy to save unnecessarily measuring actionability for all attributes in all processes. However, for situations where a new customer decision process is being planned, α is just simply not available. If the organisation has not yet implemented the decision function then the probability of an error translating into a mistake cannot be measured experimentally.

In such circumstances, it is still possible to estimate IGR. The "true information gain ratio", Z, was shown in Table 11 to be highly correlated with the IGR for specific decision functions. Recall that this measures the influence of an attribute on the "correct decision" (as opposed to the decision made by a particular model). So IGR can be found as long as a sufficient sample of correct decisions (as used, for example, in training a decision function) is available. This IGR can then be used in lieu of α in the Actionability Matrix and NPV calculation to get an order-of-magnitude estimate of value.

This section has shown how the measures, formulae and assumptions can be used to guide the design and appraisal of IQ interventions, in a way that allows analysts to focus attention on high-value solutions while discarding low-value ones. A four-step iterative sequence is outlined along with a simple matrix for tracking key values. The resulting model of costs and benefits can be expressed as Net Present Value (or related measures) as needed by the organisation.

6.7 CONCLUSION

This concludes the specification and investigation of the framework. The chapter began with a high-level conceptual model of how IQ impacts on organisational processes and a theoretically-grounded set of candidate metrics for assisting analysts in prioritising IQ improvements.

The investigation proceeded by defining and creating an environment for inducing IQ deficiencies (noise) in realistic contexts, using realistic datasets, decision tasks and algorithms. The garbling noise process was selected for use here and its effects successfully modelled as a combination of inherent properties of the data (γ) and a controllable independent variable (g). The actionability (α), or effect of errors in giving rise to mistakes, was experimentally measured in these contexts. Far too expensive to obtain in all cases in practice, the theoretical measure of IGR (information gain ratio, or Influence in this framework) was tested and shown to be a very useful proxy. Finally, based on these findings, a financial model of the effect of removing IQ deficiencies was developed. A method was proposed for analysts to use the Actionability Matrix to apply these measures in an efficient iterative search for high-value IQ interventions.

Hence, this designed artefact meets the definition for a framework outlined in Chapter 5, Section 2. It comprises of a model (the Augmented Ontological Model of IQ), a set of measures (grounded in Information Theory) and a method (based around populating the Actionability Matrix). Importantly, this framework allows analysts to make recommendation on investments in IQ improvements using the quantitative language of cost/benefit analyses. This was a key requirement identified by the practitioner interview in Chapter 4.

Chapter 7

Research Evaluation

RESEARCH EVALUATION

7.1 SUMMARY

Design Science (or Design Research) has long been an important paradigm within Information Systems research. Its primary distinction from other approaches to research in the field is the pursuit of the goal of utility, as opposed to truth (Simon 1996). As outlined in Chapter 2, the framework for the valuation of customer information quality (IQ) falls squarely within the remit of DS. This chapter explains how both the research *process* (activities) and *product* (output) constitute Design Science and draws upon published guidelines to evaluate the research.

Specifically, following best practice guidelines for research (Hevner et al. 2004), the framework is presented as an *artefact*, in this case an abstract one, and is assessed against the seven guidelines laid out in their MISQ paper. The case is made that the framework satisfies the criteria and is both rigorous and relevant, with significance for practitioners and researchers.

7.2 EVALUATION IN DESIGN SCIENCE

When evaluating Design Science research, it is necessary to establish an appropriate set of definitions, guidelines or assessment criteria. Firstly, this is used to ensure that DS is the appropriate way to conceive of and evaluate the research effort. Secondly, this set forms the basis of the evaluation proper.

Note the distinction between evaluation of the DS research – the subject of this discussion – and the evaluation of the artefact itself. Chapter 6 describes the evaluation of the artefact *per se* (Section 4c below) whereas this chapter addresses the overall research, including the process, its likely impact and contribution.

The guidelines chosen for this evaluation are those published in MISQ (Hevner et al. 2004). Their paper, *Design Science in Information Systems Research*, has the goal of offering "clear guidelines for understanding, executing, and evaluating the research". It was selected for the following reasons:

- it specifically addresses DS in an Information Systems research context,
- MISQ is the leading journal in Information Systems and this paper is widely read and cited,
- the authors have experience in conducting DS research projects and prior publications on the topic,
- the paper is contemporary and reflects current thinking,
- it offers seven clear dimensions for evaluation, with a number of examples.

This is not to say that the paper represents an absolute consensus within the IS academic community about how to define and evaluate artefacts as part of research. However, it is a credible, familiar and useful basis for discussion.

The framework was developed in Chapter 5 (with a conceptual study) and tested and refined in Chapter 6 (with simulations) with the expressed intention of solving an organisational problem. Specifically, that problem is "How can organisations efficiently and objectively value the economic contribution of IQ interventions in customer processes?" This problem statement arose from a

qualitative analysis of context interviews with practitioners and senior managers (Chapter 4), who indicated that this is an *existing*, *important* and *persistent* problem. In conjunction with a review of academic literature (Chapter 3), this is identified as an *unsolved* problem. Furthermore, it is an *Information Systems* problem, as it relates to the planning and use of IS artefacts within organisations.

[Design Science] creates and evaluates IT artefacts intended to solve identified organizational problems. Such artefacts are represented in a structured form that may vary from software, formal logic, and rigorous mathematics to informal natural language descriptions. A mathematical basis for design allows many types of quantitative evaluations of an IT artefact, including optimization proofs, analytical simulation, and quantitative comparisons with alternative designs. (Hevner et al. 2004, p77)

The approach to solving the problem consisted of asking prospective users (in this case, managers and executives) about the form a solution to such a problem would take; investigating a wide range of "kernel theories" (or reference theories) and applying skill, knowledge and judgement in selecting and combining them; and undertaking a rigorous testing/refining process of the initial conceptualisation.

This research is prescriptive, rather than descriptive. The intent is to provide practitioners and researchers with a set of (intellectual) tools for analysing and intervening in existing (or proposed) information systems. So, importantly, the goal of the research project is to increase utility. In this context, that means the framework is likely be valuable for organisations because it allows for the objective valuation of (possible) IQ interventions to be undertaken in an efficient manner. A design for a an IQ valuation framework that requires infeasible pre-conditions (in terms of time, knowledge, staff or other resources) or produces opaque, dubious or implausible valuations will not have this utility.

This project is research, as opposed to a design project, because the resulting artefact – the framework – is sufficiently generic and abstract that it can be applied to a wide range of organisational settings and situations. It also has a degree of evaluative rigour and reflection that exceeds what is required for a one-off design effort.

The artefact is a framework comprising a number of elements, including constructs, models and a method, grounded in theory.

IT artefacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems). (Hevner et al. 2004, p336)

It is worth noting that other authors, such as Walls et al. (Walls et al. 1992) and Gregor and Jones (Gregor and Jones 2007), regard the abstract artefacts (constructs, models and methods) as a special kind of artefact, dubbed an Information System Design Theory (ISDT):

*The ISDT allows the prescription of guidelines for further artefacts of the same type. Design theories can be about artefacts that are either **products** (for example, a database) or **methods** (for example, a prototyping methodology or an IS management strategy). As the word "design" is both a noun and a verb, a theory can be about both the principles underlying the form of the design and also about the **act** of implementing the design in the real world (an intervention). (Gregor and Jones 2007, p322)*

However, in keeping with the prescriptions of Hevner et al. (Hevner et al. 2004), their broader sense of artefact, which encompasses "IS design theory", will be used here:

Purposeful artefacts are built to address heretofore unsolved problems. They are evaluated with respect to the utility provided in solving those problems. ***Constructs*** *provide the language in which problems and solutions are defined and communicated (Schön 1983).* ***Models*** *use constructs to represent a real world situation – the design problem and its solution space (Simon 1996). Models aid problem and solution understanding and frequently represent the connection between problem and solution components enabling exploration of the effects of design decisions and changes in the real world.* ***Methods*** *define processes. They provide guidance on how to solve problems, that is, how to search the solution space. These can range from formal, mathematical algorithms that explicitly define the search process to informal, textual descriptions of "best practice" approaches, or some combination.* ***Instantiations*** *show that constructs, models, or methods can be implemented in a working system. They demonstrate feasibility, enabling concrete assessment of an artefact's suitability to its intended purpose. They also enable researchers to learn about the real world, how the artefact affects it, and how users appropriate it. (Hevner et al. 2004, p. 341)*

When conducting DS research, it is not necessary to produce a working IT system, such as a software package or spreadsheet as proof of concept or even a complete instantiation:

[A]rtefacts constructed in design science research are rarely full-grown information systems that are used in practice. Instead, artefacts are innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, and use of information systems can be effectively and efficiently accomplished. (Hevner et al. 2004, p349)

The primary purpose of a proof of concept or artefact instantiation is to demonstrate the feasibility of the research process and the product (framework). In this case, the feasibility of the research process is argued for by the existence of the framework itself (ie the process did produce a product). Further, feasibility of the framework is demonstrated by noting that the input measurements are either common organisational parameters (eg the discounting rate) or have been derived from the real datasets sourced for the simulations (eg information gain ratio), while the model formulae are entirely amenable to computation. In this sense, appraisals for proposed interventions can always be produced ie the framework is feasible. Whether these are likely to be useful or not is discussed in Section 4.

This research project has all the elements required to constitute Design Science research: It identifies an existing, important, persistent, unsolved Information Systems problem. The proposed solution is a novel artefact informed by reference theories, intended to be used by practitioners in solving their problems. The steps of requirements-gathering, solution design and testing/refinement constitute the construction and evaluation phases identified in DS research. It is of sufficient abstraction and rigour that its product (the framework) can be applied to a wide range of organisational settings and situations.

7.3 Presentation of Framework as Artefact

The framework is conceptualised in Chapter 5, which involves elucidating and applying the relevant "kernel theories" to the broad organisational situation mapped out during the context interviews (Chapter 4). This results in the broad constructs, candidate measures and boundaries for the framework. In Chapter 6 (Simulations), the statistical and financial models are "fleshed out", new measures are derived, tested and refined, the sequence of steps clearly articulated and a simple "tool" (Actionability Matrix) is provided for analysts' use. The resulting framework is articulated below.

