UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato di Ricerca in Ingegneria della Produzione
Dipartimento di Ingegneria Chimica, Gestionale, Informatica e Meccanica
Settore SECS-S/02 - Statistica per la Ricerca Sperimentale e Tecnologica

ADVANCED STATISTICAL TOOLS FOR SIX SIGMA AND OTHER INDUSTRIAL APPLICATIONS

IL DOTTORE
ANNA ERRORE

IL COORDINATORE
PROF. SALVATORE GAGLIO

IL TUTOR
PROF.STEFANO BARONE

CICLO XXV
ANNO 2015

Table of Contents

Abstract

Six Sigma is a methodological approach and philosophy for quality improvement in operations management; its main objectives are identifying and removing the causes of defects, and minimizing variability in manufacturing and business processes. To do so, Six Sigma combines managerial and statistical tools, with the creation of a dedicated organizational structure.

In this doctoral thesis and the three years of study and research, we have had the purpose to advance the potential applications of the methodology and its tools; with a specific attention on issues and challenges that typically prevent the realization of the expected financial and operational gains that a company pursue in applying the Six Sigma approach. Small and medium sized enterprises (SMEs), for instance, very often incur into such issues, for structural and infrastructural constraints. The overall application of the methodology in SMEs was the focus of the initial research effort and it has been studied with a case study approach.

Then, on this basis, most of our research has been turned to the rigorous methodological advancement of specific statistical tools for Six Sigma, and in a broader sense, for other industrial applications. Specifically, the core contribution of this doctoral thesis lies in the development of both managerial and/or statistical tools for the Six Sigma toolbox. Our work ranges from a decision making tool, which integrates a response latency measure with a well-known procedure for alternatives prioritization; to experimental design tools covering both planning and analysis strategies for screening experiments; to, finally, an initial effort to explore and develop a research agenda based on issues related to conjoint analysis and discrete choice experiments.

Acknowledgements

It is very difficult for me to sufficiently and properly acknowledge all the people that certainly deserve the right to be mentioned here because of their role in my life, in my doctoral program and in the research work that will be discussed in the following chapters. Some of them have clearly and directly influenced me, inspired me, helped me and sustained me, both emotionally and practically. Some others have had a more indirect, nuanced and subtle role but definitely, in different ways, as important as the one of the first group.

Conscious that my words would never be good enough to express my feelings, but confident that all these people know me sufficiently well to understand that this who I am, and to understand those feelings without needing that I explicitly say any more than a few simple words; I want to just first thank my advisor and now also my trustworthy friend, Stefano Barone. It is his merit (should I say fault?) if I applied to the PhD program at the University of Palermo; if I persisted and resisted to several moments of discouragement and lack of confidence; if I accomplished small and big goals; if I embraced the adventure of going to the US; of pursuing a Six Sigma Black Belt certification of the ASQ; and the list goes on and on … How could I ever mention all his merits? I will stop here, I'm sure he will forgive this incompleteness because he knows how long this list is.

Other two key people in my journey also cause me a bit of embarrassment in not being able to express the proper acknowledgement that they deserve. They have been my teachers, my advisors, my references in US, and so much more. I am talking about Chris Nachtsheim and William Li. Two people who have taught me so much and who may be mistakenly considered as having a similar role in my life and my time in Minnesota. The truth is that they both have played an incredibly fundamental role in my journey, but in two totally different way; and again, each of them knows exactly what I am talking about.

I'm, of course, very grateful also to all the other people who I had the honor and pleasure to learn from and to work with: Alberto Lombardo, Bradley Jones, Therese Doverholt, and all my Professors at the University of Palermo and at the Carlson School of Management.

Moreover, so many other people, my family, my friends and all the wonderful people I had the chance to cross on my path, have been precious little pieces of this puzzle.

To write this section is an extremely hard task for me, but what these people have given to me it would never be confined in a few sentences of an acknowledgement section; they are, and they still will be, part of my life and my greatest hope is to truly find my way to thank each one of them, probably my own unique way. .

I am grateful to them all and I just wish to have them always in my my life and made them proud in years to come.

Introduction

Six Sigma is a methodological approach and philosophy for quality improvement in operations management. The main objectives in the Six Sigma framework can be summarized as identifying and removing causes of defects and minimizing variability in manufacturing and business processes. The very most critical success factor in the Six Sigma approach is the well-known optimal use of a combination of managerial and statistical tools in conjunction with the creation of a dedicated organizational structure.

Six Sigma is a framework that has developed and spread out in many and different industries, it has gained its momentum in the field, and received attention in research in the operations management literature; especially in the more holistic view of the methodology, its diffusion and impact on organizational performance (both operational and financial). However, being also a huge collection of statistical tools, it is a field of great interest for statisticians as well.

The research work behind this doctoral thesis has been conducted with the purpose to advance the potential applications of the methodology and its tools; with a particular lens on issues and challenges that typically impede the realization of the expected financial and operational gains in applying the Six Sigma approach. Small and medium sized enterprises (SMEs), for instance, very often incur into such issues, for structural and infrastructural constraints.

In this spirit different phases of this research project have had different objectives. First, an initial effort was aimed towards the investigation of the extent of application and diffusion of the methodology, with particular attention to firms that face greater challenges, such as SMEs. In these cases, the potential application of such powerful methodological approach is bounded by several constraints, both financial and operational. This initial stage of research was grounded on several case studies, carried out in SMEs both in Italy and internationally. Main output of this phase was a deeper understanding of the needs and issues in a constrained industrial environment. Specific findings related to one of the case studies (a Swedish manufacturer) and general discussion on criticalities and issues for the implementation of Six Sigma in SMEs can be found in the first published work: Barone S., Doverholt T., Errore A., and Lombardo A. (2014); Six Sigma in small- and medium-sized enterprises: a Black Belt project in the Swedish steel industry. *Int. J. Six Sigma and Competitive Advantage*.

A second, and more extensive stage of research was then turned into the rigorous methodological advancement of specific statistical tools for Six Sigma, which in fact, in a broader sense, can be generally used in other industrial applications, not limited to the boundaries of Six Sigma.

A first output of this second phase is a decision making tool. Specifically, we integrated the well-known decision making tool called Analytical Hierarchy Process (AHP) with a response latency model. We modeled a new measure of degree of preference in the pairwise comparison task. The measure we use is the time taken to make a decision, we essentially replace weights calculated with a response latency model to the traditional use of a sematic scale. We conducted an empirical test of the model in a setting of service design. The use of a response latency measure, has been found to give a more nuanced measure of the decision maker's preferences and to improve the consistency of responses in the application of the AHP, in comparison with the use of the semantic scale. This work was reported in the second published paper: Barone, S., Errore, A., Lombardo, A. (2014); Prioritisation of alternatives with analytical hierarchy process plus response latency and web surveys; *Total Quality Management & Business Excellence*.

Lastly, most of the research effort in this project and most of the time in this PhD program has been dedicated to probably the most powerful tool in the Six Sigma toolbox: the Experimental Design (Design of Experiment, hereinafter DoE). DoE is a tool for exploring relationships between factors and establishing causal links and it can be leveraged and adopted to different needs and issues related to every phase of the Six Sigma framework. In the so called DMAIC cycle (Define, Measure, Analyze, Improve, and Control), experimental design is very much emphasized in the Improve Phase, but other uses of DoE can be easily found in each of the other phases, and again, in other applications in industry, beyond the Six Sigma framework. In particular, the most recent output of this research project regards the development of designs and analysis strategies for screening experiments.

Screening experiments are a special case of experimental studies, where typically there is very little knowledge about the process or product under study, there are many factors that may influence a response of interest, hypotheses on the nature of the relationships between factors is very difficult to make and assumptions are uncertain. This is very often the case in a Six Sigma project that investigates root causes of variation in a manufacturing or business process. In such a scenario, with many factors and little *a priori* knowledge, both design and analysis issues arise.

We construct small efficient experimental plans for investigation of any number of two-level factors. This work fit into the stream of research on Definitive Screening Designs (or DSDs), which were typically small and efficient designs for three-level factors. Our work has been aimed at the extension of three-level DSDs characteristics to the two-level factors case and it results in a new class of designs that we construct with a coordinate exchange algorithm. These designs have the main feature of being small and highly efficient orthogonal or nearly-orthogonal plans where the main effects are completely de-aliased by any two-factor interaction. This feature is particularly useful in the analysis stage of an experimental study. For the analysis stage we investigated several variable selection methods that are considered suitable for screening experiments where the number of potential factors of interest exceeds the number of observations in the experiment (so-called supersaturated designs); this is the case when in addition to main effects, the experimenter is interested in potentially active interaction effects, or even quadratic effects (the latter only estimable with three-level designs). We contrast and compare several variable selection methods for analyzing DSDs in a comprehensive simulation study where we evaluate the performance of different methods with different cases of active factors, signal to noise ratios, effect sparsity and heredity.

These last stage's findings can be found in two research papers under the review process in international ISI journals: Errore A., Jones B., Li W., Nachtsheim C.J., "Two-level Folded-over Efficient Screening Designs", and "Analysis Strategies for Model Selection with Definitive Screening Designs".

Finally, during the three years of the program, some effort has been devoted to issues related to experimental designs for non-linear models, applied in conjoint analysis and discrete choice experiments. Most of this work is still in progress, and it has a main purpose to extend some current research on linear models designs on the non-linear case. This entire research dedicated to experimental design issues is part of a research proposal awarded in 2013 with the Juran Fellowship Award.

Motivation, objectives and methodology

Probably, this would not fit the editorial style that one should follow in a dissertation, but I feel I need to talk personally and a bit informally about my journey in the PhD program to briefly talk about some motivation and objectives of this work, as well as the methodological approach that has been undertaken.

When I applied to the PhD program in Industrial Engineering at the University of Palermo, I had just completed my Master Program. I had spent my last semester as an exchange student within the Erasmus program at the Chalmers University of Technology in Gothenburg, Sweden. This experience was definitely crucial for me and for my desire to continue with my education in a doctoral program.

During the semester in Sweden I had the chance to take a five months academic course for Six Sigma Black Belt and I had the opportunity to work with a team on an industrial project at a company called Structo Hydraulics. This experience definitely triggered my interest in Six Sigma and applied statistics. My thesis advisor, Stefano Barone, suggested me to start thinking about a doctoral program with a project focused on Six Sigma topics.

Then, once I had the opportunity to embrace the PhD journey, my main research interest was still Six Sigma but I wanted to work on topics that could be of value both within and beyond this framework. Since the beginning most of our research effort and interest was directed toward the application of Six Sigma in small and medium companies. The motivation came from different directions. On the one hand, my personal first experience in Six Sigma was a project in a Swedish SME; during that experience, I had the chance to notice that there are different issues and challenges that a SME faces in implementing the methodology; on the other hand, the possibility to start doing this type of research in Italy, did inspire in me the idea that, in order to understand the extent of the diffusion and knowledge of the methodology in this country, it would be necessary to first better understand and issues related to the types of companies that mainly characterize the Italian industries: the small and medium enterprises.

From these type of simple initial reasoning, we decided to spend some initial time of my program in completing some work on the black belt project at Structo Hydraulics and start other projects in local small companies in Sicily. With this case study approach and some literature review we first worked on a research paper on the implementation of Six Sigma in SMEs, whose authors are Stefano Barone, Alberto Lombardo, Therese Doverholt, and me. We also took the opportunity of finishing some of these projects to discuss in which direction to move in order to accomplish the objective of taking the most advantage from the remaining part of the doctoral program.

At this point we turned from a case study research approach to a more analytical methodological approach directed towards the advancement of specific tools for the Six Sigma toolbox. Also, we wanted to work on tools of general value and applicability within the DMAIC cycle and Six Sigma projects, and beyond the framework in various other contexts of industrial applications of managerial and statistical tools for operations management, marketing, decision making. This is a paper whose authors are Stefano Barone, Alberto Lombardo and me.

A first work in this new approach is based on the development of an analytical integration of a response latency measure in the procedure of the Analytical Hierarchy Process. This research is in line with a broader research stream regarding the use of response latency measures in statistical and decision making models (Barone et al., 2007, 2012). Within this project we obtained and validated empirically a model that allows for a simpler and more intuitive procedure in the AHP method, intuitive calculations for the researchers, easy and shorter procedure for the respondents, and higher overall consistency indexes of the responses.

This work and the one mentioned in the initial phase are both published papers, some of the issues discussed in the papers are also included in the first two following chapters of this dissertation. The case studies related to these two papers are attached in appendix, should the reader be interested in the details.

Most of the remaining time and research work of this dissertation had been spent, as said, on issues related to experimental design modelling and analysis. This part of my research during the doctoral program was mainly conducted at the time I was being hosted as a visiting PhD student at the Carlson School of Management, University of Minnesota, in Minneapolis, USA.

I planned to spend in Minneapolis one academic year between fall 2012 and spring 2013, in order to accomplish part of the requirements of this PhD program and to have the opportunity to work for some months with Professors William Li and Christopher Nachtsheim. The line of research that I had the chance to join was very interesting since the beginning and it turned out to be occupy the rest of my PhD program, changing my plans to stay in Minnesota until the end of the program, December 2014.

The research related to this topic in this dissertation is part of a larger research project that we developed in 2013 and that it is still ongoing. Some of the results and the research papers written on this work are under various stages of the review process and some others are still work in progress and plans for the near future.

The whole research project being conducted with Professors Li and Nachtsheim has so far involved and continue to involve other distinguished scholars at the Carlson School and not only. To be fair with the entire project in the next few pages I will report some of the key points of the research proposal called “Definitive Screening Designs and Discrete Choice Experiments for Quality Improvement”. Two chapters will be dedicated to some current results and ongoing work. This project has been awarded by the Juran Research Center with the Juran Fellowship award in 2013.

Other than what will be mentioned in the following chapters of this doctoral thesis, other research works completed or ongoing will be just mentioned in the next section, and they will not be discussed in details. My own education as a future scholar is still ongoing and I feel like I am learning every day form my advisors and professors, and from my colleagues. By no means is this doctoral thesis the conclusion of my scholarly education or the completion of the projects we are enthusiastically working on.

Definitive Screening Designs and Discrete Choice Experiments for Quality Improvement

A substantial part of this thesis and the research work conducted during this three years program is mainly linked to methodological work in the design of experiments (DOE) for industrial applications of quality improvement projects. In quality applications, DOE plays a fundamental role in a variety of situations, from design and development of new products to product or process improvement practices (Evans and Lindsay, 2002). Of course, as we will repeat many times, DOE is one of the most important tools used in the application of the Six Sigma Methodology (Magnusson et al., 2003; Barone and Lo Franco, 2012). In particular, it is fundamental to the improvement phase of the DMAIC (Define, Measure, Analyze, Improve, and Control) cycle.

Experiments, whether statistically designed or not, are a component of the learning process. We experiment to learn. How well one succeeds will be a function of adherence to the scientific method, the most rapid means for speeding the learning process (Juran and Godfrey, 1999).

Design of experiments is a branch of applied statistics that deals with planning, execution, analysis and interpretation of a series of tests to evaluate rigorously the potential cause-and-effect relationships between inputs and outputs of a process. A well planned and executed experiment provides clear information about the effect on a response variable due to one or more factors. Many of the current statistical approaches to designed experiments originate with the work of R. A. Fisher in the early part of the 20th century. Fisher demonstrated how taking the time to seriously consider the design of an experiment prior to testing can lead to clear and valid inferences the cause-and-effect relationships between a response and a set of potential causal factors (ASQ website).

In quality applications DOE plays a fundamental role, from design and development of new products, to product or process improvement practices. These applications are found equally in both manufacturing and service operations. The large variation in products, processes, and scientific applications motivates the need for an array of tools and techniques that can both efficiently and effectively meet different scenarios' characteristics.

In applying DOE in industry, one of the most critical tasks is the choice of design. The statistical analysis to be conducted is determined by choice of design, and a well-designed experiment is generally easy to analyze. Assumptions about the nature of the cause-and-effect relationships between the factors and the response (i.e., about the operative regression model) are necessary in order to select an appropriate design. If the true model turns out to be significantly different from the assumed model, the design chosen may be inefficient and conclusions drawn from the experiment can be misleading.

The effectiveness of any experiment depends critically on the validity of assumptions made by the experimenter about the product or process that is under study. There are generally one or more quality characteristics or metrics that we wish to improve. Prior to experimentation, the experimenter must state his or her beliefs about the controllable and uncontrollable factors that may affect the response. This is frequently accomplished through the use of brainstorming sessions and the use of cause-and-effect diagrams. For the best experimental design, it is not only necessary to identify potential causal factors. It is also necessary to articulate the nature of the effect that each factor has on the quality response. For example, does the quality response change linearly with changes to a potential factor, or nonlinearly? Are interactions among factors a possibility? These prior beliefs are expressed mathematically in the form of an assumed regression model. If we consider the output to be the quality characteristic that we obtain for a given set of factor-level inputs, our model takes the form:

$$\boldsymbol{output\ (y) = f(controllable\ factor\ level\ inputs\ (x_1, \ldots, x_m))}$$

Here $\boldsymbol{f}$ represents the functional form of the regression model. Given a good guess at or approximation to the true regression model, a best experimental plan can be formulated. On the other hand, if little is known in advance about the form of the regression model, a design may turn out to be sub-optimal or completely inadequate.

This problem is especially critical when the experimenter has little prior knowledge about the product or process under study, such as during the early stages of an investigation. In such cases, screening experiments are often employed in an effort to identify the set of active factors. Subsequent, follow-up experiments employing only the active factors is then required to determine the effect of interactions and other nonlinearities. Clearly, the correct identification of the active factors is crucial.

During the early phases of research, little is generally known about which potential factors are truly important or active, and the form of the regression model. Historically, statisticians have advised investigators to posit linear, main-effects (ME) models, and to choose small designs that employ large numbers of factors in pilot studies referred to as screening experiments. The basic idea is to test many potential factors at only two levels each (e.g., low versus high, on versus off, etc.) with very few runs, in an effort to determine which factors are active. Subsequently, follow-up experiments are conducted with the active set of factors to investigate potential nonlinearities and interactions.

There are several major limitations with this approach (Jones and Nachtsheim, 2011a). First, the use of only two levels leads to the use of linear approximations, which may not be particularly accurate. Moreover, for each factor under consideration, there is often an existing factor-level setting that his thought to be best. Process engineers and product developers will generally want to include this level in the experiment, while also exploring the effects of increasing or decreasing this level. This leads naturally to a demand for three levels. Finally, if there are interactions among factors, these interactions can seriously bias the inferences made from the experiment.

Standard approaches to factor screening, such as the 2^{k-p} resolution III fractional factorial designs and Plackett and Burman designs all suffer from full or partial confounding of main effects and two-factor interactions. Resolution III 2^{k-p} directly confound main effects and two-factor interactions. Consider, for example, a resolution III design with 3 two-level factors that is to be conducted using 4 runs based on the defining relation I = ABC. For this design, the estimate of the any main effect will be completely confounded with a two-factor interaction. For instance the main effect of factor A is completely confounded with the interaction with the interaction between factors B and C. Thus if the main effect for factor A is statistically significant, we cannot tell if this was the result of a real A effect, a real BC interaction effect, or some combination of the two. Our results are ambiguous unless we can assume two-factor interactions do not exist. This undesirable feature could be avoided by switching to a design with a higher resolution, such as a resolution IV or a resolution V fractional factorial design. In a resolution IV design, all main effects are orthogonal to each other and clear of (i.e., not confounded with) any two-factor interaction; however, some two factor interaction are confounded with each other. If a resolution V design is used instead, all main effects are orthogonal to each other and clear of any two- or three-factor interactions; main effects are confounded only with four-factor or higher-order interactions.

Moreover, two-factor interactions are not confounded with each other, but may be confounded with three-factor interactions. This preferable aliasing structure comes at the cost of increasing the sample size. A rough rule of thumb is that fractional factorial designs generally double with each unit increase in resolution. For instance, in order to study 9 factors, a resolution III design would require $\mathbf{2}_{\boldsymbol{III}}^{\mathbf{9-5}} = \mathbf{16}$ runs, and a resolution IV design would require $\mathbf{2}_{\boldsymbol{III}}^{\mathbf{9-4}} = \mathbf{32}$ runs.

Definitive screening, as introduced by Jones and Nachtsheim (2011b), provides a solution to this problem that does not require a doubling of runs. DSDs provide estimates of main effects (henceforth MEs) that are unbiased by any second-order effect, require only one more than twice as many runs as there are factors, m, and avoid confounding of any pair of second order effects. Also, for designs having six factors or more, these designs project to efficient response surface designs with three or fewer factors. One limitation of these designs is that all factors must be quantitative, in order to be set at three levels, of which one is the center point. This paper represents breakthrough from the standpoint of research and practice, and it has opened an entirely new stream of research. For this reason, received two prestigious awards from the American Society for Quality. These are the 2011 Brumbaugh Award for “the paper that has made the largest single contribution to the development of industrial application of quality control,” and the 2012 Lloyd S. Nelson Award that recognizes the paper published by the Journal of Quality Technology that has had the “greatest impact on practitioners” in the preceding year.

The goal of our research is to expand upon the idea of definitive screening designs in a number of directions. Often in practice the experimenters deal with categorical factors that cannot be set at the center point as required by a standard three-level definitive screening designs. This project’s first objective is to introduce a new class of two-level screening designs. The second phase will be directed toward the development of efficient two- and three-level DSDs of higher resolution, in order to allow clear estimation of all main effects and two-factor interactions. The third phase is directed to address an even more challenging advancement, which could potentially represent the major breakthrough: the extension of the definitive screening approach to binary (discrete) responses. Examples of such responses include “defective, not defective,” “within spec, not within spec,” and so on. Design of experiments in the presence of discrete responses is more complex because the models used for estimation are nonlinear and the optimality of the designs depend on the unknown regression parameters. These designs are frequently used in quality applications, and in conjoint analysis experiments in new product development.

To summarize, this research project aims to expand the class of DSDs in three ways.

- Phase 1: Screening Designs for two-level factors. In their original paper, JN did not consider screening in the presence of two-level categorical factors. They later showed how such factors could be added to existing DSDs (Jones and Nachtsheim, 2013). Interestingly, JN never considered the case where all factors are categorical and have only two levels. Research in this phase is directed toward creating a class of two-level designs that retain many of the advantages of DSDs. We have made significant progress toward this goal and the main results will be presented in the third chapter.
- Phase 2: Definitive Screening Designs for estimation of response surface models. Here we will seek to create a new set of experimental designs for response surface exploration. These goal is to create designs that are: (1) capable of estimating all main effects, curvatures and two-factor interactions, such that (2) all second-order effects are unbiased by the potential presence of three-factor interactions. This is analogous to definitive screening designs, in which all first-order effects are unbiased by the potential presence of second-order effects.
- Phase 3: Definitive Screening Designs for Logistic Regression, with Application to Discrete Choice Experiments. When the quality characteristic is discrete, for example when the response is “defective or not defective,” the regression model is nonlinear. DSDs are not strictly

applicable in this case. However there will be designs that minimize the bias due to second-order effects when linear models are used. Our goal here is to find and characterize such designs. Initial work in this area is related to the work presented in the fourth chapter.

In Phase 1, we first conducted a literature review to identify existing designs that meet the unbiasedness requirement of DSDs. Wherever we found gaps, we employed numerical optimization methods (specifically the coordinate exchange algorithm of Meyer and Nachtsheim, 1995) to seek new designs. For many combinations of sample sizes and numbers of factors our search has identified existing designs; in other cases we have created new designs. Work in the second and third phases is just getting underway. Jones and Nachtsheim (2011b) showed how their minimal aliasing approach to design construction might be used to create robust response surface designs. We intend to build on that approach in our Phase 2 research. Phase 3 work will require new Bayesian approaches to nonlinear design. This approach will require analytical modeling and will also build on the work of Li, et al., (2013).

Structure of the thesis

After this introductory chapter, this doctoral thesis is organized as follow. Each one of the next chapters is mainly devoted to a specific research study conducted during this three year program.

The first chapter gives an overview of the Six Sigma methodology applied in Small and Medium-sized Enterprises. The related outcome of this work is a research paper recently published:

- Barone, S., Doverholt, T., Errore, A., & Lombardo, A. (2014). **Six Sigma in small–and medium–sized enterprises: a Black Belt project in the Swedish steel industry.** *International Journal of Six Sigma and Competitive Advantage*, 8(2), 125-146.

All the following chapters discuss a specific tool or set of tools developed or advanced for the Six Sigma toolbox. The second chapter introduces an integration of the Analytical Hierarchy Process and a Response Latency model. The related published paper:

- Barone S., Errore A., Lombardo A. **Prioritization of Alternatives With Analytical Hierarchy Process Plus Response Latency and Web Surveys**. *Total Quality Management & Business Excellence*, Vol. 25, 7-8. (2014).

The third chapter explores design and analysis issues for experimental design, specifically screening experiments. Two research papers currently under various states of advancement in the respective review process are:

- Errore A., Jones B., Li W., Nachtsheim C.J. **Two-level Minimal Foldover Screening Designs.**
- Errore A., Jones B., Li W., Nachtsheim C.J. **Analysis Strategies for Model Selection with Definitive Screening Designs**.

The fourth chapter explores the issues of experiments based on non-linear models, as for instance those involved in conjoint analysis studies. The related paper is a working paper:

- Errore, A., Donohue, K., Nachtsheim, C.J. **Discrete Choice Experiments: designing optimal experiments accounting for statistical efficiency and behavioral consequences**

The papers resulting from these three years of the doctoral program will be discussed in the core chapters of this thesis.

Other papers written during these three years or currently in progress, are simply listed here, but they will not be specifically discussed in the remainder of this document:

- Errore A., Linderman K., Lucianetti L. **The Use of Financial and Non-Financial Performance Measures: A Contingency Perspective**. *Joint Statistical Meeting 2013 Proceedings.*
- Barone S., Errore A., Lombardo A. **A class of Regression Models with Weighted Predictors**. Working paper.
- Errore, A., Shah, R. (2014) **Experimental research in Operations Management: opportunities and challenges.** Working paper.

References

- Barone, S., and Lo Franco, E. (2012). Statistical and managerial techniques for Six Sigma methodology: theory and application. Wiley. com.
- Barone, S., Lombardo, A., and Tarantino, P. (2007). A weighted logistic regression for Conjoint Analysis and Kansei Engineering. *Quality and Reliability Engineering International*, 23(6), 689-706.
- Barone, S., Lombardo, A., and Tarantino, P. (2012). A heuristic method for estimating attribute importance by measuring choice time in a ranking task. *Risk and Decision Analysis*, 3(4), 225-237.
- Evans, J. R., and Lindsay, W. M. (2002). The management and control of quality. Third edition
- Jones, B. and Nachtsheim, C. J. (2011a). “Efficient Designs with Minimal Aliasing”. Technometrics 53, pp. 62–71.
- Jones, B. and Nachtsheim, C. J. (2011b). “A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects”. Journal of Quality Technology 43, pp. 1–15.
- Jones, B. and Nachtsheim, C. J. (2013). “Definitive Screening Designs with Added Two-Level Categorical Factors”. Journal of Quality Technology 45, pp. 121–129.
- Juran, J. M., and Godfrey, A. B. (1999). “Juran’s quality handbook (Vol. 2)”. New York: McGraw Hill.

- Li, W., Nachtsheim, C. J., Wang, K., Reul, R., and Albrecht, M. (2013). "Conjoint Analysis and Discrete Choice Experiments for Quality Improvement". Journal of Quality Technology, 45(1).
- Magnusson, K., Kroslid, D., Bergman, B., Hyhnen, P., and Mills, D. (2003). Six sigma: the pragmatic approach. Studentlitteratur.
- Meyer, R. K. and Nachtsheim, C. J. (1995). "The Coordinate-Exchange for Algorithm Exact Constructing Optimal Experimental Designs". Technometrics 37(1), pp. 60–69.

Six Sigma in Small- and Medium-sized Enterprises

Introduction

Many big companies around the world have solid Six Sigma infrastructures and it is easy to find in literature successful case studies regarding their implementation of Six Sigma.

Conversely, small and medium-sized enterprises (SMEs) generally suffer lower attention in the literature related to this topic. Plausible explanations might be that SMEs have only more recently approached the methodology; that they have weaker connection with the academia; or they do not rigorously pursue the frameworks shown in the literature, when they introduce the Six Sigma philosophy in their business.

Our first research study is based on an industrial Six Sigma Black Belt project carried out in a Swedish medium-sized company, Structo Hydraulics AB. The company produces steel tubes mainly for hydraulic applications. The project focused on the improvement of warehouse activities, in particular related to cutting processes. This Black Belt project was also part of a Six Sigma education at the Chalmers University of Technology in Gothenburg in 2011.

The case study offers the opportunity to discuss general and specific issues that a SME has to face in the implementation of the Six Sigma methodology. Discussion triggered by this case study contributes to connect the academic debate to the practical experience in industry.

We don’t report here the case study details (which can be found in the published paper), but rather we use the general discussion on the application of Six Sigma in SMEs. This initial study gave us an overview of the issues faced by small companies wishing approach the Six Sigma methodology and philosophy. Inputs from this study were precious for the following development of this doctoral thesis.

Six Sigma: origin and definitions

Six Sigma found its origins in industry and it gathered attention in academia only in recent years. One of the main issue with research in Six Sigma is the lack of theory grounding and the disagreement in definitions.

In Behara et al. (1995) Six Sigma is defined as “the rating that signifies “best in class”, with only 3.4 defects per million units or operations”. Antony (2002) calls it a strategy, “a business performance improvement strategy that aims to reduce the number of mistakes/defects – to as low as 3.4 occasions per million opportunities”; then in Banuelas and Antony (2003) Six Sigma is identified as “a philosophy that employs a well-structured continuous improvement methodology to reduce process variability and drive out waste within the business processes using statistical tools and techniques".

In Kwak and Anbari (2006), it is also defined as "a business strategy used to improve business profitability, to improve the effectiveness and efficiency of all operations to meet or exceed customer needs and expectations".

Andersson et al. (2006) define Six Sigma as an "improvement program for reducing variation, which focuses on continuous and breakthrough improvements". Bendell (2006) says it is "a strategic, company-wide, approach ... focusing on variation reduction, projects have the potential of simultaneously reducing cost and increasing customer satisfaction". Black and Revere (2006) give the multiple view of "a quality movement, a methodology, and a measurement. As a quality movement, Six Sigma is a major player in both manufacturing and service industries throughout the world. As a methodology, it is used to evaluate the capability of a process to perform defect-free, where a defect is defined as anything that results in customer dissatisfaction". Chakrabarty and Tan (2007) use again the term "a quality improvement program with a goal of reducing the number of defects to as low as 3.4 parts per million opportunities or 0.0003 per cent".

Unclear definitions are caused by a general lack of theoretical underpinnings of Six Sigma. Linderman et al. (2003) is a first attempt to link Six Sigma to goal theory.

Schoeder et al. (2008) is a very well-known and cited article that aims at the defining Six Sigma and its underlying theory. The authors conduct a careful literature review on the available practitioner literature together with field observations through a case study approach based on 2 companies, one in the manufacturing industry, one in services; gathering field data on several successful and unsuccessful projects ran at both companies, which, moreover, have different depth of experience in Six Sigma implementation. In this paper the proposed definition of Six Sigma is “an organized, parallel-meso structure to reduce variation in organizational processes by using improvement specialists, a structured method, and performance metrics with the aim of achieving strategic objectives”.

This definition identifies four fundamental constructs: parallel-meso structure, improvement specialists, structured method, and performance metrics.

Even though a deep and exhaustive discussion on Six Sigma definitions and theory is beyond the scope here, it is important to remark that this lack of clarity and standards is partly the cause of uneven and problematic diffusion of the methodology among small and medium sized enterprises.

Literature on Six Sigma in SMEs

Most of the academic literature on Six Sigma has focused on the methodology itself and case studies are usually related to big companies such as Motorola, General Electric, AlliedSignal, Sony and ABB (Snee 2004). Only few publications (see e.g., Sarkar 2007; Bewoor & Pawar 2010) concern case studies of Six Sigma implementation in Small and Medium-sized Enterprises (SMEs henceforth).

The discussion about the implementation of the Six Sigma methodology in SMEs, concerns the applicability of the same established framework usually implemented by big companies. The methodology was in fact originally developed and successfully applied in companies of big size; therefore an issue arises: “can Six Sigma be adapted to SMEs or the company has to be a big one to successfully follow the Six Sigma pathway?”

The amount of human and financial resources necessary to build a solid Six Sigma infrastructure certainly is an advantage for big companies. However, is that a sufficient reason to be skeptical about the potential success of the Six Sigma methodology in SMEs?

Recent studies have investigated in this direction (Antony 2008; Kumar 2007; Prasada Reddy & Venugopal Reddy 2010; Thomas & Barton 2006).

The aim of this paper is to contribute to an open debate on how to implement Six Sigma in SMEs. A case study of a medium-sized Swedish company which recently approached the Six Sigma methodology, will be presented in the following Sections of the paper. The aim is to analyze this case study, discussing some of the issues that a SME has to face in order to improve the quality of its business. Some key success factors are discussed in the literature. It is useful to recognize them in the field studies and integrate this discussion with new and unexpected issues related to the daily activity of an industrial project. Some factors reveal their key role in the implementation of the Six Sigma methodology in the company, e.g. the management commitment and the training on the methodology. In this company under study, some people were already Black Belts, and the top management of the company sponsored the participation of another quality engineer to a 5-months Black Belt course running at the Chalmers University of Technology in Gothenburg. The attendance of the course gave the opportunity to carry out a project, following the DMAIC framework, and using the statistical and the managerial techniques learnt during the classes (Barone & Lo Franco, 2012; Magnusson et al., 2003).