The framework takes an organisational-wide view of customers, systems and customer processes. It includes the creation of value over time when information about those customers is used in organisational decision-making within those processes:

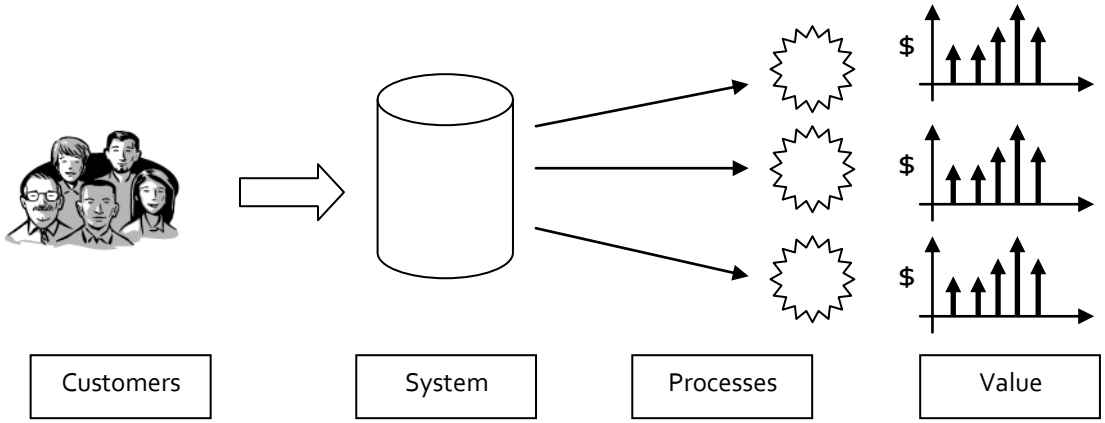

FIGURE 26 HIGH-LEVEL CONSTRUCTS IN THE FRAMEWORK

As shown, the framework assumes there is one system representing the customers (perhaps a data warehouse) shared by a number of customer processes. This "shared customer data" pattern fits many organisations.

The base conceptual model of how each process uses information is dubbed the Augmented Ontological Model, as it extends the Ontological Model of Wand and Wang (1996) to include decision-making:

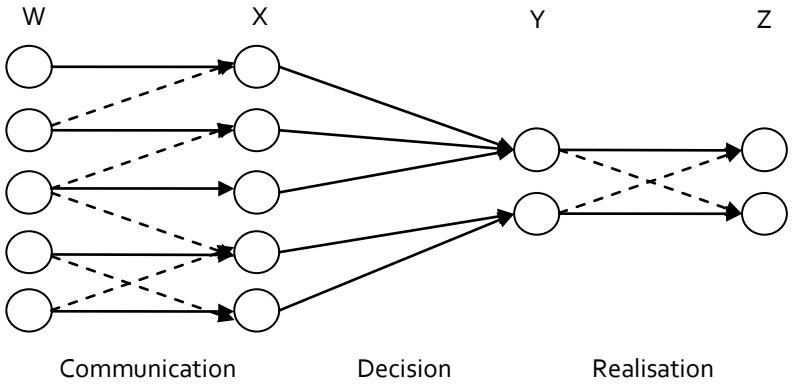

FIGURE 27 THE AUGMENTED ONTOLOGICAL MODEL

This diagram introduces a number of the key constructs used: a set of customers, each of whom exists in the external-world and exists in precisely one of a set of possible states, W. In keeping with the realist ontology throughout, these customers (and their state value) exist independently of any observation by the organisation. The organisational information system maps these customers (and their state values) onto a system representation drawn from the set of possible system states, X. In this way, each customer state is said to be *communicated* to the system.

For each customer undertaking the *process*, the system state is used by a *decision function* to select one action from a set of alternatives, Y. This action is *realised* against the optimal action, Z. ($z \in Z$ is not known at the time of the decision and $y \in Y$ is the system's best-guess.)

The impact of the realisation for each customer is expressed via a penalty matrix, Π, which describes the cost, $\pi_{y,z}$, associated with choosing action $y \in Y$ when $z \in Z$ is the optimal choice. When x=y the best choice is made so $\pi_{y,}$= 0. Each organisational process is run periodically on a portion of the customer base, generating future cash flows.

As a matter of practicality, both the customer statespace (W) and the system representation statespace (X) are decomposed into a set of a attributes A_1, A_2, …, A_a. Each attribute, A_i, has a number of possible attribute-values, so that $A_i \in \{a_1, a_2, \ldots, a_N\}$. The statespace W (and X) is the Cartesian product of these attribute sets, so that $W = A_1 \times A_2 \times \ldots \times A_a$. In practice, these attributes are generic properties of customers like gender, income bracket or post code, or organisational-specific identifiers like flags and group memberships.

The decision-function is conceived as any device, function or method for mapping a customer instance to a decision. The only formal requirement is that it is deterministic, so that the same decision is made each time identical input is presented. While this could be implemented by a person exercising no discretion, this research examines computer implementations. Examples include different kinds of decision trees, Bayesian networks and logistic model trees.

The next construct to consider is the *intervention* – either a one-off or an ongoing change to the way the external -world is *communicated* to the internal representation system. This may also involve changes to the representation system itself:

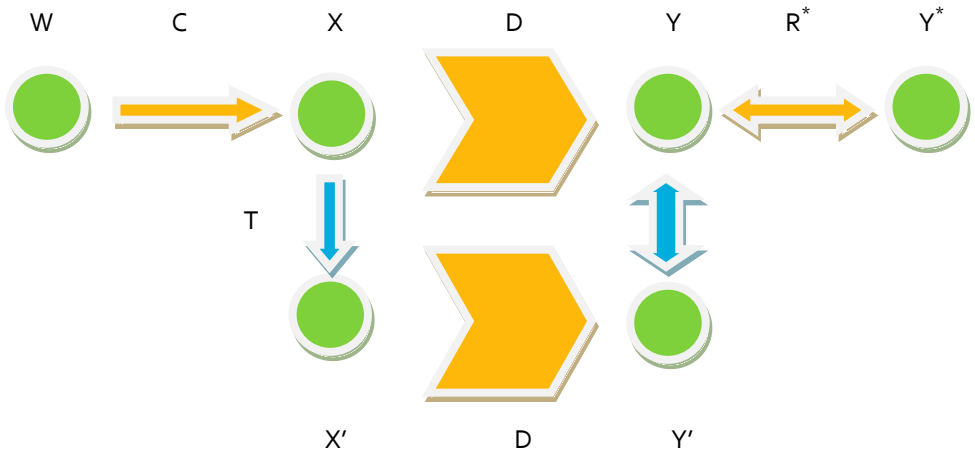

FIGURE 28 MODEL OF IQ INTERVENTIONS

In this model, the optimal decision $z \in Z$ has been replaced with $y^* \in Y^*$. This decision is not necessarily the "true optimal" decision z; it is the decision that the particular decision-function, D(•), would make if presented with perfect external-world state, $w \in W$, instead of the imperfect system state $x \in X$:

$$y = D(x)$$
$$y^* = D(w)$$

In using y^* instead of z, the effect of decision-making is "cancelled out", leaving the focus on Information Quality only, not algorithm performance.

The other extension is the intervention, T, which results in a revised system state, x'. This revised state is then presented to the same decision-function, D(•), to give a revised action, y':

$$y' = D(x')$$

The intervention is modelled as a change to how the external-world state, w, is communicated to the system, resulting in a revised system state, x' and hence a revised action, y'. This action is realised and compared with y^*. The difference between the cost of the prior decision and cost of revised decision is the benefit of the intervention.

The relations between these constructs, grounded in Information Theory and Utility Theory, was outlined in Chapter 5 (Conceptual Study):

Stake. The "value at risk" over time of the customer processes.

Influence. The degree to which an attribute "determines" a process outcome.

Fidelity. How well an external-world attribute corresponds with the system attribute.

Traction. The effectiveness of an intervention upon an attribute.

These broad measures were refined and examined in detail in Chapter 6 (Simulations). The starting point was modelling the communication between external-world and system by *noise.* An intervention can be modelled as the elimination of (a degree of) noise from this communication. This leads to the first statistical model, that of *garbling*.

A *garble* event is when a state-value is swapped with another drawn "at random". The garbling process has two parameters: γ (the garbling parameter) is an intrinsic measure of an attribute and *g* (the garbling rate) parameterises the degree of noise present. Together, these capture the notion of Fidelity by giving the probability of an *error* event, ε, a disagreement between the external-world and system state values:

$$\begin{aligned} \varepsilon &= P(x \neq w) \\ &= \gamma\, \mathrm{g} + \gamma\,(1-\mathrm{g})(1-e^{-g}) \end{aligned}$$

Some error events result in a *mistake* event (where the action does not agree with the "correct" one), with probability μ:

$$\begin{aligned} \mu &= P(y \neq y^*) \\ &= \alpha\varepsilon \\ &= \alpha\gamma[\,\mathrm{g} + \,(1-\mathrm{g})(1-e^{-g})] \end{aligned}$$

This proportion, α, of error events translating into mistake events characterises the Influence of the attribute within that process. Since α is difficult and costly to ascertain for all possibilities, the cheaper proxy measure *information gain ratio* (IGR) is used.

Mistake events are priced using the penalties in the penalty matrix, Π. The expected per-customer per-instance penalty is M. This, in turn, is converted into discounted cash flows using the parameters β (proportion of customer base going through the process), f (frequency of operation of process), n (number of years of investment) and d (appropriate discount rate). This addresses the Stake construct.

Lastly, Traction – the effectiveness of a candidate intervention in removing noise – is parameterised by χ, the net proportion of garble events removed. Combining all these elements with a cost model (κ_F for fixed costs, κ_T for testing and κ_E for editing) yields an intervention valuation formula:

$$V_a \approx \beta f n[2\gamma g\chi e^{-g} \sum_{p \in P_a} \alpha_p M_p - \kappa_F - \kappa_T - 2g\gamma\kappa_E]$$

Where an annual discount rate of d needs to be applied, the discounted total aggregated value is:

$$V_r = \sum_{i=0}^{n-1} \frac{V_a}{n}(1-d)^i$$

These parameters characterise the statistical and financial models which "flesh out" the broader conceptual constructs. The final component of the framework is a method for analysts to follow when performing the analysis in their organisation.

In Chapter 6, Section 6, the method is spelled out in some detail. The idea is to efficiently identify the key processes, attributes and interventions for analysis in order to avoid wasted analytical effort. The method comprises a sequence of iterative steps where successive candidate elements (processes, attributes and interventions) are selected based on the above parameters and their value assessed. The recommended order for analysis is *S*take, *I*nfluence, *F*idelity and *T*raction ("SIFT"), a handy mnemonic[22]. To help with this task, a tool (Actionability Matrix) is proposed and illustrated, which keeps of track of the model parameters as they are measured or derived and facilitates the selection of the next.

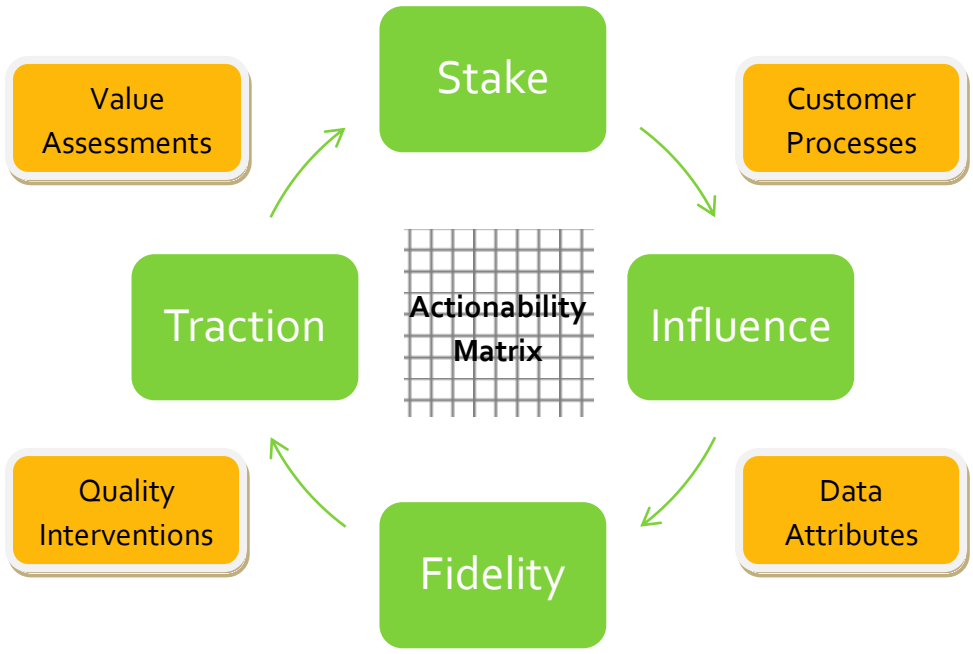

FIGURE 29 PROCESS OUTLINE FOR VALUE-BASED PRIORITISATION OF IQ INTERVENTIONS

A modified form of the "Framework for Comparing Methodologies" (Avison and Fitzgerald 2002) is used as an outline to summarise this framework:

[22] *This process of identifying and prioritising IQ interventions is akin to triage in a medical context. Interestingly, "triage" comes from the French verb "trier", which can be translated as "to sift".*

1. Philosophy
    a. Paradigm

    The underpinning philosophy in this research is Critical Realism. However, the particular ontological and epistemological stances taken for the purposes of building and evaluating this framework are not required by end-user analysts employing it in practice.

    For example, analysts may wish to adopt a scientific stance (an Realist ontology and a Positivist epistemology) for the purposes of comparing the system attribute values with the "external world" values.

    Where the framework does constrain analysts is the requirement that the situation being modelled can be decomposed into customers and processes, with well-defined states and decisions, respectively.

    b. Objectives

    The framework is not focused on developing a particular system, but improving the value to the organisation of its customer information, through its use in customer processes. This is measured through widely-used, investment-focused financial metrics.

    c. Domain

    Here, the domain addressed by the framework is at the level of organisational planning and resource allocation. It's not concerned with *how* one can (or should) improve customer IQ, but with capturing and articulating the expected benefits and costs of such initiatives.

    d. Target

    The kinds of organisations, systems and projects targeted involve large-scale, information-intensive ones with automated customer-level decision-making. Such environments would typically have call-centres, web sites, data warehouses supporting CRM activities.

2. Model

Figures 26-28 above express the model of the framework. This comprises a high-level view of the customer processes (Figure 26), the augmented Ontological Model of IQ (Figure 27) and a model of the impact of IQ interventions (Figure 28). These models are operationalised with related construct definitions and formulae.

3. Techniques and Tools

The principle techniques expressed here are the statistical sampling and measurement of the performance of the system in representing customer attributes s and the "downstream" impact on customer decision processes. The proposed metrics – Stake, Influence, Fidelity and Traction – are also tools to help the analyst assess and prioritise IQ interventions.

The Actionability Matrix tool makes the evaluation and recording of these metrics more systematic, improving the search process and assisting with re-use across time and on multiple evaluation projects.

4. Scope

The framework's method begins with a review of the organisation's set of customer decision processes and works through the "upstream" information resources and candidate interventions. It ends with a financial model of the costs and benefits of possible improvements to IQ.

This financial model is intended to be used in a business case to secure investment from the organisation's resource allocation process to implement the preferred intervention. It could also be used to track post-implementation improvements (budget vs actual) for governance purposes.

5. Outputs

The key outputs of the framework are 1) a "map" of high-value customer processes and their dependency on customer attributes, in the form of an up-to-date populated Actionability Matrix; and 2) a generic financial model of the expected costs and benefits associated with customer IQ interventions.

This section has outlined the conceptual *constructs*, statistical and financial *models* and sequence of *method* steps for the framework. The next section applies the chosen guidelines for evaluating the research.

7.4 Assessment Guidelines

This section applies the Design Science evaluation guidelines from Hevner et al.(Hevner et al. 2004) to the research project. The goal is to show how the research (including the process and the product) satisfies each of the criteria.

7.4.1 Design as an Artefact

Design-science research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation. (Hevner et al. 2004, p. 347)

As outlined in Section 3 above, the artefact is a *framework* comprising of a construct, a series of models (statistical and financial) and a method. The viability of the artefact is claimed in the fact that it can be expressed and applied to a problem domain. Further, its feasibility is argued from the computability of the models and the derivation of key metrics from real-world datasets.

7.4.2 Problem Relevance

The objective of design-science research is to develop technology-based solutions to important and relevant business problems. (Hevner et al. 2004, p. 347)

This question of problem relevance is discussed in Section 2. The extensive contextual interviews (Chapter 4) identify the inability to financially appraise customer IQ interventions as an "important and relevant business problem". Specifically, the problem is persistent, widespread – and largely unsolved. The inability to quantify the benefits, in particular, in business terms like Net Present Value (NPV) and Return on Investment (ROI) is especially problematic in situations where organisations take an investment view of such projects.

The framework does not construe a "technology-based solution" in itself. Rather, the objects of analysis (the host organisation's Information Systems and wider customer processes) are implemented using a range of technologies. The steps of analysis prescribed in the framework are made much easier using technology to populate the models.

7.4.3 DESIGN EVALUATION

The utility, quality, and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods. (Hevner et al. 2004, p. 347)

The specific evaluation of the key elements of the framework is the simulation study (Chapter 6). Hevner et al. define simulation as a type of experiment, where the researcher executes the artefact with "artificial data". Here, the artefact is executed with "synthetic data": real data with "artificial noise" added to it. In this way, the behaviour of the models (relationship between different parameters and measures) can be explored.

Chapter 2 (Research Method and Design) explains how this is the most suitable method for evaluating the artefact since access to a reference site to perform such a disruptive, sensitive and complicated analysis is not practicable. This constraint eliminates case and field studies, as well as white- and black-box testing. However, the goal of rigour (especially internal validity) requires that the underpinning mathematical and conceptual assumptions be tested. Purely analytical or descriptive approaches would not test these: the framework must be given a chance to fail. Reproducing "in the lab" the conditions found in the external-world is the best way to do this, providing sufficient care is taken to ensure that the conditions are sufficiently similar to invoke the intended generative mechanisms.

As is the norm with DS research, the evaluation and development cycles are carried on concurrently so that, for example, the "Fidelity" construct is re-cast to reflect the garbling process used to introduce controlled amounts of artificial noise. The garbling algorithm is developed to meet design criteria and is evaluated using a mathematical analysis (including the use of probability theory, combinatorics and calculus). This derivation is then checked against computer simulations with the "synthetic" data (ie real data with artificial noise). The correctness of the derivations and validity of the assumptions is proven using a range of "closeness" metrics: absolute differences, Root Mean Square Error (RMSE) and Pearson correlation.

In a similar vein, the *information gain ratio* (IGR) is used to measure the "Influence" construct instead of the *information gain* (IG), as initially proposed in Chapter 5 (Conceptual Study). This is because IGR performs better as a proxy for actionability, using a number of performance measures: Pearson correlation, Spearman (rank) correlation and the "percentage cumulative actionabilty capture" graphs.

Hevner et al. note that "[a] design artefact is complete and effective when it satisfies the requirements and constraints of the problem it was meant to solve" (Hevner et al. 2004, p352). In this context that means that a suitably-trained analyst can apply the framework to an IQ problem of interest and efficiently produce a valuation of possible interventions in a language readily understood by business: Net Present Value, Return on Investment or related investment-linked measures.

7.4.4 RESEARCH CONTRIBUTIONS

Effective design-science research must provide clear and verifiable contributions in the areas of the design artefact, design foundations, and/or design methodologies. (Hevner et al. 2004, p347)

The principal research contribution is the framework, encompassing the constructs, models and method. Hevner et al. specify that "[t]he artefact must enable the solution of heretofore unsolved problems. It may extend the knowledge base … or apply existing knowledge in new and innovative ways". They suggest looking for novelty, generality and significance across the artefact itself (research product), "foundations" (theory) and methodology (research process).

Firstly, this framework enables the valuation of IQ interventions – an important, widespread, persistent and unsolved problem in the business domain. While this has direct applicability to industry, it may also prove useful to research, in that it provides a generic, theoretically-sound basis for understanding and measuring the antecedents and consequents of IQ problems in organisations:

Finally, the creative development and use of evaluation methods (e.g, experimental, analytical, observational, testing, and descriptive) and new evaluation metrics provide design-science research contributions. Measures and evaluation metrics in particular are crucial components of design-science research. (Hevner et al. 2004, p. 347)

The formal entropy-based measurement of attribute influence (and actionability) within a decision process is of particular significance here. Further, it extends the knowledge base by introducing a practical use for Shannon's Information Theory (Shannon and Weaver 1949) into IS research and practice. Information Theory is not widely used as a "kernel theory" (or reference discipline) in Information Systems so this constitutes a novel adaptation of a large and well-developed body of knowledge to a class of IS problems.

Lastly, from a methodological perspective, the use of Critical Realism as a philosophy to underpin hybrid (qualitative and quantitative) research focused on designing an abstract framework is a contribution to the academic knowledge base. Using CR concepts like "generative mechanisms" to articulate and explain both the simulations and contextual interviews meant that a unified approach could be taken to analysing quantitative and qualitative data. A demonstration of how CR is not anathema to a carefully executed study incorporating detailed mathematical and statistical analyses is also a contribution to knowledge.