Six Sigma and SMEs: perspectives of academicians and practitioners

An interesting survey of practitioners' and academicians' perspectives on the topic of Six Sigma in SMEs can be found in Antony (2008). Some insights offered by several of those interviews can be applied and extended by the practical experience gained in the case study presented in this paper.

Antony (2008) reports an interesting starting point made by Thomas Pyzdek (Pyzdek Consulting, USA). Pyzdek points out that there are two facets of Six Sigma: the approach and the infrastructure. On one hand, in terms of possibility to build the infrastructure, big and small companies are incomparable; but, on the other hand the approach can be the same. The lack of sufficient human and financial resources that a SME can devote to Six Sigma projects could be a limitation, but the question if the approach can be equally and successfully adopted should be separately considered.

In this respect, one may start to think about how much SMEs are actually different from larger companies. In another interview, Larry Smith (Juran Institute, USA) reflects on the fact that since large companies tend to be organized into small and medium-sized departments and operations, one can actually consider that what applies to large companies also applies to SMEs. Roger Hoerl (GE Global Research, USA) suggests that not only the applicability of the methodology could be equally suitable to both large companies and SMEs, but even that in some cases the success of a Six Sigma project can be more impressive in SMEs, in terms of percentage of revenue basis. The possibility to gain substantial cost reduction is always a good incentive for a SME to start implementing Six

Sigma in their business, especially if this implementation can result in more impressive achievements when comparing the cost reduction with the revenues. Also, it can be generally demonstrated that less ambitious goals are easier to be achieved. “In my experience, the results are usually quicker and more visible in smaller companies” confirmed Matthew Hu (American Supplier Institute, USA). The challenge is to prove that those kind of achievements are possible, but perhaps even more important is to show how to do it. It is especially true for SMEs that the management commitment has the highest influence in reaching the goals of the projects. In this perspective, Jiju Antony himself (University of Strathclyde, UK) strongly advices that the senior management team in SMEs must be visibly supportive of every aspect of a Six Sigma initiative. As small companies are more agile, it is much easier to buy in management support and commitment, as opposed to large organizations.

In another interview, Ronald Snee (Tunnell Consulting, USA) identified some critical points in the implementation of Six Sigma in SMEs. Such criticalities include the difficulty for the teams to stay focused on projects, because of many other concurrent tasks of the daily work; moreover it is very rare for the Black Belts to be resources full time dedicated to these projects; many employees have several functions, unlike employees of larger organizations. It is clear that the availability of human resources is of utmost importance; the lack of specialization and employees responsible of different functions is quite a common issue for SMEs, which very often cannot dedicate a person or a team, full time to Six Sigma projects. Rick Edgeman (University of Idaho, USA) observed that SMEs are commonly confronted by limited human capital, especially in terms of specializations and training budgets. For this reason it is important to ensure that the early applications of the Six Sigma methodology have a very high probability of success. How to ensure it? In this scenario the support from the academicians and experts can be very important. According to Thong Ngee Goh (National University of Singapore) SMEs should first invite Six Sigma experts to look into their operations and potential areas of applications. Then SMEs must tailor training programs to suit the needs of their specific organization. It may be the case that certain projects can result heavy on 'DMA' and light on 'IC', at least initially. This distinction, a kind of separation between the first phases of the DMAIC cycle and the Improve and Control phases, is more common than one might think; the case study that is going to be presented highlights this issue among the others. Also, as the previous evidence states, such approach can be initially quite a good result for inspiring a deep change in the way SMEs conduct their business.

Some other interesting points for discussion can be found in a survey conducted on manufacturing SMEs in the United Kingdom (Kumar 2007). The aim of the survey was to assess the status of Six Sigma implementation in SMEs. Confirming some arguments discussed in Antony (2008), the following factors were found to be critical: top-level management commitment and direction of Six Sigma projects; decision making approach based on data, a data-focused approach able to look into all the input variables of the key processes; ability to measure the first-time pass rate for each process; being people-oriented, in the sense that the implementation must be team-based and involve the employees at the shop floor.

Among these factors, the senior management commitment is identified as the most critical element for the success of Six Sigma within SMEs. The findings of the study were in line with existing Six Sigma literature on critical success factors. The same study identified poor training and resource availability as the highest barriers.

SMEs may require more support and guidance from consultants/experts in order to effectively and efficiently integrate their daily activities with a well-organized way of implementing statistical and managerial techniques, as in Six Sigma projects. To a certain extent, when dealing with quality methods, the impediments of cost, time and relative impacts do not fully explain why SMEs have not adopted them to any significant degree (Husband & Mandal 1999). This is evidenced through the low implementation rates of common quality methods, such as quality systems and quality certification. A lack of understanding may also prevent SME owners/operators/managers and other SME interest groups from being able to justify the use of these methods.

Summarizing the reflections from academicians and practitioners, it is possible to conclude that:

- Six Sigma could be successfully applied either to SMEs or big companies. The sustainability of the approach can be independent from the ability to build a solid infrastructure.
- The amount of savings and financial results obtained by SMEs can be sometimes small, but still significant for the company.
- More than in large companies, in SMEs it is vitally important to have a strong commitment of the top management to successfully pursue the goals of Six Sigma initiatives, especially for the earliest projects.

- The involvement of right people in the projects can help overcoming the inherent difficulties incurring when introducing the approach in a SME. Involving operators of the shop floor, even if not fully committed to the projects, can help the project team to deeper understand the processes and the root causes of the problems.
- A rigorous Six Sigma training and/or support of consultants and academicians can supply the needed knowledge for a good implementation of the methodology in companies that are starting to approach Six Sigma.
- Simple projects and more easily achievable goals are a good way for SMEs to start with Six Sigma. Even unbalanced DMAIC cycles, e.g. heavier in DMA phases and lighter on the IC phases, can be a good starting point.

The case study presented in the following Section reports and highlights some of the previous issues, critical success factors and peculiarities of the implementation of the Six Sigma methodology in a Swedish manufacturing company. Along with the presentation of the case study, other aspects will contribute to the discussion, such as the issues related to data recording and the trustworthiness of the information system, and the prioritization of short and long term improvement solutions, due to time and cost constraints.

Reflections from the case study

When the theory of the Six Sigma methodology becomes practice, it is surely successful with a rational and well organized project work carried out by a skilled and motivated team. The DMAIC framework is effective as demonstrated by successful projects all over the world and across many industries. However, it is important to skip the temptation to solve everything at once, to consistently stay within the project scope.

The case study here discussed was related to an early approach to Six Sigma by the company and gave the chance to the company to foresee the systematic adoption of the methodology. It raised reflections about the strengths and weaknesses of running a Six Sigma project in a SME, confirming that SMEs have some characteristics that should be always carefully considered. In this case the company showed interest in adopting the methodology, pushed by one of its biggest customer. The aim of introducing Six Sigma in the company could not achieve the goal of building a solid Six Sigma infrastructure, but to introduce the philosophy, educate people and get immediate benefits with some projects.

A limited availability of human resources to fully dedicate to the project did not limit the ambition to achieve substantial improvement in the processes under study, to gain cost reduction and visible results. From different hierarchy levels, people in the company were willing to participate to the project, but unfortunately, the most important commitment was needed after the end of the academic course. As noted by T.N. Goh (Antony, 2008), it was easier to put more effort in the DMA part of the cycle, the hardest was to implement the solution and to follow up the results and to standardize the improvements.

On one hand, the experience here presented, confirms some initial fears of substantially adopt Six Sigma in a SME and contributes to the scientific discussion with more findings. On the other hand, and the bright side, it is useful to take advantage from the lesson learned in order to better frame how SMEs should approach the methodology. It is certainly good to initially have the support of experts, other practitioners and academicians, in order to start in the proper way. However, much more important is the continuous commitment of the management during and after the project. The formal completion of the project ended with recommendations for improvements and plan for control phase. It is remarkable to notice that even though SMEs can struggle with several limitations and raising issues in their daily business, they are willing to adopt new methodologies, to follow the best practices and to gain quality improvements in their activities.

It is important to coordinate and gather people who are highly skilled in the subject of the Six Sigma methodology, as well as people who work in the company and are in strict contact with the operators and all workers involved in the processes under study. This is fundamental because both theory and technicalities of Six Sigma, as well as in-depth knowledge of how things happen in the daily life of the company's activities, are necessary ingredients of a good project.

Acknowledgements

The authors, who mutually acknowledge an equal contribution to the paper, wish to thank the managers and all people of the company Structo Hydraulics who participated to the project and Marcus Assarlind, of the Chalmers University of Technology for his support as academic tutor, during the project.

References

- Andersson, R., Eriksson, H. and Torstensson, H. (2006), "Similarities and differences between TQM, Six Sigma and lean", *The TQM Magazine*, Vol. 18 No. 3, pp. 282-96.

- Antony, J. (2002). Design for Six Sigma: a breakthrough business improvement strategy for achieving competitive advantage. *Work Study*, 51(1), 6-8.
- Antony, J., (2008). ‘Can Six Sigma be effectively implemented in SMEs?’ *International Journal of Productivity and Performance Management*, 57(5), pp. 420–423.
- Banuelas, R. and Antony, J. (2003), “Going from Six Sigma to design for Six Sigma: an exploratory study using analytic hierarchy process”, *The TQM Magazine*, Vol. 15 No. 5, pp. 334-44.
- Barone, S. and Lo Franco, E., (2012). *Statistical and Managerial Techniques for Six Sigma Methodology*. Wiley.
- Behara, R. S., Fontenot, G. F., & Gresham, A. (1995). Customer satisfaction measurement and analysis using Six Sigma. *International Journal of Quality & Reliability Management*, 12(3), 9-18.
- Bendell, T. (2006), “A review and comparison of Six Sigma and the lean organizations”, *The TQM Magazine*, Vol. 18 No. 3, pp. 255-62.
- Bewoor, A.K. and Pawar, M.S., (2010). ‘Mapping macro/micro level critical links for integrating Six Sigma DMAIC steps as a part of company’s existing QMS: an Indian SME case study’. *International Journal of Six Sigma and Competitive Advantage*, 6(1/2), p. 105.
- Black, K. and McGlashan, R., (2006). ‘Essential characteristics of Six Sigma Black Belt candidates: a study of US companies’. *International Journal of Six Sigma and Competitive Advantage*, 2(3), p. 301.

- Black, K. and Revere, L. (2006), “Six Sigma arises from the ashes of TQM with a twist”, *International Journal of Health Care Quality Assurance*, Vol. 19 No. 2/3, pp. 259-66.
- Chakrabarty, A. and Tan, K. (2007), “The current state of Six Sigma application in services”, *Managing Service Quality*, Vol. 17 No. 2, pp. 194-208.
- European Commission, (2003). ‘The new SME definition’. *Official Journal of the European Union*, p. 124.
- Harry, M, & Schroeder, R. *Six Sigma*, Currency, 2000.
- Husband, S. and Mandal, P., (1999). ‘A conceptual model for quality integrated management in small and medium size enterprises’. *The International Journal of Quality & Reliability Management*, 16(7), pp.699–713.
- Kumar, M., (2007). ‘Critical success factors and hurdles to Six Sigma implementation: the case of a UK manufacturing SME’. *International Journal of Six Sigma and Competitive Advantage*, 3(4), p. 333.
- Kwak, Y.H. and Anbari, F.T. (2006), “Benefits, obstacles and future of Six Sigma approach”, *Technovation*, Vol. 26, pp. 708-15.
- Linderman, K., Schroeder, R. G., Zaheer, S., & Choo, A. S. (2003). Six Sigma: a goal-theoretic perspective. *Journal of Operations Management*, 21(2), 193-203.
- Magnusson, K., Kroslid, D., Bergman, B., Häyhänen, P., Mills, D., (2003). *Six Sigma: The Pragmatic Approach*. Studentlitteratur.

- Malliga, P. and Srinivasan, S.P., (2007). ‘The stock service improvement by the deployment of Six Sigma’. *International Journal of Six Sigma and Competitive Advantage*, 3(2), p. 103.
- Rao, K.P. and Girija Rao, K., (2007). ‘Higher management education should Six Sigma be added to the curriculum?’ *International Journal of Six Sigma and Competitive Advantage*, 3(2), p. 156.
- Reddy, G.P.P. and Reddy, V.V., (2010). ‘Process improvement using Six Sigma – a case study in small-scale industry’. *International Journal of Six Sigma and Competitive Advantage*, 6(1/2), p. 1.
- Sarkar, B.N., (2007). ‘Capability enhancement of a metal casting process in a small steel foundry through Six Sigma: a case study’. *International Journal of Six Sigma and Competitive Advantage*, 3(1), p. 56.
- Snee, R.D., (2004). ‘Six-Sigma: the evolution of 100 years of business improvement methodology’. *International Journal of Six Sigma and Competitive Advantage*, 1(1), p. 4.
- Thomas, A. and Barton, R., (2006). ‘Developing an SME-based Six Sigma strategy’. *Journal of Manufacturing Technology Management*, 17(4), p. 417.

Managerial tools for Six Sigma: Prioritization of alternatives with AHP and response latency

Introduction

Six Sigma is toolbox that perfectly combines managerial and statistical tools and methods. Typical managerial tools include project management tools and techniques, decision making methods, process management diagrams, and many more.

We developed another managerial tool for the Six Sigma toolbox that can be used in decision making and voice of the customer applications. This work has been first presented at the QMOD conference 2013, then selected among the best papers of the conference and published in April 2014 in the academic journal Total Quality Management and Business Excellence (Barone, Errore and Lombardo, 2014).

Our paper introduces an integration of the response latency with the AHP procedure. The main purpose is to overcome some critical aspects of the traditional AHP. This idea is not totally new in the literature (Feinstein, 2000; Feinstein & Lumley, 2001). However we use an analytical model of response latency previously validated in other recent research projects (Barone, Lombardo, & Tarantino, 2007, 2012). Moreover the original contribution of this work in respect to the state of the art is the use of the integrated method AHP-response latency in web surveys on currently available web platforms. A case study is used as a pilot test of the proposed method. It relates to the development of a new formula for a tourism service. The investigation has been performed using the online survey platform Qualtrics®. This platform was chosen because it is one that allows – among other options – to record and use the choice times.

We include here the literature review on which the new proposed integrative approach is grounded and the new method. In the published paper we also discuss an application of the method to the development of a tourism service that provides an initial test. This new method can be profitably adopted in web surveys where it is easy to measure and record response latencies. Findings show that the respondent effort, fatigue and time consumption are drastically reduced, making the survey much simpler and faster. The obtained results seem to be very reliable in terms of judgment consistency.

There are still open issues and opportunities for future research work, as discussed in the final sections of the paper.

Tools for prioritizing alternatives in decision making

The prioritization of alternatives based on the determination of weights of relative importance represents a cross-field area of interest. This topic has been investigated in many research fields, such as marketing research (Kwong & Bai, 2003; Neslin, 1981), decision sciences (Barron & Barrett, 1996), experimental psychology (Zakay, 1985), operational research (Weber & Borcherding, 1993), management sciences (Wittink, Krishnamurthi, & Nutter, 1982).

A well-established method for such purpose is the Analytical Hierarchy Process (AHP), originally developed by Thomas Saaty (Saaty, 1986, 1990). Several applications of the AHP can be found in vary fields. It is still nowadays a very well recognized and adopted approach. For instance, in Quality Management the AHP has been adopted in combination with the Quality Function Deployment to prioritize the Voice of Customer and the technical responses (Wang, Xie & Goh, 1998; Raharjio *et al*. 2007), to prioritize the critical dimensions in Supply Chain Quality Management (Kueia, Madua & Linb, 2008), to quantify the weights for success factors for TQM in a case study involving Korean firms (Yoo, 2003) and to examine the relative importance of TQM practices in the service industry (Talib *et al*. 2011); these are just few examples. An illustrative presentation of the AHP in the context of Six Sigma is provided in Barone & Lo Franco (2012).

However, despite this method is widely recognized by the scientific community, it has always led to debate and criticism (Belton & Gear, 1983, 1997; Holder, 1990).

On the other hand, the concept of *response latency* has gained a certain attention in research and practice, especially in the last years. The general basic idea is that the time needed for a human being to make a decision (for example a choice between two alternatives) is somehow related to the degree of complexity of the decision he/she is facing. The response latency can be related to the similarity between two stimuli, or alternatives to choose from (Thurmond & Alluisi, 1963). In other words, when the time taken to make a decision is long, this means that the proposed alternatives are very similar for the respondent, so their *relative importance weights* are essentially the same, which is equivalent to say that the respondent attaches the same utility to the alternatives. Vice versa if the alternatives are very different, the favorite one will be chosen sooner.