7.4.5 RESEARCH RIGOUR

Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artefact. (Hevner et al. 2004, p. 347)

The rigour of the research relates to the research process and how it impacts upon the resulting claims to knowledge. In the context of the empirical work, in the interviews (Chapter 4) the research project determined the nature of the problem, the existing "state of the art" and requirements and constraints for any proposed solution. This research follows prescribed qualitative sampling, data capture and analysis methods. Chapter 6 (Simulations) employed careful experimental procedures to reduce the possibility of unintended mechanisms interfering with the invocation of the generative mechanisms under study. It also uses rigorous mathematical models and tested assumptions and approximations, complying with the expected conventions in doing so. In both cases, sufficient detail is provided to allow subsequent researchers to reconstruct (and so verify through replication) the findings, or to make an assessment as to their scope, applicability or validity.

In the conceptual work, rigour is applied in the design of the overall study in the selection of Critical Realism as a unifying philosophical "lens" and the adoption of Design Science to describe and evaluate the study. The literature review (Chapter 2) and conceptual study (Chapter 5) draw upon existing knowledge bases and synthesise a framework from them. This chapter (Evaluation) involves reflecting on the research as a whole, how its constituent parts fit together and how it fits within a wider knowledge base.

Throughout the research project, the articulated ethical values were upheld, adding to the rigour of the process. These values include, for example, fair and honest dealings with research participants and stakeholders and academic integrity in acknowledging other people's contributions and reporting adverse findings.

7.4.6 Design as a Search Process

The search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment. (Hevner et al. 2004, p. 347)

The framework can be conceived as the result of an extensive search process. The conceptual study (Chapter 5) surveys a wide range of candidate concepts from engineering, economics and philosophy (introduced in the Literature Review, Chapter 3) and synthesises a suitable sub-set of them into a broad outline, the conceptual framework. This is further refined iteratively through the Simulations (Chapter 6) to produce the constructs, models and method that comprise the framework. Doing so requires the judicious selection, use and testing of mathematical assumptions, approximations, techniques and other formalisms focused on the "desired ends" (in this case, a transparent value model of IQ costs and benefits).

Throughout, the constraints of the problem domain (organisational context) are kept in mind. This includes access to commercial information, measurability of operational systems, mathematical understanding of stakeholders and the expectations and norms around organisational decision-making (ie business case formulation). For example, the contextual interviews (Chapter 2) show clearly that decision-makers in organisations expect to see value expressed as Net Present Value or Return on Investment (or related discounted cash flow models).

Seen in this light, the framework is a "satisficing" solution (Simon 1996) to the problem of valuing investments in customer IQ interventions. In particular, the use of the information gain ratio as a proxy for actionability to yield quicker, cheaper (though less accurate) value models constitutes a pragmatic approach to the problem.

7.4.7 Communication as Research

Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences. (Hevner et al. 2004, p. 347)

Chapter 2 (Research Method and Design) explains why, at this stage of the research, it is not feasible to present the framework to practitioners (of either orientation) to assess the effectiveness of the presentation, how readily it could be applied or inclination to deploy it. In short, the time available to present the complex material means rich, meaningful discussion is unlikely.

So, as a piece of scholarly research, the framework is presented to an academic audience for the purpose of adding to the knowledge base. (See "Research Contributions" above.) There is sufficient detail in the presentation of the constructs, models and method to allow researchers to reproduce the mathematical analyses and derivations and computer simulations, in order to verify the reported results and extend or refine further the artefact.

It is unlikely (and not the intent) that "technology-oriented" audiences would understand, evaluate or apply the framework as it is not directly concerned with databases, programming, networking or servers. Rather, the technical skills required (or "technologies", in a very broad sense) are in business analysis: modelling customers, processes and data, measuring performance, preparing cost/benefit analyses and working with allied professionals in marketing, operations and finance. The framework presented above is amenable to assessment for deployment within a specific organisation, providing the analyst has a strong mathematical background – for example, from operations research, statistics or data mining.

"Management-oriented" audiences are the intended beneficiaries of this framework. For them, the framework can be treated as a "black box": proposals, assumptions and organisational

measurements go in and investment metrics (NPV, ROI) come out. They may appreciate the constructs at a high-level, but the detail underpinning the statistical and financial models is not required to use the value models produced by the framework. These goals and constraints emerged during the contextual interviews (Chapter 4) with managers and executives, and so were guiding requirements for the development and refinement of the framework.

This section has assessed the research process and product against the seven DS guidelines advanced by Hevner et al. The research is found to meet the criteria for DS as it has the requisite features or elements:

- produced an artefact (the framework with its constructs, models and method),
- that tackles an important, widespread and persistent problem (IQ investments),
- through an iterative development/refinement cycle (conceptual study and simulations),
- with a rigorous evaluation of the artefact (mathematical analysis and simulations),
- which draws upon and adds to the knowledge base (Information Theory),
- resulting in a purposeful, innovative and generic solution to the problem at hand.

Chapter 8

Conclusion

CONCLUSION

8.1 SUMMARY

The quality of information is a key challenge in designing, developing and maintaining any large-scale information system. In particular, customer information for use in enterprise systems supporting functions like Customer Relationship Management (CRM) systems requires significant organisational resources. Compared with other information systems projects, Information Quality (IQ) initiatives struggle to compete for these scarce organisational resources. This is due in part to the difficulty of articulating a quantitative cost/benefit analysis in financial terms.

The SIFT framework developed and evaluated in the preceding chapters offers analysts a set of *constructs*, *measures* and a *method* to help them assess, prioritise and present disparate IQ improvements as sound organisational investments. Further, this framework allows them to do so in an efficient manner with minimal disruption of operational systems.

The research process followed a Design Science approach, underpinned by a Critical Realist philosophy. Here, the framework was cast as an abstract *artefact* with utility as the goal. Motivated by a series of practitioner interviews, this artefact was developed through a conceptual study. This was followed by a quantitative investigation using computer simulations and mathematical derivations which provided further detail and empirical support. Finally, the framework was successfully evaluated against a leading set of criteria from the field of Design Science.

8.2 RESEARCH FINDINGS

Within Information Systems research, the sub-field of Information Quality has seen a large number of conceptual and practitioner-oriented approaches. However, few efforts are directed at measuring the value created for organisations by investing in customer information quality improvements. Many of the frameworks are burdened by a lack of theoretical rigour, giving rise to unclear definitions, misleading measures and an inability to operationalise constructs that renders them impractical.

The large body of knowledge residing in Information Theory and Information Economics has had little bearing on Information Quality research. The Semiotic IQ Framework (Price and Shanks 2005a), organises a hierarchy of IQ concepts including the Ontological Model for IQ (Wand and Wang 1996) at the semantic level, which is isomorphic to Shannon and Weaver's model of information (Shannon and Weaver 1949). This presented an opportunity to bridge knowledge from these different domains in order to tackle the difficult IQ investment problem.

A series of semi-structured interviews with practitioners was required to assess the current state of the art within industry and to capture the intended requirements of the framework. This involved interviewing 15 practitioners (analysts, managers and executives) with a range of backgrounds and experiences, for one to two hours about their experiences with IQ investment, measurement and business cases.

The key finding was that while IQ is recognised as an important enabler of value creation within organisations, there is a widespread inability to employ accepted value measurements in accordance

with standard organisational practices. Many subjects argued that while they believe they know what is required to remedy specific IQ deficiencies and that doing so would be in their organisation's interests, the lack of value measures means IQ initiatives cannot compete for and win funding against more traditional IS projects. As a result, this frustrated articulation of the benefits gives rise to widespread under-investment in IQ. The way many IQ initiatives secure funding and advance is from a mandate from a sufficiently senior sponsor. This can also lead to a perceived misallocation of the organisation's resources.

This need for practitioners to support their business cases with financial arguments provided the central design goal of the framework. Drawing on the quantitative approaches from the earlier Literature Review (Chapter 3), the framework was conceived with customer-level decision-making at its core. This facilitated the use of a variety measurement approaches from the related fields of Decision Theory, Machine Learning and Information Theory. In this framework, CRM processes are re-cast as customer classifiers and their performance can be assessed similarly. More importantly, the effect of IQ deficiencies on classifier performance can be described quantitatively too.

To take the framework to a richer level and allow empirical validation, some real-world scenarios were found in the Data Mining literature. By recreating the target contexts (large-scale automated customer-level decision-making) and inducing the IQ deficiencies with a "garbling" noise process, the consequences of the deficiencies could be explored. In this way, the simulations allowed empirical verification of mathematical modelling of the noise process and its effect on observed errors.

Further, the simulations demonstrated that the proposed Influence metric (Information Gain Ratio, or IGR) is a very good proxy for *actionability* (the propensity of a representational error to translate into a decision mistake). This is important because IGR is a property of the decision-making function and can be assessed much more quickly and cheaply – and with minimal disruption – compared with the experimental requirements of directly measuring actionability.

Along with Stake, Fidelity and Traction, the Influence metric was embedded in a method for financial modelling of cash flows arising from the costs and benefits and candidate IQ interventions. The method employed an Actionability Matrix to guide the search of processes and attributes to determine efficiently the most valuable opportunities for improvement.

Finally, the framework – comprising the model, measures and method – was evaluated against the set of seven criteria proposed in MISQ (Hevner et al. 2004) by several leading Design Science scholars. The framework (as an abstract artefact) was found to be a purposeful, innovative and generic solution to an important, widespread and persistent problem.

8.3 Limitations and Further Research

The primary limitation of this research is the lack of empirical testing of the framework in its entirety in a realistic situation. Ideally, this would involve the application of the framework by the target users (business analysts within an organisation) to build a business case for customer IQ improvement across a range of customer processes. As discussed in Section 2.6.3, access to an organisation willing to support research like this would likely be difficult, owing to the large commitment in time and resources.

That said, the research evaluation undertaken as part of this project lowers the risk for prospective industry partners, making future collaboration more likely. Specifically, the development and testing of the use of IGR to speed up the search makes further field testing a more attractive proposition.

Undertaking further field trials would allow researchers to assess and understand:

- **The output**. How acceptable are the kinds of financial models, valuations and recommendations produced by the framework? To what extent are decision-makers likely to give credence to these outputs, given the complexity and opacity of the underlying model? Does this change with familiarity? What are the determinants of acceptability for these decision-makers?

- **The process**. What is the effectiveness and efficiency of actually using the framework? What are the determinants of successful use of the framework? What kinds of improvements could be made? Could it be made more generic or does it need to be more focused?

- **The context.** What motivates organisations to adopt this framework? What background or preparation does an analyst need to use it successfully? How can the outputs be used by the organisation to manage IQ over time or across organisational boundaries?