With the use of platforms for conducting surveys, running on the internet, it is today possible to record the time taken for the decision. This approach could be non-intrusive for the respondent and costless for the researcher.

The Analytical Hierarchy Process

Determining the importance weights for the customer requirements is an essential and crucial process. The Analytical Hierarchy Process (AHP) has been used to determine the importance weights for product planning.

The AHP is a well-established method that allows to simplify complex decisional problems. The methodology can be summarized in three phases:

1. Decomposition of the problem into basic elements and hierarchy of the elements;
2. Pairwise comparisons, and construction of the pairwise comparison matrix;

3. Summary of results and prioritizing between the elements of the decision-making process.

A critical phase in the procedure is the pairwise comparisons of the basic elements. This is usually made by human subjects (hereinafter called *respondents*), who are the decision makers involved in the process. Through a well-defined procedure, the respondents are asked to undergo pairwise comparisons between the basic elements. The assignment of priorities between the elements directly derives from the procedure with rather simple mathematical calculations.

According to Saaty, who introduced the method in the late 1970s, it is natural for a respondent to compare pairs of elements, by establishing a relation of relative importance between the two. The pairwise comparison is considered as an unconscious process of the human brain when making any decision concerning more alternatives. In the AHP, respondents are asked to express the relative importance of pairs of alternatives through a semantic rating scale, whose levels are translated into numbers ranging from 1 to 9 (see **Error! Reference source not found.**).

Table 1: Rating scale for the AHP (Saaty, 1990)

Score	**Definition**
1	Equal importance between the two alternatives
3	Alternative *i* is moderately more important than the alternative *j*
5	Alternative *i* is sensibly more important than the alternative *j*
7	Alternative *i* is much more important than the alternative *j*
9	Alternative *i* is absolutely more important than the alternative *j*
2,4,6,8	To be used as half points of the scale

Each pairwise comparison provides an element of the pairwise comparisons matrix:

$$A = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} \qquad (1)$$

Where a_{ij} is the score assigned by the respondent on the comparison of the alternative i with the alternative *j*. If n is the number of elements, assuming a_{ij} = *1/a_{ji}* (reciprocity) and a_{ii} = 1, the number of judgments that the respondent is asked to make is $n(n-1)/2$.

The judgments are said to be consistent when the *transitivity property* holds. For example, in case of three alternatives, having the first alternative judged to be twice more important than the second alternative and the second alternative twice more than the last alternative, it follows that the first alternative is judged four times better than the third. According to this condition, the pairwise comparison matrix should be ideally *consistent*. Mathematically, this condition is verified if the maximum eigenvalue of the matrix *A* is equal to its dimension, $\lambda_{max} = n$.

Perfect consistency is seldom achieved in practice, mainly because of the natural limits of the human rationality. The level of inconsistency of a respondent can be measured by the consistency index, defined as:

$$\mathrm{CI} = \frac{\lambda_{\max} - n}{n-1} \qquad (2)$$

In order to make this measure meaningful in respect to a certain value, Saaty randomly generated pairwise comparison matrices for different *n* and calculated the average consistency index, calling it Random Index (RI). By making the ratio between the consistency index CI and the Random Index, a Consistency Ratio is obtained:

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}} \qquad (3)$$

According to Saaty, the upper bound for CR to consider a respondent "sufficiently" consistent is 0.1 for n-dimensional matrices, with n bigger than 5. Only for n = 3 the CR upper bound is equal to 0.05 and for n = 4 the upper bound is 0.08.

, An iterative procedure is used to calculate the final relative weights of the alternatives. This procedure requires more steps when the respondent's consistency is poor. It spreads the inconsistency over the pairwise comparisons matrix. Such procedure is here reported for completeness.

At the iteration step k:

1. Calculate

$$A_k = A_{k-1} \times A_{k-1} \qquad (4)$$

where "×" indicates the row-by-column matrix multiplication.

2. Calculate the relative importance weights of the alternatives:

$$w_i = \frac{\sum_{j=1}^{n} a_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}} \quad \forall i = 1, \dots, n \qquad (5)$$

3. Compare the obtained weights with those of the previous step and stop procedure if they are the same (with a preset tolerance).

Ishizaka & Lusti (2004) state that *consistency is not necessary, and certainly not perfectly achievable by human subjects, but it is definitely desirable for the reliability of the results*.

Limitations and criticisms on the AHP

Several authors noted some weaknesses of the AHP and proposed reviews of the method (see, e.g. Belton & Gear, 1983; Dyer, 1990; Holder, 1990). One of the main reasons of criticism versus the AHP is the use of the rating scale. Some researchers state that the pairwise comparison is an ambiguous action, especially for intangible assets, due to the difficulty to express subjective estimates on a scale of relationships (Dyer, 1990). This scale is considered restrictive because respondents may not find what actually fits their opinion among the nine points of the scale (e.g. an alternative considered *twenty* times more important than another). Therefore, the cut-off of the scale at "9" is an unnecessary limitation (Murphy 1993; Belton & Gear 1982; Belton 1986). Moreover it may be difficult for the respondent to discriminate between two adjacent scores of the scale. Considering these drawbacks, the transitive property seems to be only ideal and consequently the respondent is very seldom consistent.

Finally, it has to be highlighted that, when the respondent is asked to compare two alternatives using the rating scale, he/she has to practically answer two questions at the same time:

1. Which of these two alternatives do you prefer?

2. How much do you prefer the chosen one to the second? Or in other terms, what is your degree of preference?

On this double question are grounded the major criticisms on the method, which results to be long and tedious for the respondent, especially when the number of pairwise comparisons grows with the increase of the total number of alternatives. A large number of comparisons might be considered itself another cause of inconsistency.

Response latency

The foundations of the use of response latency can be traced back to 1890, when Joseph Jastrow showed that *every mental process takes time and that this time is increasing with the complexity of the decision to be taken* (Jastrow, 1890). Terms as "reaction time", "time to choose", “choice time” are used in the so-called "preference uncertainty theory” (Fischer *et al*., 2000). In this theory it is stated that the more a subject is uncertain about the relative value between two alternatives, the greater is the response latency. This theory has some similarity with the position of Thurmond and Alluisi, who showed that the choice time is proportional to the similarity of the perceived stimuli when evaluating alternatives (Thurmond & Alluisi, 1963). In particular, the lower the perceived difference between the two alternatives, the longer the time required to choose one of the two. According to these theories, there is a relationship between the time to choose an alternative and its importance relative to the other one.

In marketing research, the use of the response latency metrics is nowadays rather consolidated as an indicator of latent processes linked to memory and attitudes (Otter, Allenby, & Zandt, 2008). In survey research, response latency is also used as an indicator of data quality. Answers given too quickly can be seen as lacking the necessary attention, while too high response latency may indicate respondents distracted by other concurrent activities.

The increased interest on the response latency metrics contrasts, however, with a modest number of quantitative models using such data (Haaijer, Kamakura, & Wedel, 2000). One reason may be that response latency may be affected by many latent factors, so that the interpretation and the use of such data need to be careful.

Based on these concepts, Barone *et al*. (2007) proposed the formulation of a simple analytical model relating the response latency to the relative importance weights of two alternatives as a function of the response latency:

$$\frac{w_1}{w_2} = f(t_c) \qquad (6)$$

Where:

$w_1 \in [0,1]$ is the relative importance weight of the first (chosen) alternative;

$w_2 \in [0,1]$ is the relative importance weight of the second alternative;

$f(t_c)$ is a mathematical function of the response latency t_c (time to choose).

It is assumed that:

$$w_1 + w_2 = 1 \qquad (7)$$

The mathematical function $f(t_c)$ is derived from *boundary* conditions: if the response latency (ideally) tends to infinity, this means that the respondent is absolutely undecided, so the two alternatives have the same importance ($w_1 = w_2 = 0.5$).

$$\lim_{\mathrm{t_c}\to\infty} \frac{\mathrm{w_1}}{\mathrm{w_2}} = 1 \quad (8)$$

If the response latency (ideally) tends to zero, this means that the respondent considers the chosen alternative greatly more important than the second one. In this case $w_1 = 1$ and $w_2 = 0$:

$$\lim_{t_c\to 0} \frac{w_1}{w_2} = +\infty (9)$$

To make dimensionless the right hand side of the equation (6) and to take into consideration that different respondents may have different reaction times to the same stimulus, a *reference time t** is introduced in the formulation. The simplest model meeting these boundary conditions and completely dimensionless is:

$$\frac{\mathrm{w_1}}{\mathrm{w_2}} = 1 + \frac{\mathrm{t^*}}{\mathrm{t_c}} \qquad (10)$$

When it is possible to measure the times t^* and t_c for a respondent, the unknowns w_1 and w_2 can be calculated by solving a simple system of two equations:

$$\begin{cases} \frac{w_1}{w_2} = 1 + \frac{t^*}{t_c} \\ w_1 + w_2 = 1 \end{cases} \qquad (11)$$

Hence, w_1 and w_2 are given by:

$$\begin{cases} w_1 = \frac{t_c + t^*}{2t_c + t^*} \\ w_2 = \frac{t_c}{2t_c + t^*} \end{cases} \qquad (12)$$

This simple model can be implemented in more complex models aiming to incorporate the response time in the calculation of relative weights of importance and consequent prioritization.

A proposed integration: AHP plus response latency

When it is possible to record the response latency, this measure can be used to overcome the limitations of the AHP related to the use of the rating scale. Through the previously formulated model it is possible to determine the relative importance weights in the pairwise comparisons and to introduce them in the pairwise comparisons matrix:

$$A = \begin{vmatrix} 1 & {}^{w_1}/_{w_2} & \cdots & {}^{w_1}/_{w_n} \\ {}^{w_2}/_{w_1} & 1 & \cdots & {}^{w_2}/_{w_n} \\ \cdots & \cdots & \cdots & \cdots \\ {}^{w_n}/_{w_1} & {}^{w_n}/_{w_2} & \cdots & 1 \end{vmatrix} \qquad (13)$$

When the matrix has been fulfilled with the pairwise relative weights computed by the response latency model, then the AHP procedure continues as usual, so the relative importance weights for all alternatives are finally determined.

The idea is not new. Jerald Feinstein (2000) already used the response latency for the calculation of the relative weights of alternatives in a decision-making process of pairwise comparison judgments. He used the reciprocal of the response time (measured in seconds) as a measure of the degree of preference of the alternative selected. Feistein made an experiment where respondents were subjected to a sequence of pairwise comparisons with two tasks: the choice between the two options, and the expression of the degree of preference on the rating scale 1-9. The calculation of the relative importance weights of the alternatives was made by the standard procedure, using the explicit judgments of the respondent as a_{ij} or alternatively the reciprocal of the response latency. The experimental study showed that the matrices obtained with the response latency had higher consistency (on average) compared to the matrices constructed with the rating scale data.

For the experimental test of the integrated method explained above, it is necessary to have a suitable platform for the interviews, i.e. a tool for researchers to easily build a survey and to manage the results simply and quickly. A software interface was initially developed for the purpose (Barone et al., 2007). This software was conceived for face-to-face interviews which would allow more control on the respondent by the interviewer. Unfortunately face-to-face interviews imply small sample sizes.

To test the method on larger samples, an extensive search of software already available, led to identify a web-based platform that proved suitable to the purposes of the research. This is the *Qualtrics® Research Suite* platform, which has all basic functions for conducting on-line surveys and other additional features. One of these features is the timing function that allows to record the time taken by respondents during each phase or screen of the survey.

The survey can be conducted online via the circulation of a hyperlink through mailing lists or, to reach more extensive populations, social networks or blogs and forums related to the research topic can be used for the purpose.

The decision to conduct an online survey is motivated by several advantages:

- speed and ease of circulation: the survey is spread out with few clicks; possibility to reach a population of respondents without size and geographical limits;

- availability of survey results in real time: data are automatically uploaded on the server and they are continuously available to the researcher;
- low cost compared with traditional survey techniques;
- scalability of the investigation: it is possible to increase the sample size at any time;
- absence of some noise factors – e.g. the presence of interviewer – which may influence the respondent.

Obviously, one of the major drawbacks of online surveys is the fact that the researcher partially loses control on the respondent, being unable to check who is filling the survey, and to assess whether he/she is paying the necessary attention to the survey or is distracted by other concurrent activities.

Qualtrics® allows the researchers to manage the survey with several tools and features. With a registered account the researcher designs the survey, sends it to a target population of respondents, and saves automatically the data for the subsequent analysis. Several features are available and customizable to build the survey, utilizing descriptive text and pictures, single-answer questions, multiple choice questions, matrix questions, sliders and so on. Moreover a special feature, as mentioned above, named *timing,* allows recording the response time.

For each page of the survey, the server records four times (at millisecond resolution):

- *First Click time* (t_{fc}): time between the page load and the first click of the respondent on the page. It can be considered as a measure of the time taken by the respondent to read the question, reflect and make a first decision.

- *Last Click Time* (t_{lc}): time between the page load and last click of the respondent on the page before clicking the “Next” button. It is the latest time when the respondent expresses his/her final decision. In case of no mind changes, $t_{lc}= t_{fc}$.
- *Page Submit Time* (t_{ps}): time between the page load and the click on “Next” button. It represents the final decision to go ahead with the next question, so it separates the two moments of the last click (decision) and a confirmation of it.
- *Click Count:* number of respondent’s clicks on the page. Between the first and last click the respondent may be uncertain about his/her decision and sometimes is going to revise it. A measure of this uncertainty could be the *click count*.

The timing option can be added to each question that needs to be tracked. For the application of the AHP, needing several pairwise comparisons, each comparison appears on a separate page. The clock counter is not visible to respondents.

With the aim of introducing the response latency into the AHP procedure the survey must be built in such a way that, after recording some respondent’s demographics (e.g. country of origin, age, gender), each pairwise comparison is made by a question asking which one of the following two alternatives is the preferred one. Two pictures show the alternatives and the respondent can click on either one or the other to express his/her preference; then the *page submit* button leads to the next comparison.

To apply the response latency model on Qualtrics survey’s results, it is important to decide what time measurements to use, in particular for t_c and t^* in equation (3). In the application discussed in the next section the *Last Click Time* has been chosen as t_c, while the mean difference between *Page Submit Time* and *Last Click Time* is used for an estimate of the reference time t^*.

Application and open issues

A recurrent task in product and service development is the identification of an optimal *profile*, i.e. the best setting of *attributes* i.e. features of the product or service to be implemented. Attributes may assume different modalities (*attribute levels*). An attribute is considered important if the perception of the potential customer on this attribute determines a change of his/her judgment on the entire product/service profile and eventually the purchasing decision. Hence, surveys on customer preferences have a fundamental importance to anticipate which attributes have greatest impact on influencing customer opinion, and which levels of these attributes may determine the greatest satisfaction.

The so-called agri-tourism – intended as a type of accommodation for tourists in farms - is a very popular tourism format in countries like Italy. This service was chosen as field application. In particular the study concerns the initial phases of development of a new service. We skip here the details about this application (which can be found in the paper), and just remark some findings and discussions.

Beyond the specific application, there are methodological aspects which need to be highlighted. One of them is related to the consistency of the pairwise comparison matrices. The recorded times revealed how to discriminate surveys made by respondents paying the proper attention. By analyzing the pairwise comparison matrices of the final sample of respondents they show rather good values of the consistency ratios. For a high percentage of respondents in the final sample it was not necessary to “force” the consistency in the iterative procedure presented in Section 2.

The estimation of relative importance of alternatives has been widely discussed in the scientific literature and applied in many fields. Some methods are widespread and consolidated, although there are criticisms and limitations. Among them, the AHP has always stimulated scientific debate. Innovative models and more efficient and effective methods have been developed and proposed over the years and often already undergone several validations. However, these methods often suffer of an inadequate circulation in the academic community and industrial applications. Among those, the methods using response latency, as the one discussed in this paper.

With methodological arguments supported by the results of a pilot investigation, a response latency model is proposed in this paper with the aim of improving the AHP procedure. The purpose was to highlight a possibility of integration that allows lightening the burden of the AHP procedure for the respondent with the use of response latency in the elicitation of pairwise comparisons. With response latency, the step in which the respondents should express judgments on a scale of preference is eliminated. The procedure is simplified and the overall completion time of the survey is strongly reduced, with a beneficial effect on the respondent consistency. Moreover thanks to the availability of response times, a data pre-filtering is also possible to eliminate cases of surely unreliable respondents.

Further research is needed to investigate the appropriate choice of the recorded times to be used in the mathematical model of response latency, formulated in this paper and to deeper refine and validate the method.