All of these questions can only be tackled by deploying the framework in a real-world situation, perhaps as a field experiment (Neuman 2000). Practically, to do this kind of research would require an uncommon level of access with an industry partner. Given the immaturity of the framework plus the confidentiality and commercial constraints, an action research approach (eg. Burstein and Gregor 1999) to evaluating the artefact may be a better fit.

The second limitation (and hence opportunity for further research) is the noise process used in the simulations. To recap, a "garbling" noise process was used, which involves iterating through a set of records and swapping a particular field's values with the value from another record selected at random. The garbling rate, g, was used to control the amount of noise introduced.

The effect of this noise process was modelled to a high level of precision by its mathematical derivation. However, in a practical sense, it represents a particularly "disastrous" noise event, where all traces of the correct value are lost. Such a garbling event might arise from a field value being erased or a database index getting "jumbled" or a data entry error by a clerk.

While this kind of event happens in practice and warrants study in its own right, other kinds of noise will have different properties. Specifically, some will retain some information about the original value. For instance, miscoding a telephone number by one digit is likely to result in a telephone number in the same geographical region. For IQ deficiencies arising from currency issues, the "stale" value is not entirely unrelated to the correct value. For example, with customer residential addresses, customers on the whole are likely to move to postcodes or regions with similar socio-economic conditions. Decisions made on the basis of such stale information may frequently be the same as with correct information. So the errors introduced by the stale address are not as complete or "disastrous" as the ones modelled by the garbling process, which represents the worst-case.

It may be that further research could yield an alternative noise process that captures some of the information-retaining characteristics of more realistic noise. Ideally, any new noise processes would still be amenable to the kind of empirical analysis undertaken here. If not, new methods could be developed for investigation, perhaps based on other areas of applied mathematics.

References

REFERENCES

Alter, S. "A Work System View of DSS in its Fourth Decade," *Decision Support Systems* (38:3) 2004, pp 319-327.

Alter, S., and Browne, G. "A Broad View of Systems Analysis and Design," *Communications of the Association for Information Systems* (16:50) 2005, pp 981-999.

Applegate, L.M. "Rigor and Relevance in MIS research - Introduction," *MIS Quarterly* (23:1) 1999, pp 1-2.

Arnott, D. "Cognitive Biases and Decision Support Systems Development: A Design Science Approach," *Information Systems Journal* (CCCB:1) 2006, pp 55-78.

Arrow, K.J. *The Economics of Information* Blackwell, Oxford [Oxfordshire], 1984.

Asuncion, A., and Newman, D.J. "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2007.

Avison, D.E., and Fitzgerald, G. *Information Systems Development: Methodologies, Techniques and Tools* McGraw-Hill, London, 2002.

Ballou, D., Wang, R., Pazer, H., and Tayi, G.K. "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science* (44:4) 1998, pp 462-484.

Ballou, D.P., Chengalur-Smith, I.N., and Wang, R.Y. "Sample-Based Quality Estimation of Query Results in Relational Database Environments," *IEEE Transactions on Knowledge & Data Engineering* (18:5) 2006, pp 639-650.

Ballou, D.P., and Pazer, H.L. "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science* (31:2) 1985, pp 150-162.

Ballou, D.P., and Pazer, H.L. "Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff," *Information Systems Research* (6:1) 1995, p 51.

Ballou, D.P., and Pazer, H.L. "Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts," *IEEE Transactions on Knowledge & Data Engineering* (15:1) 2003, p 240.

Ballou, D.P., and Tayi, G.K. "Reconciliation Process for Data Management in Distributed Environments," *MIS Quarterly* (9:2) 1985, pp 97-108.

Ballou, D.P., and Tayi, G.K. "Methodology for Allocating Resources for Data Quality Enhancement," *Communications of the ACM* (32:3) 1989, pp 320-329.

Barone, D., Cabitza, F., and Grega, S. "HDQ: A Meta-Model for the Quality Improvement of Heterogeneous Data," in: *Second International Conference on Digital Information Management*, IEEE, Lyon, France, 2007, pp. 418-423.

Batini, C., and Scannapieco, M. *Data Quality: Concepts, Methodologies And Techniques* Springer, Berlin, 2006.

Bayón, T., Gutsche, J., and Bauer, H. "Customer Equity Marketing: Touching the Intangible," *European Management Journal* (20:3) 2002, pp 213-222.

Becker, S. "A Practical Perspective on Data Quality Issue," *Journal of Database Management* (9:1), Winter98 1998, p 35.

Berger, P., and Nasr, N.I. "Customer Lifetime Value: Marketing Models and Applications," *Journal of Interactive Marketing* (12:1) 1998, pp 17-30.

Bhaskar, R. *A Realist Theory of Science* Books, [York], 1975.

Bhaskar, R. *The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences* Harvester Press, Brighton, Sussex, 1979.

Bhaskar, R. *Reclaiming Reality: A Critical Introduction to Contemporary Philosophy* Verso, London, 1989.

Brobrowski, M., and Soler, S.V. "DQ Options: Evaluating Data Quality Projects Using Real Options," in: *Ninth International Conference on Information Quality (IQ 2004)*, MIT, 2004, pp. 297-310.

Burstein, F., and Gregor, S. "The Systems Development or Engineering Approach to Research in Information Systems: An Action Research Perspective," *Proceedings of the10th Australasian Conference on Information Systems*) 1999.

Cai, Y., and Shankaranarayanan, G. "Managing Data Quality in Inter-Organisational Data Networks," *International Journal of Information Quality* (1:3) 2007, pp 254-271.

Cappiello, C., Ficiaro, P., and Pernici, B. "HIQM: A Methodology for Information Quality Monitoring, Measurement, and Improvement," in: *ER (Workshops)*, Springer, Tucson, USA, 2006, pp. 339-351.

Cappiello, C., Francalanci, C., and Pernici, B. "Time-Related Factors of Data Quality in Multichannel Information Systems," *Journal of Management Information Systems* (20:3), Winter2003 2003, pp 71-91.

Carlsson, S.A. "Advancing Information Systems Evaluation (Research): A Critical Realist Approach," *Electronic Journal of Information Systems Evaluation* (6:2) 2003a, pp 11-20.

Carlsson, S.A. "Critical Realism: A Way Forward in IS Research," in: *European Conference on Information Systems*, Naples, Italy, 2003b.

Carlsson, S.A. "Critical Realism in IS Research," in: *Encyclopedia of Information Science and Technology (I)*, Idea Group, 2005a, pp. 611-616.

Carlsson, S.A. "A Critical Realist Perspective on IS Evaluation Research," European Conference on Information Systems, Regensburg, Germany, 2005b.

Chandler, D. *Semiotics: The Basics* Routledge, London, 2007.

Chauchat, J.-H., Rakotomalala, R., Carloz, M., and Pelletier, C. "Targeting Customer Groups using Gain and Cost Matrix : A Marketing Application," in: *Data Mining for Marketing Applications, PKDD'2001*, Freiburg, Germany, 2001, pp. 1-13.

Chengalur-Smith, I.N., and Ballou, D.P. "The Impact of Data Quality Information on Decision Making: An Exploratory Analysis," in: *IEEE Transactions on Knowledge & Data Engineering*, IEEE, 1999, p. 853.

Courtheoux, R.J. "Marketing Data Analysis and Data Quality Management," *Journal of Targeting, Measurement & Analysis for Marketing* (11:4) 2003, p 299.

Cover, T.M., and Thomas, J.A. *Elements of Information Theory* J. Wiley, Hoboken, N.J., 2005.

Dariusz, K., Tadeusz, L., Maciej, S., and Bogdan, T. "Investigation of Application Specific Metrics to Data Quality Assessment," in: *ER (Workshops)*, Springer Berlin / Heidelberg, Tuscon, USA, 2007.

Davenport, T.H., and Markus, M.L. "Rigor vs. Relevance Revisited: Response to Benbasat and Zmud," *MIS Quarterly* (23:1) 1999, pp 19-23.

Davies, M.A.P. "Perceived Information Quality: An Information Processing Perspective," *Journal of International Consumer Marketing* (13:4) 2001, p 29.

DeLone, W.H., and McLean, E.R. "Information Systems Success: The Quest for the Dependent Variable," *Information Systems Research* (3:1) 1992, pp 60-95.

DeLone, W.H., and McLean, E.R. "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update," *Journal of Management Information Systems* (19:4) 2003, pp 9-30.

Dobson, P.J. "The Philosophy of Critical Realism - An Opportunity for Information Systems Research," *Information Systems Frontiers* (3:2) 2001, pp 199-210.

Drummond, C., and Holte, R.C. "Cost curves: An Improved Method for Visualizing Classifier Performance," *Machine Learning* (65:1) 2006.

Dvir, R., and Evans, S. "A TQM Approach to the Improvement of Information Quality," in: *Conference on Information Quality (IQ 1996),* R.Y. Wang (ed.), MIT, 1996, pp. 207-220.

Eckerson, W. "Excerpt from TDWI's Research Report - Data Quality and the Bottom Line," *Business Intelligence Journal* (8:1) 2001.

English, L.P. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits* Wiley, New York, 1999.

English, L.P. "Information Quality: Critical Ingredient for National Security," in: *Journal of Database Management*, 2005, pp. 18-32.

Eppler, M., and Helfert, M. "A Framework For The Classification Of Data Quality Costs And An Analysis Of Their Progression," in: *Ninth International Conference on Information Quality (IQ 2004)*, MIT, 2004, pp. 311-325.

Eppler, M., and Mengis, J. "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines," *Information Society* (20:5) 2004, pp 325-344.

Evans, P. "Scaling and Assessment of Data Quality," in: *Acta Crystallographica: Section D*, Blackwell Publishing Limited, 2006, pp. 72-82.

Even, A., and Shankaranarayanan, G. "Utility-Driven Assessment of Data Quality," *DATA BASE* (38:2) 2007a, pp 75-93.

Even, A., and Shankaranarayanan, G. "Utility-Driven Configuration of Data Quality in Data Repositories," *IJIQ* (1:1) 2007b, pp 22-40.

Even, A., Shankaranarayanan, G., and Berger, P.D. "Economics-Driven Data Management: An Application to the Design of Tabular Data Sets," *IEEE Transactions on Knowledge & Data Engineering* (19:6) 2007, pp 818-831.

Fawcett, T. "An Introduction to ROC Analysis," *Pattern Recognition Letters* (27:8) 2006, p 861.

Fellegi, I.P., and Sunter, A.B. "A Theory for Record Linkage," *Journal of the American Statistical Association* (64:328) 1969, pp 1183-1210.

Fischer, P., Schulz-Hardt, S., and Frey, D. "Selective Exposure and Information Quantity: How Different Information Quantities Moderate Decision Makers' Preference for Consistent and Inconsistent Information," *Journal of Personality & Social Psychology* (94:2) 2008, pp 231-244.

Fisher, C.W., Chengalur-Smith, I., and Ballou, D.P. "The Impact of Experience and Time on the Use of Data Quality Information in Decision Making," *Information Systems Research* (14:2) 2003, pp 170-188.

Fjermestad, J., and Romano, N.C. "Introduction to the Special Section: Advances in Electronic Commerce Customer Relationship Management," *International Journal of Electronic Commerce* (7:2) 2002, pp 7-8.

Fox, C., and Redman, T. "The Notion of Data and its Quality Dimensions," in: *Information Processing & Management*, 1994, pp. 9-20.

Frank, A. "Analysis of Dependence of Decision Quality on Data Quality," *Journal of Geographical Systems* (10:1) 2008, pp 71-88.

Freeman, P., and Seddon, P.B. "Benefits from CRM-Based Work Systems," in: *Proceedings of European Conference on Information Systems (ECIS)*, Regensburg, 2005.

Freeman, P., Seddon, P.B., and Scheepers, R. "Explaining Organisational Investments in CRM Point Solutions.," in: *Proceedings of the 15th European Conference on Information Systems*, St Gallen, Switzerland, 2007.

Gartner "CRM Data Strategies: The Critical Role of Quality Customer Information," Gartner Inc.

Gartner "Using Business Intelligence to Gain a Competitive Edge: Unleashing the Power of Data Analysis to Boost Corporate Performance," Gartner Inc., 2004.

Ge, M., and Helfert, M. "A Theoretical Model to Explain Effects of Information Quality Awareness on Decision Making," in: *International Conference on Enterprise Information Systems*, Madeira, Portugal, 2007, pp. 164-169.

Ge, M., and Helfert, M. "Data and Information Quality Assessment in Information Manufacturing Systems," in: *International Conference on Business Information Systems*, Springer, Innsbruck, Austria, 2008, pp. 380-389.

Glass, R.L. "Rigor vs. Relevance: A Practitioner's Eye View of an Explosion of IS Opinions," *The Communications of the Association for Information Systems* (6) 2001.

Gregor, S. "The Nature of Theory in Information Systems," *MIS Quarterly* (CCCB CCCBC:3) 2006, pp 611-642.

Gregor, S., and Jones, D. "The Anatomy of a Design Theory," *Journal of the Association for Information Systems* (8:5) 2007, p 335.

Gustafsson, P., Lindström, A., Jägerlind, C., and Tsoi, J. "A Framework for Assessing Data Quality - From a Business Perspective," *Software Engineering Research and Practice*) 2006, pp 1009-1015.

Han, J., and Kamber, M. *Data Mining : Concepts and Techniques*, (2nd ed.) Morgan Kaufmann, San Francisco, CA, 2006.

Hand, D.J. *Construction and Assessment of Classification Rules* Wiley, Chichester ; New York, 1997, p. 214.

Heinrich, B., Kaiser, M., and Klier, M. "Metrics for Measuring Data Quality - Foundations for an Economic Data Quality Management," *ICSOFT (ISDM/EHST/DC)*) 2007, pp 87-94.

Heller, W.P., Starr, R.M., Starrett, D.A., and Arrow, K.J. *Uncertainty, Information, and Communication* Cambridge University Press, Cambridge [Cambridgeshire], 1986.

Henderson, I., and Murray, D. "Prioritising and Deploying Data Quality Improvement Activity," *Journal of Database Marketing & Customer Strategy Management* (12:2) 2005, pp 113-119.

Hevner, A.R., March, S.T., Park, J., and Ram, S. "Design Science in Information Systems Research," *MIS Quarterly* (28:1) 2004, pp 75-105.

Hill, G. "An Information-Theoretic Model of Customer Information Quality," in: *IFIP WG8.3 International Conference on Decision Support Systems* R. Meredith (ed.), Monash University, Prato, Italy, 2004.

Huang, K.-T., Lee, Y.W., and Wang, R.Y. *Quality Information and Knowledge* Prentice Hall PTR, Upper Saddle River, N.J., 1999.

Hughes, A.M. *Strategic Database Marketing* McGraw-Hill, New York, 2006.

Ishaya, T., and Raigneau, J. "Data Quality for Effective E-Commerce Customer Relationship Management," in: *International Conference on Enterprise Information Systems*, Madeira, Portugal, 2007, pp. 92-100.

Jacaruso, B. "Improving Data Quality," *Library & Information Update* (5:11) 2006, pp 34-35.

James, W. *Pragmatism: A New Name for Some Old Ways of Thinking : Popular Lectures on Philosophy* Longmans, Green, London, 1907.

Jayaganesh, M., Shanks, G., and Jagielska, I. "CRM System Benefits: Encouraging Innovative Use of CRM Systems," in: *IFIP International Conference on Creativity and Innovation in Decision Making and Decision Support (CIDMDS 2006)*, London, 2006.

Johnson, J.R., Leitch, R.A., and Neter, J. "Characteristics of Errors in Accounts Receivable and Inventory Audits," *Accounting Review* (56:2) 1981, p r81.

Jörg, B., Björn, N., and Christian, J. "Socio-Technical Perspectives on Design Science in IS Research," *Advances in Information Systems Development*) 2007.

Joseph, B., Robert, L.G., Eric, S., and Zhao, T. "An Event Based Framework for Improving Information Quality that Integrates Baseline Models, Causal Models and Formal Reference Models," in: *Proceedings of the 2nd International Workshop on Information Quality in Information Systems*, ACM, Baltimore, Maryland, 2005.

Kaboub, F. "Roy Bhaskar's Critical Realism: A Brief Overview and a Critical Evaluation."

Kahn, B.K., Strong, D.M., and Wang, R.Y. "Information Quality Benchmarks: Product and Service Performance," *Communications of the ACM* (45:4) 2002, pp 184-192.

Kahneman, D., and Tversky, A. "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* (47:2) 1979, pp 263-291

Kaomea, P. "Valuation of Data Quality: A Decision Analysis Approach ", MIT.

Kaplan, D., Krishnan, R., Padman, R., and Peters, J. "Assessing Data Quality in Accounting Information Systems," *Communications of the ACM* (41:2) 1998, pp 72-78.

Karr, A.F., Sanil, A.P., Sacks, J., and Elmagarmid, A. "Workshop Report: Affiliates Workshop on Data Quality," 177, National Institute of Statistical Sciences.

Keynes, J.M. *A Treatise on Probability* Rough Draft Printing, 1923.

Khalil, O.E.M., Strong, D.M., Kahn, B.K., and Pipino, L.L. "Teaching Information Quality in Information Systems Undergraduate Education," in: *Informing Science*, 1999, pp. 53-59.

Klein, B.D., and Callahan, T.J. "A Comparison of Information Technology Professionals' and Data Consumers' Perceptions of the Importance of the Dimensions of Information Quality," *International Journal of Information Quality* (1:4) 2007, pp 392-411.

Kock, N.F., McQueen, R.J., and Scott, J.L. "Can Action Research Be Made More Rigorous in a Positivist Sense? The Contribution of an Iterative Approach.," *Journal of Systems and Information Technology* (1:1) 1997, pp 1-24.

Kohavi, R. "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," in: *KDD96*, 1996, pp. 202-207.

Kononenko, I., and Bratko, I. "Information-Based Evaluation Criterion for Classifier's Performance," *Machine Learning* (6:1) 1991, pp 67-80.

Krogstie, J., Sindre, G., and Jorgensen, H. "Process Models Representing Knowledge for Action: A Revised Quality Framework," *European Journal of Information Systems* (15:1) 2006, pp 91-102.

Kuechler, B., and Vaishnavi, V. "On Theory Development in Design Science Research: Anatomy of a Research Project," *European Journal of Information Systems* (17:5) 2008, pp 489-504.

Lawrence, D.B. *The Economic Value of Information* Springer, New York, 1999.

Layder, D. *New Strategies in Social Research: An Introduction and Guide* Polity Press, Cambridge, UK, 1993.

Lee, Y.W., Pipino, L., Strong, D.M., and Wang, R.Y. "Process-Embedded Data Integrity," *Journal of Database Management* (15:1) 2004, pp 87-103.

Lee, Y.W., and Strong, D.M. "Knowing-Why About Data Processes and Data Quality," *Journal of Management Information Systems* (20:3) 2003, pp 13-39.

Lee, Y.W., Strong, D.M., Kahn, B.K., and wang, R.Y. "AIMQ: A Methodology for Information Quality Assessment," in: *Information & Management*, Elsevier Science Publishers B.V., 2002, p. 133.

Levitin, A., and Redman, T. "Quality Dimensions of a Conceptual View," in: *Information Processing & Management*, 1995, pp. 81-88.

Levitin, A.V., and Redman, T.C. "Data as a Resource: Properties, Implications, and Prescriptions," *Sloan Management Review* (40:1) 1998, pp 89-101.

Lindland, O.I., Sindre, G., and Solvberg, A. "Understanding Quality in Conceptual Modeling," *IEEE Software* (11:2) 1994, p 42.

Madnick, S., Wang, R., Chettayar, K., Dravis, F., Funk, J., Katz-Haas, R., Lee, C., Lee, Y., Xian, X., and Bhansali, S. "Exemplifying Business Opportunities for Improving Data Quality From Corporate Household Research," Massachusetts Institute of Technology (MIT), Sloan School of Management, Working papers: 4481-04, 2004.

Madnick, S., Wang, R., and Xiang, X. "The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality," *Journal of Management Information Systems* (20:3) 2003, pp 41-69.

Mandke, V.V., and Nayar, M.K. "Cost Benefit Analysis of Information Integrity," in: *Seventh International Conference on Information Quality (IQ 2002)*, 2002, pp. 119-131.

March, S.T., and Smith, G.F. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4) 1995, pp 251-266.

Marschak, J. "Economics of Information Systems," *Journal of the American Statistical Association* (66:333) 1971, pp 192-+.

Marschak, J. *Economic Information, Decision, and Prediction: Selected Essays* D. Reidel, Dordrecht, 1974.

Marschak, J., Radner, R., McGuire, C.B., and Arrow, K.J. *Decision and Organization. A Volume in Honor of Jacob Marschak* North-Holland Pub. Co., Amsterdam, 1972.

Marsh, R. "Drowning in Dirty Data? It's Time to Sink or Swim: A Four-Stage Methodology for Total Data Quality Management," *Journal of Database Marketing & Customer Strategy Management* (12:2) 2005, pp 105-112.