Acknowledgments

The data in this work were collected thanks to the online survey platform Qualtrics®. The authors, who mutually acknowledge an equal contribution to this work, had a temporary full license, which allowed recording response latency measures. For this reason, the authors express gratitude to the company Qualtrics and especially to Mr. Jason Harris.

References

- Barone, S., and Lo Franco, E. (2012). *Statistical and managerial techniques for Six Sigma methodology*. Wiley.
- Barone, S., Lombardo, A., and Tarantino, P. (2007). A weighted logistic regression for Conjoint Analysis and Kansei Engineering. *Quality and Reliability Engineering International*, 23(6), 689-706.
- Barone, S., Lombardo, A., and Tarantino, P. (2012). A heuristic method for estimating attribute importance by measuring choice time in a ranking task. *Risk and Decision Analysis*, 3(4), 225-237.
- Barron, F.H., and Barrett, B.E. (1996). Decision quality attribute using weights ranked. *Management Science*, 42(11), 1515-1523.
- Belton, V., and Gear, T. (1983). On a shortcoming of Saaty's Analytic Hierarchy Process. *Omega*, 11(3), 228-230.
- Belton, V., and Gear, T. (1997). Discussion on the meaning of relative importance. *Journal of Multi-criteria Decision Analysis*, 6, 335-338.
- Dyer, J.S. (1990). Remarks on the Analytic Hierarchy Process. *Management Science*, 36 (3), 249-258.

- Feinstein, J.L. (2000). Comparing response latency and self-report methods for estimating levels of certainty in knowledge elicitation for rule-based expert systems. *Expert Systems*, 17(5), 217-225.
- Feinstein, J.L., and Lumley, R. (2001). Theoretical and practical aspects of AHP using a scale derived from the time it takes to decide between two choices instead of one derived from “1-9” estimates. *Proceedings 6th ISAHP 2001 Berne, Switzerland* (pp. 93-100).
- Fischer, G.W., Luce, M.F., Jia, J., Frances, M., Jianmin, L., and Carolina, N. (2000). Attribute conflict and preference uncertainty: effects on judgment time and error. *Management Science*, 46(1), 88-103.
- Haaijer, R., Kamakura, W., and Wedel, M. (2000). Response latencies in the analysis of conjoint choice experiments. *Journal of Marketing Research*, 37(3), 376-382.
- Holder, R. D. (1990). Some comments on the Analytic Hierarchy Process. *Journal of Operational Research Society,* 41(11), 1073-1076.
- Ishizaka, A., and Lusti, M. (2004). An expert module to improve the consistency of AHP matrices. *International Transactions in Operational Research*, 11, 97-105.
- Jastrow, J. (1890). Time-relation of mental phenomena. *Science,* 16(397), 142-150.
- Kueia C.H., Madua C.N., and Linb C. (2008). Implementing supply chain quality management. *Total Quality Management*, 19(11), 1127–1141.

- Kwong, C.K., and Bai, H. (2003). Determining the importance weights for the customer requirements in QFD using a fuzzy AHP with an extent analysis approach. *IIE Transactions*, 35, 619-626.
- Neslin, S.A. (1981). Linking product features to perceptions: self-stated versus statistically revealed importance weights. *Journal of Marketing Research*, 18(1), 80-86.
- Otter, T., Allenby, G.M., and Zandt, T.V.A.N. (2008). An integrated model of discrete choice and response time. *Journal of Marketing Research,* 45 (October), 593-607.
- Raharjo H., Xie M., Goh T.N., and Brombacher A.C. (2007). A methodology to improve Higher Education quality using the Quality Function Deployment and Analytic Hierarchy Process. *Total Quality Management*, 18(10), 1097–1115.
- Saaty, T.L. (1986). Axiomatic foundation of the Analytic Hierarchy Process. *Management Science*, 32 (7), 841-855.
- Saaty, T.L. (1990). How to make a decision: The Analytic Hierarchy Process. *European Journal of Operational Research*, 48(1), 9-26.
- Talib F., Rahman Z. and Qureshi M.N. (2011). Prioritizing the practices of Total Quality Management: an Analytic Hierarchy Process analysis for the service industries. *Total Quality Management & Business Excellence*, 22:12, 1331-1351.

- Thurmond, J.B., and Alluisi, E. (1963). Choice time as a function of stimulus dissimilarity and discriminability. *Canadian Journal of Psychology*, 17 (3), 326-37.
- Wang, H., Xie M., and Goh T.N. (1998). A comparative study of the prioritization matrix method and the Analytic Hierarchy Process technique in Quality Function Deployment, *Total Quality Management*, 9(6), pp. 421–430.
- Weber, M., and Borcherding, K., (1993) Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research*, 67 (1),1-12
- Wittink, D.R., Krishnamurthi, L., and Nutter, J.B. (1982). Comparing derived importance weights across attributes. *Journal of Consumer Research*, 8(4), 471-474.
- Yoo H. (2003). A study on the efficiency evaluation of total quality management activities in Korean companies. *Total Quality Management*, 14(1), 119–128.
- Zakay, D. (1985). Post-decisional confidence and conflict experienced in a choice process. *Acta Psychologica*, 58, 75-80.

Statistical tools for Six Sigma: experimental designs for screening

Introduction

The very basic principle behind experimental investigation is the acknowledgement that correlation does not imply causation. As opposed to observational studies, experimental studies actively manipulate independent variables that are under the experimenter's control, in order to investigate their effect on the dependent variable of interest.

As a mean of active manipulation, design of experiment (DOE) is the most powerful tool in the Six Sigma toolbox. Moreover, this is a tool that a Black Belt can leverage on all the phases of the define-measure-analyze-improve-control (DMAIC process).

DEFINE - When defining the progress metrics in the define phase, a fundamental role is played by the Voice of the Customer (internal or external). Customer focus must be the leading principle of any Six Sigma project. If the measure isn’t on the customer’s radar screen, it probably shouldn’t be on yours (Jones and Nachtsheim, 2003).Typical applications of DOE in the define phase are designed customer surveys.

MEASURE - In the measure phase, DOE can be used to determine process capability, compare alternative measurement methods and establish the validity of selected metrics.

ANALYZE – in investigating root causes of process variation

IMPROVE - In this phase the project team identifies the specific changes that potentially can yield the desired improvements and reduce the major sources of variation. The key process variables are identified through statistically designed experiments; Black Belts then use these data to establish what ‘knobs’ must be adjusted to improve the process (Harry and Schroeder, 2000).

CONTROL - Identified improvement actions need to be standardized, communicated and implemented throughout the organization. DOE is used in this step in a variety of ways, for instance performing sensitivity analysis, scale-ups, and fine-tuning studies.

The earlier DOE is actively used in the production cycle, the more effective can be the cost savings. Design for Six Sigma (DFSS), is the integration of DOE and Six Sigma from the outset. DOE becomes the analytical tool of choice for new product and new process development. The potential savings from DFSS may far exceed the gains in simply optimize a product or process developed in traditional way.

The applications of DOE are not unique to Six Sigma. In the purpose of this doctoral thesis, experimental designs are framed as a powerful tool for Six Sigma projects, and they can be proved to be essential tools in every one and each of the DMAIC phases. However, it's important to highlight that the value of research in this area goes far beyond the edges of the Six Sigma methodology.

Since the dawn of experimental design, in the early 20th century, this research area has made contributions to virtually every area of science and technology. It is widely recognized that designed experiments require more effort than observational studies, and critical success factors of such studies go beyond the mere collection and analysis of data. They also require some up-front investment in time and resources.

In the following paragraphs we explore some issues related to a peculiar case of experimental studies, screening experiments. This work comprises what is included in two papers currently under the review process in top international journals. The content reported here is mainly taken from early versions of the papers, respectively.

Screening Designs for Two-Level Factors

Jones and Nachtsheim (2011b) introduced a new class of three-level designs called Definitive Screening Designs (DSDs). These designs have a number of appealing statistical properties. For example, they provide estimates of main effects that are unbiased by any second-order effects; they require only one more than twice as many runs as there are factors; and they avoid confounding of any pair of second order effects. These authors later showed how two-level categorical factors could be added to existing DSDs (Jones and

Nachtsheim, 2013). Interestingly, the authors never considered the case where all factors are categorical and have only two-levels. This paper is directed toward creating and exploring a class of two-level designs that retain many of the advantages of DSDs. We employ an algorithm for the construction of designs for varying run sizes and numbers of factors, and we characterize the statistical properties of these designs.

Screening experiments are often useful tools during the early phases of an empirical investigation, when an experimenter has little prior knowledge about the product or process under study. At this stage, subject matter experts may have identified a relatively large set of potential factors that may have cause-and-effect relationships with the response (or responses) of interest. Screening experiments are usually small, two-level, main effects designs that aim to identify the much smaller set of active factors. Follow-up experiments employing only the active factors are then employed to determine the nature of interactions or other non-linearities in a sequential process.

Definitive screening designs (DSDs), as introduced by Jones and Nachtsheim (2011b), provide a new approach for screening experiments at three levels. DSDs provide estimates of main effects that are unbiased by any second-order effect, require only one more than twice as many runs as the number of factors, m, and avoid confounding of any pair of second-order effects. The use of three levels for each factor allows the investigator to identify quadratic effects. Also, for designs having six factors or more, these designs project to efficient response surface designs with three or fewer factors. Jones and Nachtsheim used numerical methods to construct DSDs and found that these designs were orthogonal for 6, 8, and 10 factors. In a key advance, Xiao et al. (2012), showed how to use conference matrices to construct orthogonal DSDs for most even numbers of factors.

One limitation of the original DSDs is that all factors must be quantitative. Recently Jones and Nachtsheim (2013) showed how to augment a three-level DSD with two-level categorical factors. They accomplish the construction of these mixed-level designs with two alternative methods: the DSD-augment method retains the fold-over structure of the DSD so as to provide highly efficient designs with all main effects unbiased by any active second-order effects. The ORTH-augment method leads instead to designs that are orthogonal linear main effects plans, even though some partial aliasing between main effects and interactions involving the categorical factors is present.

Interestingly, the construction of DSDs in which all factors have two-levels has not been considered. Technically, it's not possible to retain all the properties of the three-level DSDs when dealing with only two-level factors. For this reason it is not proper to call this expansion two-level DSDs; however our investigation on two-level design follow the same spirit, yet with its limitations, of the three-level case.

Of course, there is a vast literature on screening designs for two-level factors, and standard practice is to employ resolution III or resolution IV fractional factorial designs, or non-regular designs such as Plackett-Burman designs. Margolin (1969) considered the use of full fold-overs of small non-orthogonal designs that have desirable aliasing properties for screening purposes, because of their desirable aliasing structure. As will be seen in Section 2, Margolin's work is very closely aligned with our objectives here.

The remainder of the paragraph is organized as follows. We provide an overview of the related literature. Then we give a simple construction method that leads to two-level factors designs whose estimated main effects are completely independent of second-order effects. We use the algorithm of Section 3 to characterize a class of two-level screening designs. We then introduce a compound optimization procedure which leads to refinements for some of our designs. Conclusions and discussion are contained in the final section.

Literature on screening experiments

Standard approaches to factor screening, such as the use of 2^{k-p} resolution III fractional factorial designs and/or Plackett-Burman designs, all suffer from full or partial confounding of main effects and two-factor interactions. Resolution III 2^{k-p} fractional factorial directly confound main effects and two-factor interactions. These undesirable features can be avoided by moving to a design of a higher resolution, such as a resolution IV or a resolution V fractional factorial design. In a resolution IV fractional factorial design, all main effects are not confounded with any other main effects or two-factor interactions; however, some two-factor interactions are confounded with each other. In a resolution V fractional factorial design, neither main effects nor two-factor interactions are confounded with any other main effects or two-factor interactions. The preferable aliasing structure associated with resolution IV and V designs comes at the cost of significant increases to the sample size.

Margolin (1969) was perhaps the first researcher to consider non-orthogonal designs that have desirable aliasing structures as a useful alternative to orthogonal designs in screening experiments. Margolin sought designs that were highly efficient for estimation of main effects with the property that all main effects are completely independent of two-factor interactions. This condition leads to estimates of main effects that are unbiased by any potential two-factor interactions. For certain number of factors m he suggested the use of the so called “$2^m // 2m$” designs, which are $2m$-run designs obtained from fold-overs of *m*-run weighing designs (more details on weighing designs can be found in Hotelling, 1944). This approach is unique, because these designs are generally not completely orthogonal for main effects. Perhaps for this reason, the Margolin designs have been largely ignored until recently, when Miller and Sitter (2005) studied the robustness and other properties of these designs and advocated their use in screening applications. Although Jones and Nachtsheim (2011b) did not reference Margolin's $2^m//2m$ designs, DSDs are closely related. Both are based on fold-overs of m-run designs and, for this reason, both approaches produce designs that provide estimates of main effects that are unbiased by two-factor interactions. DSDs, of course, are based on three-levels and hence provide estimates of curvatures. They are also frequently orthogonal for main effects.

Since the publication of Jones and Nachtsheim (2011b), some extensions to DSDs have begun to appear. Xiao et al. (2012) showed how conference matrices can be used to construct three-level DSDs for most values of even m. Building on the work of Xiao et al., Jones and Nachtsheim (2013) showed how it is possible to construct DSDs with added two-level categorical factors. In another extension, Jones and Nachtsheim (2013) developed methods for constructing orthogonally-blocked DSDs.

The purpose of our investigation is to identify and characterize a new class of two-level DSDs. The intention here is to expand upon the three-level DSDs and mixed two and three-level DSDs of Jones and Nachtsheim, as well as the $2^m//2m$ designs proposed by Margolin. Margolin restricted his class of designs to fold-overs of weighing designs that require run size to be 2*m*. Our approach does not have such a requirement. Using numerical optimization techniques, we identify optimal two-level fold-over designs for sample sizes equal to 2*m* and larger.

Methodology

Consider the linear main effect model (*M E*)

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i \quad i = 1, \ldots, n \qquad (1)$$

where m is the number of factors, the parameters $\beta_0, \ldots, \beta_m$ are unknown constants (of which many are zero by the sparsity of effects assumption), and the $\{\varepsilon_i\}$ are *iid* $\mathrm{N}(0, \sigma^2)$. Similarly, the main effects plus bi-linear interactions model (*M E* + *I*) is:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \beta_{jk} x_{ij} x_{ik} + \varepsilon_i \qquad i = 1, \ldots, n \quad (2)$$

Let ***d*** denote the $n \times m$ design matrix

$$\boldsymbol{d} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

where x_{ij} = ±1, and let $\boldsymbol{x}_i'$ = $(x_{i1}, \ldots, x_{im})$ denote the i-th row of **d**. Let

$$\mathbf{f}_1'(\mathbf{x}_i, \mathbf{d}) = (1, x_{i1}, \ldots, x_{im}) \qquad (3)$$

$$\mathbf{f}_2'(\mathbf{x}_i, \mathbf{d}) = \left(x_{i1}x_{i2}, x_{i1}x_{i3}, \ldots, x_{i(m-1)}x_{im}\right) \qquad (4)$$

Then the *M E* model is:

$$y_i = \mathbf{f}_1'(\mathbf{x}_i, \mathbf{d})\beta_1 + \varepsilon_{1\,,}$$

and the *M E* + *I* model is:

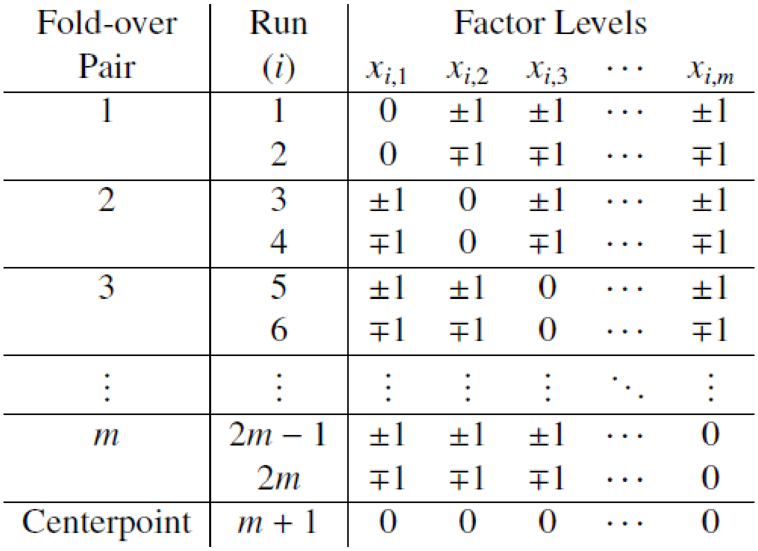
$$y_i = \mathbf{f}_1'(\mathbf{x}_i, \mathbf{d})\beta_1 + \mathbf{f}_2'(\mathbf{x}_i, \mathbf{d})\beta_2 + \varepsilon_{1\,,}$$

Let

$$\mathbf{X}_k = \begin{vmatrix} \mathbf{f}'_k(\mathbf{x}_1, \mathbf{d}) \\ \vdots \\ \mathbf{f}'_k(\mathbf{x}_n, \mathbf{d}) \end{vmatrix} \quad k = 1,2; \qquad \mathbf{Y} = \begin{vmatrix} y_1 \\ \vdots \\ y_n \end{vmatrix}, \qquad \varepsilon = \begin{vmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{vmatrix}.$$

In matrix form, the *M E* model is $\mathbf{Y} = \mathbf{X_1}\boldsymbol{\beta_1} + \boldsymbol{\varepsilon}$, where $\mathbf{X}_1$ is the $n \times (m + 1)$ model matrix for the intercept term and the linear main effects in $\boldsymbol{\beta}_1$. Similarly, the *M E* + *I* model in matrix form is $\mathbf{Y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}$, where $\mathbf{X}_2$ is the n x t model matrix for the interactions terms, $t = \binom{m}{2}$, and $\boldsymbol{\beta_2}$ is the vector of interaction effects.