Meltzer, M. "CURARE Drives CRM," in: *Information Management Direct*, 2002.

Messner, W. "The Beauty and Importance of Quality Customer Information," *Marketing Review* (4:3), Fall2004 2004, pp 279-290.

Michnik, J., and Lo, M.-C. "The Assessment of the Information Quality with the Aid of Multiple Criteria Analysis," *European Journal of Operational Research* (195:3) 2009, pp 850-856.

Miller, H. "Information Quality and Market Share in Electronic Commerce," *Journal of Services Marketing* (19:2) 2005, pp 93-102.

Mingers, J. "The Contribution of Critical Realism as an Underpinning Philosophy for OR/MS and Systems," *The Journal of the Operational Research Society* (51:11) 2000, pp 1256-1270.

Mingers, J. "A Critique of Statistical Modelling from a Critical Realist Perspective," in: *ECIS*, 2003.

Mingers, J. "Critical Realism and Information Systems: Brief Responses to Monod and Klein," *Information and Organization* (14:2) 2004a, p 145.

Mingers, J. "Real-izing Information Systems: Critical Realism as an Underpinning Philosophy for Information Systems," *Information and Organization* (14:2) 2004b, p 87.

Missi, F., and, S.A., and Fitzgerald, G. "Why CRM Efforts Fail? A Study of the Impact of Data Quality and Data Integration," in: *38th Hawaii International Conference on System Sciences (HICSS-38 2005)*, 2005.

Moody, D.L., and Shanks, G.G. "Improving the Quality of Data Models: Empirical Validation of a Quality Management Framework," *Information Systems* (28:6) 2003, p 619.

Moody, D.L., Shanks, G.G., and Darke, P. "Improving the Quality of Entity Relationship Models: Experience in Research and Practice," *ER (Workshops)* (1507) 1998.

Moody, D.L., and Walsh, P.A. "Measuring The Value Of Information: An Asset Valuation Approach," in: *Guidelines for Implementing Data Resource Management,* B. Morgan and C. Nolan (eds.), DAMA International Press, Seattle, USA, 2002.

Motro, A., and Rakov, I. "Estimating the Quality of Data in Relational Databases," In Proceedings of the 1996 Conference on Information Quality, MIT, 1996, pp. 94--106.

Myers, M., and Newman, M. "The Qualitative Interview in IS research: Examining the Craft," *Information and Organization* (17:1) 2007, pp 2-26.

Neuman, W.L. *Social Research Methods : Qualitative and Quantitative Approaches*, (4th ed. ed.) Allyn and Bacon, Boston, MA, 2000.

Neumann, J.V., and Morgenstern, O. *Theory of Games and Economic Behavior*, (60th anniversary ed ed.) Princeton University Press, Princeton, N.J. ; Woodstock, 2004.

Orr, K. "Data Quality and Systems Theory," *Communications of the ACM* (41:2) 1998, pp 66-71.

Paradice, D.B., and Fuerst, W.L. "An MIS Data Quality Methodology Based on Optimal Error Detection," *Journal of Information Systems* (5:1) 1991, pp 48-66.

Parasuraman, A., Valarie, A.Z., and Leonard, L.B. "A Conceptual Model of Service Quality and Its Implications for Future Research," *The Journal of Marketing* (49:4) 1985, pp 41-50.

Parssian, A. "Managerial Decision Support with Knowledge of Accuracy and Completeness of the Relational Aggregate Functions," *Decision Support Systems* (42:3) 2006, pp 1494-1502.

Parssian, A., Sarkar, S., and Jacob, V.S. "Assessing Data Quality for Information Products," in: *International Conference on Information Systems*, 1999.

Parssian, A., Sarkar, S., and Jacob, V.S. "Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product," *Management Science* (50:7) 2004, pp 967-982.

Pawson, R., and Tilley, N. *Realistic Evaluation* Sage, London, 1997.

Peffers, K.E.N., Tuunanen, T., Rothenberger, M.A., and Chatterjee, S. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3) 2007, pp 45-77.

Piatetsky-Shapiro, G. "KDnuggets : Polls : Data Mining / Analytic Software Tools (May 2007)," 2007a.

Piatetsky-Shapiro, G. "KDnuggets : Polls : Data Mining Methods (Mar 2007)," 2007b.

Piatetsky-Shapiro, G., and Steingold, S. "Measuring Lift Quality in Database Marketing," *SIGKDD Explor. Newsl.* (2:2) 2000, pp 76-80.

Pipino, L.L., Lee, Y.W., and Wang, R.Y. "Data Quality Assessment," *Communications of the ACM* (45:4) 2002, pp 211-218.

Piprani, B., and Ernst, D. "A Model for Data Quality Assessment," *OTM Workshops* (5333) 2008, pp 750-759.

Pradhan, S. "Believability as an Information Quality Dimension," in: *ICIQ*, MIT, 2005.

Prat, N., and Madnick, S.E. "Evaluating and Aggregating Data Believability across Quality Sub-Dimensions and Data Lineage," Massachusetts Institute of Technology (MIT), Sloan School of Management, Working papers, 2008a.

Prat, N., and Madnick, S.E. "Measuring Data Believability: A Provenance Approach," Massachusetts Institute of Technology (MIT), Sloan School of Management, Working papers, 2008b.

Price, R., Neiger, D., and Shanks, G. "Developing a Measurement Instrument for Subjective Aspects of Information Quality," *Communications of AIS* (2008:22) 2008, pp 49-74.

Price, R., and Shanks, G. "A Semiotic Information Quality Framework: Development and Comparative Analysis," *Journal of Information Technology (Palgrave Macmillan)* (20:2) 2005a, pp 88-102.

Price, R., and Shanks, G.G. "Data Quality Tags and Decision-Making: Improving the Design and Validity of Experimental Studies," *CDM* (176) 2008, pp 233-244.

Price, R.J., and Shanks, G.G. "Empirical Refinement of a Semiotic Information Quality Framework," in: *38th Hawaii International Conference on System Sciences*, IEEE Computer Society, 2005b.

Provost, F., Fawcett, T., and Kohavi, R. "The Case Against Accuracy Estimation for Comparing Induction Algorithms," in: *The Fifteenth International Conference on Machine Learning*, Morgan Kaufmann, 1997, pp. 445-453.

Quinlan, R.J. *C4.5: Programs for Machine Learning* Morgan Kaufmann, 1993.

Rao, L., and Osei-Bryson, K.-M. "An Approach for Incorporating Quality-Based Cost-Benefit Analysis in Data Warehouse Design," *Information Systems Frontiers* (10:3) 2008, pp 361-373.

Redman, T.C. "Improve Data Quality for Competitive Advantage," *Sloan Management Review* (36:2) 1995, pp 99-107.

Redman, T.C. "The Impact of Poor Data Quality on the Typical Enterprise," *Communications of the ACM* (41:2) 1998, pp 79-82.

Redman, T.C. *Data driven: Profiting from Your Most Important Business Asset* Harvard Business School Press, Boston, Mass., 2008.

Romano, N.C., and Fjermestad, J. "Electronic Commerce Customer Relationship Management: An Assessment of Research," *International Journal of Electronic Commerce* (6:2) 2001, pp 61-113.

Romano, N.C., and Fjermestad, J. "Electronic Commerce Customer Relationship Management: A Research Agenda," *Information Technology & Management* (4:2/3) 2003, pp 233-258.

Sarkar, P. "A Paragon of Quality," *Intelligent Enterprise* (5:16) 2002, pp 39-42.

Seddon, P.B. "A Respecification and Extension of the DeLone and McLean Model of IS Success," *Information Systems Research* (8:3) 1997, p 240.

Shankaranarayanan, G., and Cai, Y. "Supporting Data Quality Management in Decision-Making," *Decision Support Systems* (42:1) 2006, pp 302-317.

Shankaranarayanan, G., and Even, A. "Managing Metadata in Data Warehouses: Pitfalls and Possibilities," *Communications of AIS* (2004:14) 2004, pp 247-274.

Shankaranarayanan, G., and Even, A. "The Metadata Enigma," *Communications of the ACM* (49:2) 2006, pp 88-94.

Shanks, G., Arnott, D., and Rouse, A. "A Review of Approaches to Research and Scholarship in Information Systems," Monash University, Melbourne, Australia.

Shanks, G.G., and Darke, P. "Understanding Data Quality and Data Warehousing: A Semiotic Approach," in: *Information Quality,* L. Pipino (ed.), MIT, 1998, pp. 292-309.

Shannon, C.E. "A Mathematical Theory of Communication," *Bell System Technical Journal* (27:3) 1948, pp 379-423.

Shannon, C.E., and Weaver, W. *The Mathematical Theory of Communication*, (Illini Books ed. ed.) University of Illinois Press, Urbana, Ill., 1949, p. 125.

Simon, H.A. *The Sciences of the Artificial* MIT Press, Cambridge, Mass., 1996.

Smith, M.L. "Overcoming Theory-Practice Inconsistencies: Critical Realism and Information Systems Research," *Information and Organization* (16:3) 2006, p 191.

Solomon, M.D. "It's All About the Data," *Information Systems Management* (22:3) 2005, pp 75-80.

Stamper, R., Liu, K., Hafkamp, M., and Ades, Y. "Understanding the Roles of Signs and Norms in Organizations - A Semiotic Approach to Information Systems Design," *Behaviour & Information Technology* (19:1) 2000, pp 15-27.

Stigler, G.J. "The Economics of Information," *The Journal of Political Economy* (Vol. 69:No. 3) 1961, pp 213-225

Stiglitz, J.E. "The Contributions of the Economics of Information to Twentieth Century Economics," *Quarterly Journal of Economics* (115:4) 2000.

Strong, D.M. "IT Process Designs for Improving Information Quality and Reducing Exception Handling: A Simulation Experiment," *Information & Management* (31:5) 1997, p 251.

Strong, D.M., and Lee, Y.W. "10 Potholes in the Road to Information Quality," *Computer* (30:8) 1997, p 38.

Strong, D.M., Lee, Y.W., and Wang, R.Y. "Data Quality in Context," *Communications of the ACM* (40:5) 1997, pp 103-110.

Stvilia, B. "A Workbench for Information Quality Evaluation," in: *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, ACM, Pittsburgh, PA, USA, 2008, p. 469.

Stvilia, B., Gasser, L., Twidale, M.B., and Smith, L.C. "A Framework for Information Quality Assessment," *JASIST* (58:12) 2007, pp 1720-1733.

Tayi, G.K., and Ballou, D.P. "Examining Data Quality," *Communications of the ACM* (41:2) 1998, pp 54-57.

Theil, H. *Economics and Information Theory* North-Holland Pub. Co., Amsterdam, 1967.