Table 2: Design structure for three-level *m*-factor DSD



Fold-over Pair	Run (i)	Factor Levels				
		$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$\cdots$	$x_{i,m}$
1	1	0	± 1	± 1	$\cdots$	± 1
	2	0	∓ 1	∓ 1	$\cdots$	∓ 1
2	3	± 1	0	± 1	$\cdots$	± 1
	4	∓ 1	0	∓ 1	$\cdots$	∓ 1
3	5	± 1	± 1	0	$\cdots$	± 1
	6	∓ 1	∓ 1	0	$\cdots$	∓ 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	$2m - 1$	± 1	± 1	± 1	$\cdots$	0
	$2m$	∓ 1	∓ 1	∓ 1	$\cdots$	0
Centerpoint	$m + 1$	0	0	0	$\cdots$	0

If we assume that the reduced model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ is used for ordinary least squares estimation, it is well known that $\mathrm{E}\{\widehat{\boldsymbol{\beta}}_1\} = \boldsymbol{\beta_1} + \mathbf{A}\boldsymbol{\beta_2}$, where $\mathbf{A} = (\mathbf{X}'_1\mathbf{X}_1)^{-\mathbf{1}}\mathbf{X}'_1\mathbf{X}_2$ is the alias matrix. If either $\mathbf{A} = \mathbf{0}$ or $\boldsymbol{\beta_2} = \mathbf{0}$, then the estimate of $\boldsymbol{\beta_1}$ is unbiased.

All DSDs proposed in the literature employ fold-over pairs in order to obtain alias matrices having most or all elements equal to zero. For example, three-level DSDs impose the structure shown in Table 2. To construct the two-level DSDs proposed in this article, we start with a two-level design $\mathbf{X}$, and then impose a similar fold-over structure. The design matrix, $\mathbf{X}_1$, is given by

$$\mathbf{X}_1 = \begin{vmatrix} \mathbf{1} & \mathbf{X} \\ \mathbf{1} & -\mathbf{X} \end{vmatrix}. \qquad (5)$$

This structure guarantees that all main effects are orthogonal to any two-factor interactions. To see this we first note that X2 has the following structure:

$$\mathbf{X}_2 = \begin{vmatrix} \mathbf{X}_I \\ \mathbf{X}_I \end{vmatrix}, \qquad (6)$$

where $\mathbf{X}_I$ represents the *t* interaction columns obtained from the cross-products of main effects columns in $\mathbf{X}$. Because of the fold-over structure of $\mathbf{X}_1$, the lower submatrix of $\mathbf{X}_2$, obtained from cross-products of columns in **-X**, also results in $\mathbf{X}_I$.

Consequently, the alias matrix:

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \begin{vmatrix} \mathbf{1}'\mathbf{X}_I \\ \mathbf{0} \end{vmatrix}. \qquad (7)$$

In summary, the construction method of (5) guarantees that no confounding will exist between main effects and two-factor interactions. Some confounding between the intercept term and two-factor interactions may be present for a general value of *n = 2k.* However, when n is a multiple of 4 (i.e., *k* is even), designs may exists such that **A** = 0.

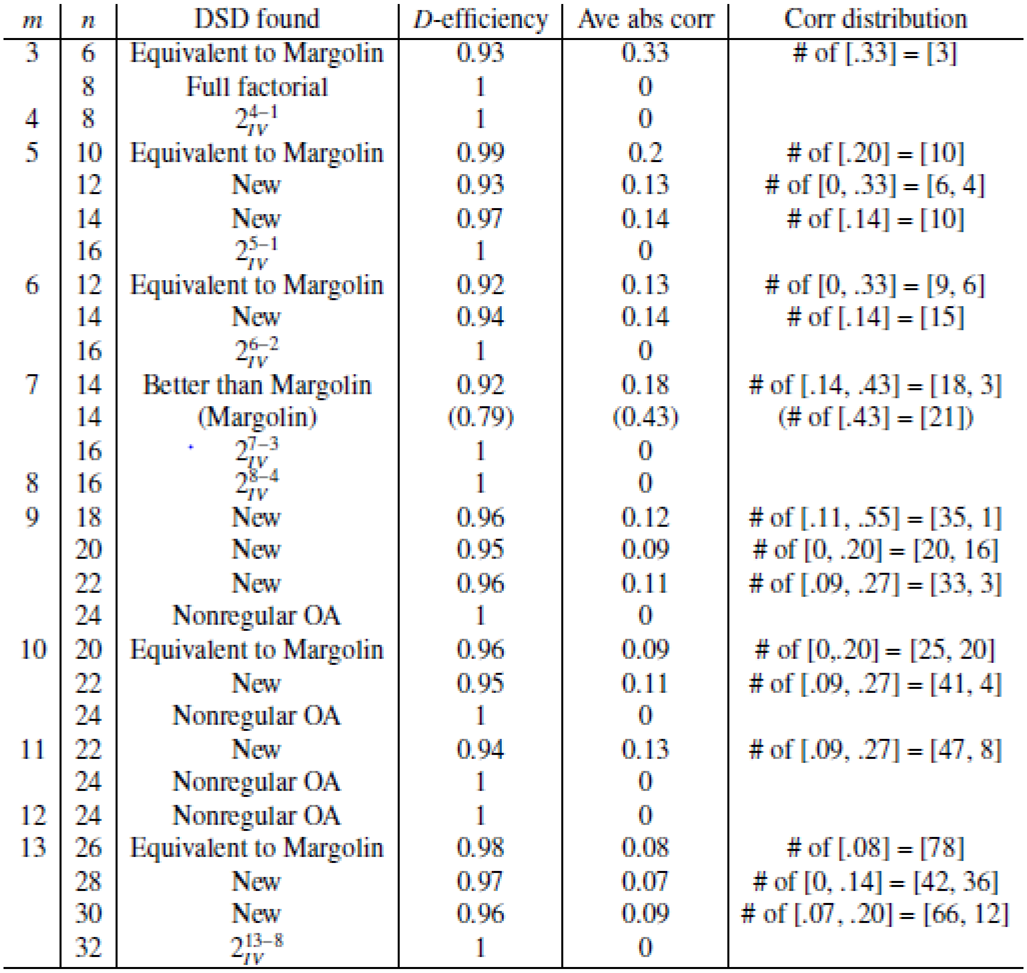
In general, a design is said to be an m-factor two-level DSD if the design maximizes$|\mathbf{X}_1'\mathbf{X}_1|$, where $\mathbf{X}_1$ is constrained to follow the structure implied by (5). Following Jones and Nachtsheim (2011b), we use the coordinate exchange algorithm (Meyer and Nachtsheim, 1995) to search for the optimal design of $\mathbf{X}$ such that $|\mathbf{X}_1'\mathbf{X}_1|$ is maximized. The resulting DSDs may or may not be orthogonal, and the first row of the alias matrix may or may not be zero. We explore the characteristics of the class of two-level DSDs in the next section.

We note that for certain values of n, the heuristic procedure based on modifying conference matrices given by Jones and Nachtsheim (2013) can be employed. Jones and Nachtsheim developed this procedure for obtaining mixed-level DSDs with $m_3 > 0$ three-level factors and $m_2 > 0$ two-level categorical factors. The procedure also works, without modification, when all factors are two-level categorical factors (i.e., $m = m_2$ and $m_3 = 0$). We compared the results from use of the Jones and Nachtsheim (2013) heuristic to optimization of the determinant via coordinate exchange, as described above, and found that the results were roughly comparable. Because of its generality, in what follows, we report results from use of the coordinate exchange algorithm.

Results

We construct two-level DSDs for $3 \leq m \leq 13$ and for selected even values of $n \geq 2m$. The results are summarized in Table 3. We compare our designs with alternative designs, if existing, in the table. We also show the D-efficiency of the obtained design, the average absolute correlation between main effects, and the correlation distributions in the last three columns of the table, respectively. In the following sections we identify some notable features of the proposed designs.

Table 3: DSD Table



m	n	DSD found	D-efficiency	Ave abs corr	Corr distribution
3	6	Equivalent to Margolin	0.93	0.33	# of [.33] = [3]
	8	Full factorial	1	0	
4	8	2^{4-1}_{IV}	1	0	
5	10	Equivalent to Margolin	0.99	0.2	# of [.20] = [10]
	12	New	0.93	0.13	# of [0, .33] = [6, 4]
	14	New	0.97	0.14	# of [.14] = [10]
	16	2^{5-1}_{IV}	1	0	
6	12	Equivalent to Margolin	0.92	0.13	# of [0, .33] = [9, 6]
	14	New	0.94	0.14	# of [.14] = [15]
	16	2^{6-2}_{IV}	1	0	
7	14	Better than Margolin	0.92	0.18	# of [.14, .43] = [18, 3]
	14	(Margolin)	(0.79)	(0.43)	(# of [.43] = [21])
	16	2^{7-3}_{IV}	1	0	
8	16	2^{8-4}_{IV}	1	0	
9	18	New	0.96	0.12	# of [.11, .55] = [35, 1]
	20	New	0.95	0.09	# of [0, .20] = [20, 16]
	22	New	0.96	0.11	# of [.09, .27] = [33, 3]
	24	Nonregular OA	1	0	
10	20	Equivalent to Margolin	0.96	0.09	# of [0,.20] = [25, 20]
	22	New	0.95	0.11	# of [.09, .27] = [41, 4]
	24	Nonregular OA	1	0	
11	22	New	0.94	0.13	# of [.09, .27] = [47, 8]
	24	Nonregular OA	1	0	
12	24	Nonregular OA	1	0	
13	26	Equivalent to Margolin	0.98	0.08	# of [.08] = [78]
	28	New	0.97	0.07	# of [0, .14] = [42, 36]
	30	New	0.96	0.09	# of [.07, .20] = [66, 12]
	32	2^{13-8}_{IV}	1	0	

Orthogonal 2-level DSDs

One attractive feature of the proposed 2-level DSDs is that they can be orthogonal when n is a multiple of four. There are several such designs shown in Table 3, where n = 8, 16, 20, 24, and 32. For n = 8, our algorithm found the full factorial 23 design for m = 3 and the maximum resolution fractional factorial 2^{4-1}_{IV} for m = 4.

For n = 16, Sun, Li, and Ye (2008) obtained and cataloged all non-isomorphic orthogonal designs for all values of m. An interesting research question arises here. For a given m, how many 16 × m designs in the catalog satisfy the requirement of (7)? That is, what proportion of orthogonal designs have main effects are orthogonal to all two-factor interactions? A careful check of the catalog of Sun et al. (2008) reveals that most of the 16-run designs have a resolution 3 (for regular designs) or 3.5 (for non-regular designs), which do not satisfy the requirement of 2-level DSDs to have alias structure (7). For the cases of *m* = 6; 7, and 8 considered in Table 3, there is only one resolution-IV design that is DSD-eligible. Our algorithm found the corresponding designs in all cases. For *m* = 5, there are two designs meeting the requirement of (7): a resolution-V design and a resolution-IV design. It can be easily proven that the resolution-V design does not have the fold-over structure of (5). Thus, it is not surprising that our algorithm did not produce the resolution-V design.

For (*n; m*) = (32; 13), the DSD is the same as the minimum aberration 32-run design given in Wu and Hamada (2008).

The cases for *n* = 20 and 24 are also interesting. For *m* = 9 and *n* = 20, we did not find an orthogonal design. This should come as no surprise, as Sun et al. (2008) showed that all 20-run orthogonal designs have a resolution of either 3.8 (for $m \leq 10$) or 3.4 (for $11 \leq m \leq 19$). However, the 20 × 9 DSD obtained does not sacrifice much in the way of orthogonality. It has a D-efficiency of .96 and the average absolute correlation between columns of .09. The last column of Table 3 shows that, among the 36 pairs of columns, 20 have a correlation of 0, and 16 have a correlation of .20.

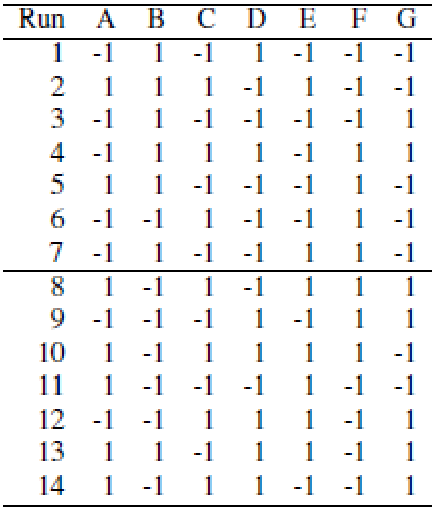
For *n* = 24, and *m* = 9, 10, 11, and 12, the DSDs found are all fold-overs of the corresponding 12-run Plackett-Burman designs with m factors, respectively. These results were consistent with those in Miller and Sitter (2001), who advocated the use of fold-overs of 12-run Plackett-Burman designs.

DSDs vs. Margolin’s Designs

Margolin’s approach requires that $n = 2m$. In all such cases considered in Table 5, our approach found the designs that are equivalent to or better than Margolin’s designs. We call two designs equivalent if they have the same D-efficiency and correlation distributions. For $m = 3; 5; 6; 10$ and 13, the DSDs constructed are equivalent to those of Margolin’s. For instance, when $m = 10$ and $n = 20$, the DSD has $D = .96$. Among the 45 pairs of columns, most (41 out of 45) have a correlation of only .09, and the remaining four pairs have a correlation of .27. Overall, the lack of orthogonality appears to be moderate.

There is one case in which we obtained better results than Margolin’s approach. Consider the case having $m = 7$ factors and $n = 14$ runs. Our design, as shown in Table 4, has a *D*-efficiency of 0.91 and an average absolute correlation of 0.18. This represents substantial improvement over Margolin’s 2^7 / / 14 design, which has $D = .79$ and an average absolute correlation of 0.43. Figure 1 provides a correlation cell plot (for correlations among main effects and two-factor interaction columns) for the 14 × 7 DSD.

Table 4: Two-level DSD for *m* = 7 and *n* = 14



Run	A	B	C	D	E	F	G
1	-1	1	-1	1	-1	-1	-1
2	1	1	1	-1	1	-1	-1
3	-1	1	-1	-1	-1	-1	1
4	-1	1	1	1	-1	1	1
5	1	1	-1	-1	-1	1	-1
6	-1	-1	1	-1	-1	1	-1
7	-1	1	-1	-1	1	1	-1
8	1	-1	1	-1	1	1	1
9	-1	-1	-1	1	-1	1	1
10	1	-1	1	1	1	1	-1
11	1	-1	-1	-1	1	-1	-1
12	-1	-1	1	1	1	-1	1
13	1	1	-1	1	1	-1	1
14	1	-1	1	1	-1	-1	1

We can compare this design with some alternative designs for studying seven factors, in which both main effects and 2-factor interactions are of interest. A resolution III fractional factorial design would require only eight runs, but main effects would be completely confounded with two-factor interactions. Similarly, a non-regular Plackett-Burman design would require 12 runs, but all main effects are partially with some two-factor interactions.

A 16-run design would have better confounding structure, as shown in Table 3, but two extra runs would be needed. Thus, the 14 × 7 DSD or Table 4, like other DSDs proposed here, appears to have a good balance of the trade-offs between run size, orthogonality, and confounding structure.

New DSDs

Ten of the 27 DSDs summarized in Table 3 are identified as “New”, in the sense that there are no existing designs constructed satisfying the requirement of (7). Consider, for instance, the case for *m* = 9 and *n* = 18, for which no existing design is available. (Note that Margolin's approach is not applicable to all cases for *n* = 2*m*.) The DSD obtained has *D* = .96. The average absolute correlation is .12. Again, the loss of orthogonality is moderate. In exchange, it has completely eliminated any aliasing between main effects and two-factor interactions.