Tozer, G.V. *Information Quality Management* NCC Blackwell, Oxford, 1994.

Vaishnavi, V., and Kuechler, W. "Design Research in Information Systems," ISWorld, 2004.

Walls, J.G., Widmeyer, G.R., and El Sawy, O.A. "Building an Information System Design Theory for Vigilant EIS," *Information Systems Research* (3:1), March 1, 1992 1992, pp 36-59.

Wand, Y., and Wang, R.Y. "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM* (39:11) 1996, pp 86-95.

Wand, Y., and Weber, R. "An Ontological Model of an Information System," *IEEE Transactions on Software Engineering* (16:11) 1990, pp 1282-1292.

Wang, R., Allen, T., Harris, W., and Madnick, S. "An Information Product Approach For Total Information Awareness," Massachusetts Institute of Technology (MIT), Sloan School of Management, Working papers: 4407-02, 2003.

Wang, R.W., and Strong, D.M. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4) 1996, pp 5-33.

Wang, R.Y. "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering* (7) 1995.

Wang, R.Y. "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:2) 1998, pp 58-65.

Wang, R.Y., Yang, W.L., Pipino, L.L., and Strong, D.M. "Manage Your Information as a Product," *Sloan Management Review* (39:4) 1998, pp 95-105.

Wang, R.Y., Ziad, M., and Lee, Y.W. "Extending the Relational Model to Capture Data Quality Attributes," in: *Advances in Database Systems*, Springer, 2001, pp. 19-35.

Wang, S., and Wang, H. "Information Quality Chain Analysis for Total Information Quality Management," *IJIQ* (2:1) 2008, pp 4-15.

Wang, Y.R., Kon, H.B., and Madnick, S.E. "Data Quality Requirements Analysis and Modelling," in: *Ninth International Conference of Data Engineering*, Vienna, Austria, 1993.

Welzer, T., Golob, I., Brumen, B., Druzovec, M., Rozman, I., and Jaakkola, H. "The Improvement of Data Quality - A Conceptual Model," in: *European-Japanese Conference on Information Modelling and Knowledge Bases*, IOS Press, Yyteri, Pori, Finland, 2007, pp. 276-281.

Willshire, M.J., and Meyen, D. "A Process for Improving Data Quality," *Data Quality* (3:1) 1997.

Wixom, B.H., and Watson, H.J. "An Empirical Investigation of the Factors Affecting Data Warehousing Success," *MIS Quarterly* (25:1) 2001, pp 17-41.

Yao, Y.Y., Wong, S.K.M., and Butz, C.J. "On Information-Theoretic Measures of Attribute Importance," in: *The Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, Springer-Verlag, 1999, pp. 133--137.

Zhu, B., Shankar, G., and Cai, Y. "Integrating Data Quality Data into Decision-Making Process: An Information Visualization Approach," *HCI (8)* (4557) 2007, pp 366-369.

Appendices

APPENDIX 1

SOURCE CODE FOR GARBLING NOISE PROCESS

The below source code (in JavaScript) implements the "garbling" noise process described in Section 6.3.3.

The code was executed in Internet Explorer as a web page. It accepts a dataset by pasting the values into a HTML form. This JavaScript code then iterates over the dataset, garbling the rows on an attribute-by-attribute basis for a range of different garbling levels. The resulting garbled datasets are output to disk in industry-standard ARFF files for subsequent analysis by the RapidMiner data mining workbench.

```
<script type="text/javascript">
// Global variable
DataSet = new Array();                                  // dataset is two dimensional array
GarbleDS = new Array();
Imputed = new Array();
Header = "";
function loadData() {
// Load data from web page
        alert('loading data ...');
        data = document.details.dataset.value;
        iter = document.details.iter.value;

        var cols=[];
        var rows=data.split('\n');
        alert('Found '+rows.length+' rows');

        DataSet=[];
        for(var r=0; r<rows.length; r++) {
                cols=rows[r].replace(/[\n\r\s]/ig,'').split(',');
                DataSet.push(cols);
        }
        alert('Found '+DataSet[0].length+' cols');
        displayData(DataSet);
        return;
}

function displayData(d) {
```

```
// Display sample data on web page
        var m=document.details.rowmax.value;
        var t=document.getElementById('datatable');
        var tr, td, cb;
        var ch;
        while(ch=t.firstChild)                          // delete existing rows
                t.removeChild(ch);
        tr=document.createElement('tr');                // insert ordinal value checkboxes
        for(var a=0; a<d[0].length; a++) {
                td=document.createElement('td');
                cb=document.createElement('input');
                cb.type="checkbox";
                cb.id="cb"+a;
                td.appendChild(cb);
                tr.appendChild(td);
        }
        t.appendChild(tr);
        for (var r=0; r<m; r++) {
                tr=document.createElement('tr');
                for(var c=0; c<d[r].length; c++) {
                        td=document.createElement('td');
                        td.innerHTML=d[r][c];
                        tr.appendChild(td);
                }
        t.appendChild(tr);
        }
        return;
}

function imputeData() {
// Estimate and replace missing values (if required)
        var tr, t, filename, iv;
        var maxiter=document.details.maxiter.value;
        var d=document.details.droprate.value/100;
        var ord=0;
        var cat={}, Cats=[];
        var catmax;
        var gCount=0, dCount=0;
```

```
alert('Calculating imputed values ...');

for (var a=0; a<DataSet[0].length; a++) {                    // get imputed value
        if (document.getElementById('cb'+a).checked) {    // is it ordinal or nominal?
                ord=0;                                                // ordinal
                for (var r=0; r<DataSet.length; r++)
                        if (DataSet[r][a].search(/[?]/)==-1) {   // test for missing value
                                ord+=parseFloat(DataSet[r][a]);
                                Imputed[a] = ord/DataSet.length;
                                                                      // get mean value
                        }
                else {
                        cat ={};                                      // categorical
                        cat['!temp!']=-1;
                        for (var r=0; r<DataSet.length; r++)
                                if(cat[DataSet[r][a]])
                                        cat[DataSet[r][a]]++;
                                else
                                        cat[DataSet[r][a]]=1;
                                catmax='!temp!'; Cats[a]="";
                                for (var c in cat) {
                                        Cats[a]+=c+", ";
                                        if (cat[c]>cat[catmax])
                                        catmax=c;
                                }
                        Cats[a]=" {"+Cats[a].replace('!temp!,','').replace('?,','')+"}";
                        Cats[a]=Cats[a].replace(/\n|\r/ig,'');
                        Cats[a]=Cats[a].replace(/,\s*}/ig,'}\n');
                        Imputed[a] = catmax;                  // get mode value
                }
        }
alert('Inserting imputed values ...');
var t=document.getElementById('datatable');
tr=document.createElement('tr');                    // insert imputed values
for(var a=0; a<DataSet[0].length; a++) {
        td=document.createElement('td');
        iv=document.createElement('input');
        iv.type="text";
        iv.id="iv"+a;
        iv.value=Imputed[a]
```

```
                if(iv.value.length>5)
                        iv.size="5";
                else
                        iv.size=iv.value.length;
                td.appendChild(iv);
                tr.appendChild(td);
        }
        t.appendChild(tr);
        alert('Building ARFF header ...');
        Header="";
        for (var a=0; a<DataSet[0].length; a++) {
                Header+="@ATTRIBUTE a"+a;
                if (document.getElementById('cb'+a).checked)       // is it ordinal or nominal?
                        Header+=" NUMERIC\n";
                else
                        Header+=Cats[a];
                }
        Header+="@DATA\n";
        alert('Header: '+Header);
        return;
}
function garbleData() {
// Function to apply garbling noise process
        alert('garbling data ...');
        var maxiter=document.details.maxiter.value;
        var d=document.details.droprate.value/100;
        for (var a=0; a<DataSet[0].length-1; a++)  {          // for each attribute (exclude class)
                for(i=1; i<=maxiter; i++) {
                        // for each iteration, starting with 1/maxiter probability of garbling
                        gCount=0, dCount=0, eCount=0;
                        GarbleDS=[];
                        for(var r=0; r<DataSet.length; r++) {         // clone original dataset
                                row=DataSet[r].toString();
                                GarbleDS[r]=row.split(',');
                        }
                        for(var r=0; r<DataSet.length; r++) {                  // for each row
                                if (Math.random()<=d ||
                                GarbleDS[r][a].toString().search(/[?]/)==0) {
                                                                        // if "success" or ? then drop
                                        GarbleDS[r][a]=Imputed[a];// insert imputed value
```

```
                                dCount++;
                        }
                var p=i/maxiter;
                if (Math.random()<=p) {                 // if "success" then swap
                        do {
                                var t=Math.floor(Math.random()*DataSet.length)
                                                                // pick target
                                temp=GarbleDS[t][a]; // swap with current with target
                                GarbleDS[t][a]=GarbleDS[r][a];
                                GarbleDS[r][a]=temp;
                        }  while (document.details.toggle.checked && t==r)
                        gCount++;
                        if (GarbleDS[t][a]!=GarbleDS[r][a])
                                eCount++;
                }
        }

        document.details.attr.value=a;
        document.details.iter.value=i;
        document.details.garbles.value=gCount;
        document.details.drops.value=dCount;
        document.details.errors.value=eCount;
        filename=document.details.setname.value+"-a"+a+"-i-"+i;
        document.details.outfile.value=filename;

        if (document.details.writefile.checked)
                writeToFile(GarbleDS.join('\n').toString());
    }
return;
}

function writeToFile(writeStr) {
// Code for writing tables as text file
// IE specific
// http://www.webreference.com/js/tips/001031.html

    var fname= document.details.basedir.value+document.details.outfile.value+'.arff';
    //alert('filename: '+fname+'\nString: '+Header+'\n\n'+writeStr);
    var TristateFalse = 0;
```

```
        var ForWriting = 2;
        try {
                var myActiveXObject = new ActiveXObject("Scripting.FileSystemObject");
        }
        catch(e) {
                alert('Cannot write to file - failed creating ActiveXObject.');
                document.details.writefile.checked=false;
                return;
        }

        myActiveXObject.CreateTextFile(fname);
        var file = myActiveXObject.GetFile(fname);
        var text = file.OpenAsTextStream(ForWriting, TristateFalse);
        //text.Write('@RELATION
        '+document.details.outfile.value+'\n'+Header+'\n\n'+DataSet.join('\n').toString()+'\n'+
        writeStr);   // include original
        text.Write('%eCount                    '+document.details.errors.value+'\n'+'@RELATION
        '+document.details.outfile.value+'\n'+Header+'\n\n'+writeStr);   // garbled only
        text.Close();
}
</script>
```