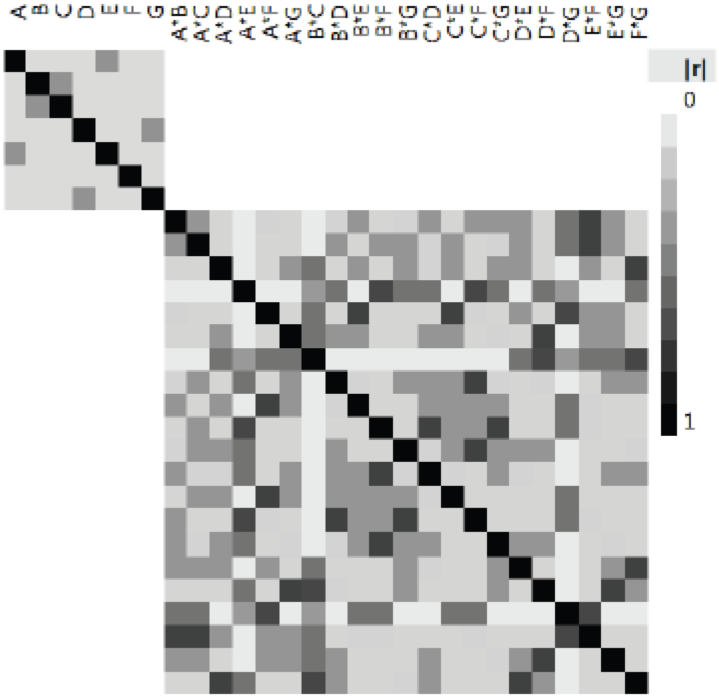


Figure 1: Color map of correlations for a two-level DSD with *m* = 7 and *n* = 14

Eliminating fully aliased two-factor interactions

In the previous sections we considered the optimization of the $|\mathbf{X}_1'\mathbf{X}_1|$ for the main effects model subject to the constraint that the design is a DSD. As illustrated in Table 3, this approach results in a number of designs that perform quite well with respect to the *D*-efficiency of the main effects model. In addition to estimating the main effects, an aim of investigators using these designs might be to estimate a small number of non-negligible two-factor interactions without ambiguity. To accomplish this goal it is necessary that no two columns of $\mathbf{X}_2$ be identical. Such pairs of interactions are completely confounded, which means that their effects cannot be separated using any data driven methodology. One potential limitation of the class of DSDs summarized in Table 3 is that a few of them had some two-factor interactions fully aliased with other two-factor interactions.

In this section, we develop a methodology for constructing DSDs that avoid full aliasing of pairs of second-order interactions. To break the aliasing between two-factor interaction pairs, when it exists, we use a multiple objective (or compound) function optimization as advocated by Jones and Nachtsheim (2011a) and as implemented in the procedure described by Jones (2013). The procedure involves creating designs that maximize a weighted average of two criteria. The primary criterion is approximate *D*-efficiency:

$$C_1 = \frac{|\mathbf{X}_1'\mathbf{X}_1|^{\frac{1}{m+1}}}{n}$$

The secondary criterion, *C*2, seeks to find small values of the off diagonal elements of the $t \times t$ covariance matrix **M =** $\mathbf{X}_2'\mathbf{X}_2$, where $\mathbf{X}_2$ is defined in (6).

Let $\{c_k\}$, for k *1, … ,* g denote the set of $g = t(t-1)/2$ elements of **M** that lie above the diagonal. The secondary (maximization) criterion is given by the inverse of the *L*r norm for *r* ≥ 1:

$$C_2 = \left[\sum_{k=1}^{g} |c_k|^r\right]^{\frac{-1}{r}} \qquad (8)$$

We discuss the choice of r below. For $r \geq$, this secondary minimization criterion penalizes designs having pairs of columns in $\mathbf{X}_2$ with large covariances. The overall objective function is to maximize, for a specified weight *w*, *(*0 ≤*w*≤ 1*)*, a weighted average of the above two criteria:

$$C_w = \text{w} * C_1^s + (1 - w) * C_2^s \qquad (9)$$

where C_1^s and C_2^s are scaled values of C_1 and C_2 . To find the scaling, following Jones (2013), we first maximize C_1 . Let C_1^{max} be the resulting *D*-efficiency. We then optimize C_2 yielding C_2^{max}. Subsequently, for any design, the scaled primary and secondary criteria are given by:

$$C_i^s = \frac{C_i}{C_i^{max}}$$

For i = 1; 2. To produce our compromise designs, we repeatedly choose w at random from (0,1) and find the design that optimizes C_w. We then find the Pareto front of non-dominated designs and choose a final design by requiring that the design avoid any confounding between pairs of two-factor interaction columns while maintaining a relative *D*-efficiency of greater than 95% with respect to the D-optimal DSD.

For secondary criterion (8), we initially chose r = 1. However, with this choice we did not consistently eliminate the full aliasing among all two-factor interaction pairs. We observed that minimizing the average absolute covariance was not sufficient, and that minimization of the maximum absolute covariance was of greater relevance to the goal of eliminating confounded pairs. Since the mini-max criterion results for $r \rightarrow \infty$, we simply experimented with larger values of *r*, finding that choice of r = 4 was sufficient.

Of the two-level DSDs produced in Table 3, eight had two-factor interaction pairs fully aliased. We applied the compound optimization algorithm in each of these cases. The results are summarized in Table 5. Note that in every case, aliasing between pairs of two-factor interactions has been eliminated. Decoupling these interactions has come at a very slight cost in terms of D-efficiency. The drop in D-efficiency for the compound designs ranges from 2% to 4%. For comparison, we include in Table 5 the average and maximum of the absolute correlations between pairs of two-factor interactions. It is interesting to note that the average absolute correlations for the compound DSDs have increased slightly, while the maximum is reduced in every case. This is typical of the mini-max criterion, where minimizing the maximum of a function usually comes at the expense of an increase in the average.

We illustrate the use of our compound optimization procedure by applying to the case with m = 9 factors and n = 22 runs. To do so, we provide a plot of the two criteria for designs on the Pareto frontier of non-dominated designs in Figure 2. The point plotted at the upper left corresponds to the D-optimal DSD while the point in the lower right corresponds to the design having the maximum value of the secondary criterion C_2. Note that of 10; 000 designs generated with the previously described random weights, w, we found only four designs on the Pareto front. The design we prefer corresponds to the point at the top right. Its D-efficiency is 0.995 while its efficiency with respect to the secondary criterion is 0:99. For this design, all the factor columns have a pairwise correlation of 1/11th with other columns. The maximum correlation among pairs of two-factor interaction columns is 7/15th (0.47). There are 126 such pairs. There are 252 pairs of two-factor interaction columns having correlations of 4/15th (0.27) and a final 252 pairs of two-factor interaction columns having correlations of 1/10th.

By contrast, the D-optimal DSD confounds three pairs of two-factor interactions. It also has 18 pairs of interactions with correlations greater than a half. On the other hand its main effects have standard errors that are around 5% smaller. In addition, the average correlation of two-factor interaction pairs is also about 5% smaller.

Figure 3 provides a graphical comparison of the above results. The large correlations for the D-optimal DSD have been reduced through use of the compound optimization. The darkest off-diagonal cells in Figure 3(a) correspond to an absolute correlation of 1.0, whereas the darkest off-diagonal cells in Figure 3(b) correspond to an absolute correlation 0.46.

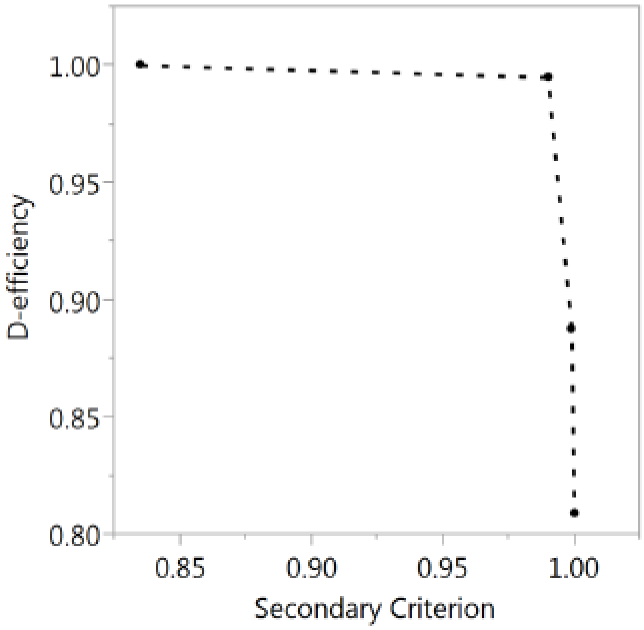
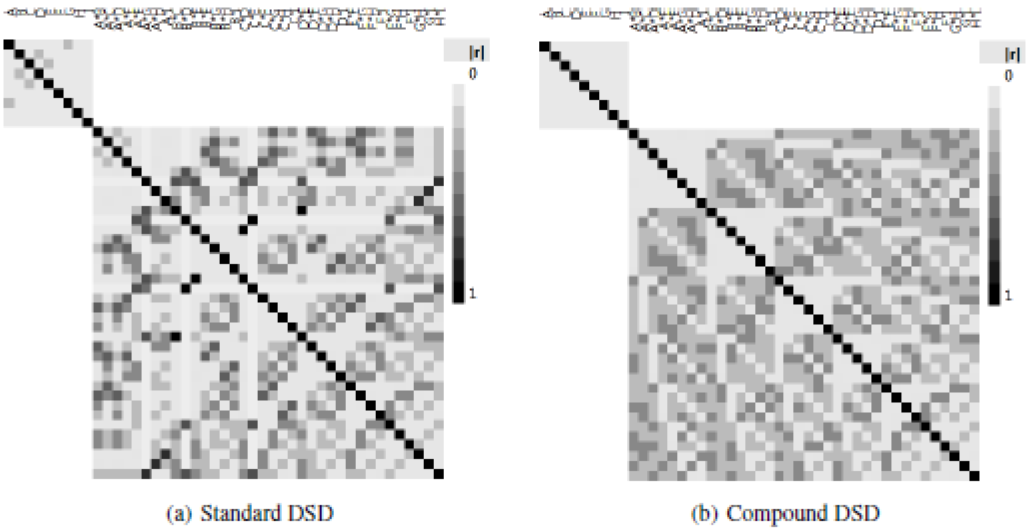


Figure 2: Pareto frontier of criterion values for non-dominated designs for *m* = 9 and *n* = 22



(a) Standard DSD

(b) Compound DSD

Figure 3: Correlation cell plots for *m* = 9 and *n* = 22

Discussion

For screening designs there are three particularly desirable features: orthogonality of the main effects, no aliasing (or partial aliasing) of main effects with two-factor interactions, and a small run size. It is often the case that these characteristics cannot be fully satisfied simultaneously, and trade-offs among them must be considered. For example, Plackett-Burman designs have small run sizes and are orthogonal, but they exhibit partial aliasing among two-factor interactions. Resolution III fractional factorial designs are also orthogonal for main effects, but some main effects and two-factor interactions are fully aliased. Resolution IV fractional factorial designs are another orthogonal alternative, however two-factor interactions are completely confounded with other two-factor interactions.

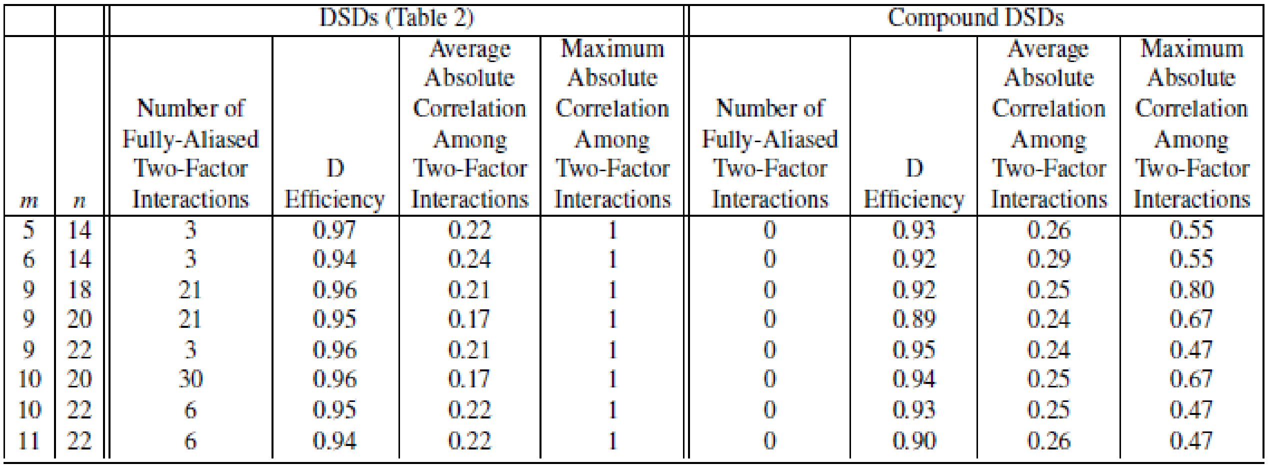
Two-level DSDs represent a different kind of tradeoff. These are small designs that sacrifice full orthogonality of main effects for complete independence between main effects and two factor interactions. This approach was introduced by Margolin (1969), and strongly advocated by Miller and Sitter (2005), who studied the robustness properties of Margolin's designs. Two-level DSDs improve upon and extend the class of Margolin-type designs. In addition, using the methodology of Section 5, they can be constructed in such a way that two-factor interactions are not completely confounded with other two-factor interactions. This provides the experimenter with the opportunity to identify, not only key main effects, but also active two-factor interactions in the presence of sparsity.

Lack of orthogonality has two related consequences: the parameter estimates have longer confidence intervals than an orthogonal design based on the same number of runs, and the power to detect an effect is reduced, again in comparison to an orthogonal design of the same size. Consider the cases of non-orthogonal DSDs in Table 6. This table shows the fractional increase in the maximum standard error of all the parameter estimates of the main effects compared to the standard error for an orthogonal design (assuming one exists, and in most of these cases no such design exists).

One way to reduce such impact is minimize non-orthogonality of the design. For this purpose one could employ $E(s^2)$ as a driving criterion (subject to a lower bound on the D-efficiency of the design). We have used this approach for the cases considered in Table 3 and have found some designs that have better correlation structures.

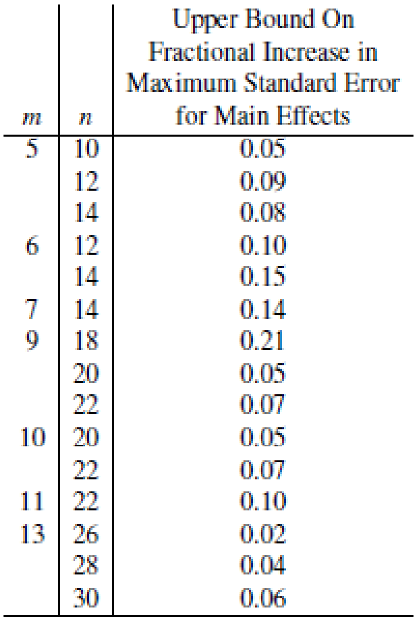
Consider, for example, the 14 × 7 DSD reported in Table 3. According to Table 2, this design has $D = 0.92$, and 18 pairs of the seven factors have an absolute correlation of 0.14, and the remaining 3 pairs have an absolute correlation of 0.43. By using the $E(s^2)$ criterion, we constructed a design for which all 21 pairs have absolute correlation equal to 0.14. The *D*-efficiency of this design is slightly lower at $D = 0.89$. Considering the drop in the maximum absolute correlation from 0.43 to 0.14, the latter design may be preferred by some practitioners.

Table 5: Compound DSDs from selected cases in Table 3[1]



		DSDs (Table 2)				Compound DSDs			
m	*n*	Number of Fully-Aliased Two-Factor Interactions	D Efficiency	Average Absolute Correlation Among Two-Factor Interactions	Maximum Absolute Correlation Among Two-Factor Interactions	Number of Fully-Aliased Two-Factor Interactions	D Efficiency	Average Absolute Correlation Among Two-Factor Interactions	Maximum Absolute Correlation Among Two-Factor Interactions
5	14	3	0.97	0.22	1	0	0.93	0.26	0.55
6	14	3	0.94	0.24	1	0	0.92	0.29	0.55
9	18	21	0.96	0.21	1	0	0.92	0.25	0.80
9	20	21	0.95	0.17	1	0	0.89	0.24	0.67
9	22	3	0.96	0.21	1	0	0.95	0.24	0.47
10	20	30	0.96	0.17	1	0	0.94	0.25	0.67
10	22	6	0.95	0.22	1	0	0.93	0.25	0.47
11	22	6	0.94	0.22	1	0	0.90	0.26	0.47

Table 6: Upper bound on fractional increase in the maximum standard error of main effects compared to the standard error for an ideal orthogonal design



m	*n*	Upper Bound On Fractional Increase in Maximum Standard Error for Main Effects
5	10	0.05
	12	0.09
	14	0.08
6	12	0.10
	14	0.15
7	14	0.14
9	18	0.21
	20	0.05
	22	0.07
10	20	0.05
	22	0.07
11	22	0.10
13	26	0.02
	28	0.04
	30	0.06

[1] The table below refers to the designs in Table 2, however, in the enumeration of this document it has been so far referred as Table 3.

Analysis Strategies for Model Selection with Definitive Screening Designs

Definitive Screening Designs (DSDs) were recently introduced by Jones and Nachtsheim (2011). These designs offer a new way to conduct screening experiments and optimization in one step. The use of three-level factors and the desirable aliasing structure of the DSDs make them suitable for identifying main-effects and second-order terms in one stage of experimentation.

However, a comprehensive investigation of the best methods for analysis of these designs has not yet been conducted. This paper explores various variable selection methods in a variety of settings for a simulation study and methods comparisons.

Let us remark again what are the purposes of screening experiments and what are the properties of three-level DSDs, as this work is focused on that class of designs.

Screening experiments are used to identify a set of active effects within a large set of potential factors. In fact, some of the factors under investigation may actively influence a certain response of interest, whereas other may not. Because screening experiments are typically performed when a large number of factors is to be investigated, they usually have the primary goal of identifying active main-effects. Small, orthogonal main-effects plans have been traditionally preferred in early stages of experimentation. After the initial analysis, follow-up experiments are frequently used to identify interactions or higher-order terms in response surface models. However, this traditional two-stage approach can be very cost inefficient, or in some experimental conditions, unsuitable.

Definitive Screening Designs (DSDs), introduced by Jones and Nachtsheim (2011), provide a new approach for screening experiments that potentially allows the experimenter to comprehensively study all the active first- and second-order effects in one stage of experimentation. They used numerical methods to construct DSDs and found that these designs were orthogonal for 6, 8, and 10 factors. Xiao, Lin, and Bai (2012) showed how to use conference matrices to construct orthogonal DSDs for most even numbers of factors and how to easily construct the DSDs for odd number of factors starting from a DSD with even factors. A typical DSD is formed by m pairs of fold-over rows in which one factor is kept at the central level and the others are at levels ±1 (respectively $\mp$ in the foldover rows); a final row has all factors at the center point. For *m* factors, these designs require 2*m*+1 runs and have the structure described in Table 1, where $x_{i,j}$ denotes the setting of the j-th factor for the i-th run. The fold-over structure and the settings of the center points for one factor in each pair of rows have, as a consequence, the orthogonality of all main-effects with respect to all second-order terms; consequently the estimation of the main-effects is unbiased by any second-order term, and moreover, any pair of second-order terms is not fully aliased. A DSD having six factors or more projects to efficient response surface designs with three or fewer factors.

Because of their properties, DSDs are not meant to be used for the estimation of the main effects only, but also higher-order effects. With respect to models that include higher-order effects, a DSD can be considered as a supersaturated design. Generally, a supersaturated design is a design in which the number of columns exceeds the number of rows. This means that *p*, the number of parameters of interest, is greater than *n*, the number of observations, or design points. This concept usually applies to designs in which the number of factors investigated is larger than the available degrees of freedom. However, this same idea applies when the *p* parameters of interest are not only the main-effects, but also some interactions and curvature effects. For the main-effects model, a DSD has enough degrees of freedom for the estimation of the main-effects and the error term. However, as noted above, if the experimenter wishes to investigate higher-order effects, then the design becomes supersaturated. In fact, the number of rows is less than the number of parameters that would need to be estimated if a full second-order model is to be fitted. Given the small number of runs, not all effects can be estimated by the same model. This gives rise to a model selection problem that could be handled with different variable selection methods.

The main research question that we address in this work concerns the identification of the best methods for analysis of a DSD. We assess how well different model selection methods work in a variety of situations, such as different numbers of active effects, or different sizes of the effects. On the one hand, screening experiments are frequently carried out assuming sparsity; some empirical papers have demonstrated how sparsity and other empirical principles often hold in practice (see Li, Sudarsanam, & Frey, 2006). For this reason, we conduct a simulation study that assumes sparsity, strong heredity, and hierarchy of the effects. On the other hand, it can be the case that there are no-heredity effects, or heredity is weak (when in order for an interaction effect to be active one of the involved main-effects has to also be in the model). To make our study more comprehensive, we investigated different cases, relaxing the strong heredity assumption. We aim to find in what situations the analysis of each method fails to correctly identify the active terms. The remainder of the paragraph is again is organized as in the paper. We provide an overview of the related literature on the construction and analysis of supersaturated designs for screening experiments. Then, we detail our study design, we summarize the results of our simulation study. Finally, we conclude with a discussion and suggestions for practitioners.

Previous work

Designs for screening experiments

Screening is often the first step for an experimentation when there is a large number of potential causal factors in a system but only a few are expected to influence the response of interest. As explained by Wu and Hamada (2008), these experiments tend to be economical in the sense that they are usually small main-effects orthogonal designs, where a few degrees of freedom are left for estimating error variance and higher-order terms, such as quadratic effects or interactions. Once the important variables are identified, a follow-up experiment is sometimes conducted for a thorough exploration of their effects on the response. This second step of experimentation frequently falls into the category of response surface exploration. Response surfaces are usually based on larger designs that allow the estimation of all linear and quadratic main-effects and all two-factor (linear-by-linear) interactions effects. In order to estimate all of these effects simultaneously it is necessary to fit a full second-order model.

Traditionally, screening experiments have been based on small main-effects orthogonal designs, such as resolution III regular Fractional Factorial Designs or irregular designs such as Plackett-Burgman Designs. These designs are two-level designs, and for this reason they do not allow the estimation of quadratic effects. Two-factor interactions are not estimable in this first stage because they are either fully (in the case of regular resolution III designs) or partially (in the case of irregular designs) confounded with main-effects. For these reasons, follow-up experiments are frequently required to obtain a complete analysis.

DSDs offer an opportunity to combine the two stages described above in a single stage of experimentation. These designs, as described in the introduction, combine a desirable aliasing structure that allows an unbiased estimation of main-effects, and a strategic use of center points that allows the estimation of quadratic effects, while avoiding full confounding between pairs of second-order effects. The partial confounding between pairs of second-order effects makes it possible to identify the active terms if the true model is sufficiently sparse, since the sample size is smaller than the number of potential effects.

A supersaturated design is characterized by fewer runs than effects to be estimated. Thus, for instance, a design can be unsaturated when only main-effects are of interest but supersaturated when interactions are also to be investigated (Dragulji et al., 2014). In the case of DSDs the designs are unsaturated for the main-effects but supersaturated for the estimation of interactions; however, the orthogonality of the main-effects in respect to any second-order terms and the partial confounding of pair of second-order terms, give hope to the experimenter to conduct a more accurate model selection. In addition, if three or fewer main-effects are active, these designs projects to a highly efficient response surface design, and model selection becomes much easier. In this case, the full quadratic model is estimable. For supersaturated designs columns are not all pairwise orthogonal. For this reason, standard analysis techniques for orthogonal designs such as half-normal plots are not applicable. However, by the nature of the construction some interactions cannot be entertained, and a careful assignment of the factors to the columns may minimize aliasing between important factors. Supersaturated designs are a good option for screening experiments because they are constructed based on the sparsity-of-effect practice. Sparsity refers to the observation that the number of relatively important effects in a factorial experiment is generally small (Box & Meyer, 1986). This is also called the Pareto Principle in Experimental Design. In addition to the effect sparsity principle, the hierarchical principle states that the lower-order effects are more likely to be significant than the higher-order effects, and the heredity principle indicates that an interaction is significant one if one or both of its parent effects are significant, which are respectively called the weak and strong heredity (Wu & Hamada, 2008). This regularity can strongly influence sequential, iterative approaches to experimentation. Heredity can also provide advantages in analyzing data from experiments with complex aliasing patterns, enabling experimenters to identify likely interactions without resorting to high-resolution designs (Chipman, Hamada, & Wu, 1997).

Li et al. (2006) find that these well-known empirical principles very often hold in practice. The authors analyze a set of experimental designs results that confirm the general principles of sparsity, heredity, and hierarchy. In respect to heredity, this study finds that the probabilities that second-order effects are active depend on how likely the corresponding main-effects are active. If both main-effects involved in an interaction effect are active, this interaction effect follows the strong heredity principle; if only one of the two is an active main-effects, the effect is said to follow the weak heredity principle; if any of the main-effects involved in the interaction are active, then there is no heredity in the second-order term. Chipman et al. (1997) consider such probabilities 0.25 for strong heredity effects, 0.10 for weak heredity effects, and 0.01 for no heredity effects. Li et al.

(2006) empirically find these probabilities to be 0.33, 0.045, and 0.0048 respectively; moreover, they find that the probability of a main-effect to be active is 0.40.

Analysis, model selection methods and comparison studies

DSDs represent a promising alternative to traditional screening approaches from a design perspective. However, a comprehensive investigation of the best methods for analysis of these designs has not yet been conducted. The literature offers some studies of the analysis of other classes of supersaturated designs. Most previous studies concern two-level designs, but they explore alternative variable selection methods that are easily adapted to the analysis of DSDs.

For screening designs, Dragulji et al. (2014) presented a comprehensive comparison of screening strategies. They conducted a simulation study that compares two screening approaches, supersaturated designs and group screening, with several variable selection methods: LASSO, Smoothly Clipped Absolute Deviation (SCAD), Gauss-Dantzig selector, Simulated Annealing Model Search (SAMS), Bayesian Model Selection, and Maximum A Posteriori (MAP) estimation. The literature on variable selection methods proposed for the analysis of supersaturated designs is very diverse. Wu (1993) suggested the use of forward selection methods to identify active main-effects. Lin (1993) suggested the use of stepwise selection procedures. Westfall, Young, and Lin (1997) utilized a modified forward model selection; Lu and Wu (2004) suggested a stepwise selection based on staged dimensionality reduction. Chipman et al. (1997) proposed a Bayesian variable selection method; and Beattie, Fong, and Lin (2002) used a two-stage Bayesian approach for model selection. Li and Lin (2002) introduced a variable selection method via nonconvex penalized least squares that employs an iterative ridge regression. Holcomb, Montgomery, and Carlyle (2003) proposed contrast-based methods; Zhang, Zhang, and Liu (2007) a method based on partial least squares. Phoa, Pan, and Xu (2009) used a simulation study to examine the application of the Dantzig selector described by Candes and Tao (2007). Marley and Woods (2010) performed a simulation study that compares two classes of supersaturated designs and three methods for variable selection: forward stepwise regression (Miller, 2002), Gauss-Dantzig selector (Candes & Tao, 2007; Poha et al. 2009), and their proposed model averaging procedure.

Variable selection methods that directly address principles of sparsity, heredity, and hierarchy were also considered in the literature. Yuan, Joseph, and Lin (2007) proposed a modification of Least-Angle Regression (LARS) (Efron, Hastie, Johnston, & Tibshirani, 2004) to account for the empirical principles. Choi, Li, and Zhu (2010) proposed a method called Strong Heredity Interaction Model (SHIM) that automatically enforces the strong heredity constraints in the penalty function. Later Li and Zhu (2014) proposed a similar method, called Weak Heredity Interaction Model (WHIM), to allow also weak heredity effects.

Study model

The purpose of this study is to evaluate the efficiency of the analysis of screening experiments conducted through Definitive Screening Designs. We compare several popular and promising model selection methods using computer simulation in a variety of settings. We vary the number of factors in the design, the level of sparsity, and signal-to-noise ratio.

Number of factors and sparsity

Based on the existing literature, we set up our simulation study for different levels of sparsity and in one setting we assumed strong heredity; this assumption may look restrictive, for this reason we considered also the case that allows all kinds of second-order effects to be active. We conducted two simulations: in the first, we varied an increasing number of active main-effects in the true model, and for each number of active main-effects we varied the sparsity of the second-order effects, as will be explained. In the second, we let the number of active second-order-effects vary randomly according to certain probabilities, based on the results of the empirical study of Li et al. (2006). Results of the latter will be discussed and compared to our first simulation settings in the next section.

In terms of design size, we investigate the analysis of DSDs having 6, 10, and 14 factors. In each design, the number of active factors varies from 2 to the total number of factors, in steps of size two. Note that if we follow the empirical evidence of Li et al. (2006), we should expect 40% of the factors in the design to be active main-effects. Our simulation considers all these possibilities for sparsity of main-effects in combination with varied levels of sparsity of the second-order effects. Regarding sparsity in second-order effects, we evaluate three cases:

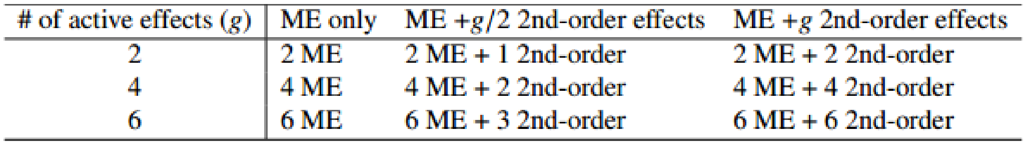
1. Only main-effects are active;

2. Some second-order effects are active, and their number is half the number of active MEs;

3. Some second-order effects are active, and their number is the same as the number of active MEs.

These settings led to all cases examined in our first simulation, in a full factorial manner, in combination with the signal-to-noise ratio cases. For instance, when evaluating the DSD with 6 factors, Table7 displays the nine cases for the true model:

Table 7: Effects in the nine true models for the 6-factor DSD



# of active effects (g)	ME only	ME +$g/2$ 2nd-order effects	ME +g 2nd-order effects
2	2 ME	2 ME + 1 2nd-order	2 ME + 2 2nd-order
4	4 ME	4 ME + 2 2nd-order	4 ME + 4 2nd-order
6	6 ME	6 ME + 3 2nd-order	6 ME + 6 2nd-order

Signal-to-noise ratio

We also evaluated the designs and the different sparsity cases for different levels of signal-to-noise ratio. Consider the model Y = Xβ + **ε**, we allowed the β of the active terms in the true model to vary according to a cascading function. Assume there are *k* active factors. The nonzero regression coefficients are chosen such that: (1) the mean of the absolute values of the *k* regression coefficients is given by the signal size; and (2) for *i* ≥ > 1, the absolute value of the *i*-th largest coefficient is $0.75^{(i-1)}$ times absolute value of the largest coefficient. For example, if *k* = 4 and the signal level is 3, the absolute values of the four coefficients, in descending order, are: 4.3886, 3.2914, 2.4686, and 1.8514. The signs of the coefficients are chosen at random.

Variable selection methods and selection criteria

We explored different options for variable selection. The goal of variable selection is to identify the smallest subset of the set of potential predictors that clearly explain the data (Wu & Hamada, 2008). One strategy is to use a model section criterion to evaluate all possible subsets of the covariates and select the model with the best value for the criterion.

This approach is called Best Subset Regression or All Possible Models (for a deeper exposition of subset selection procedures see Miller, 1990). With *p* parameters of interest, there are 2^p possible models to evaluate – fewer if one considers only models that follow the strong heredity principle, or models that follow the weak heredity principle. Clearly, as *p* becomes large (e.g. greater than 20), Best Subset results in an infeasible strategy. Nonetheless, we ran one simulation including this method, and we found minimal difference in performance with respect to other methods that would have justified the extremely high computing time.

Another popular approach to model selection is stepwise regression. The classical forward stepwise procedure starts with the null model and adds the most significant factor main-effect at each step according to an F-test (Miller, 2002). The procedure continues until the model is saturated or no further factors are significant. The evidence required for the entry of a variable is controlled by the probability to enter, denoted by the α level. Stepwise regression can also be conducted as a backward procedure, starting with all terms in the model and then dropping one at each step, but the approach is unfeasible for SS designs, where $n \leq p$. Stepwise regression can be also applied in conjunction with other selection criteria, as will be described, such as *AIC*, or *BIC*. With these approaches, the idea is to add the variable, at each step, that leads to a new modeling having the best value by that particular criterion. When reach the *n*-term model, stop and choose the model that led to the best criterion value.

LASSO, least angle shrinkage and selector operator, introduced by Tibshirani (1996) with modifications introduced by Efron (2006), is another option for model selection that has become very popular. This method and the Dantzig selector (Candes & Tao, 2007) are two methods that we compare to stepwise regression.

When the model consists of both main effects and two-factor interactions, and the true model follows heredity, Choi et al. (2010) and Li and Zhu (2014) proposed effective model selection methods for dealing with strong and weak heredity, respectively. Denote the linear model with p main effects and their two-factor interactions by

$$y = \beta_0 + \beta_1 x_1 + \ \dots + \beta_p x_p + \ \alpha_{12} x_1 x_2 + \ \dots + \ \alpha_{p-1,p} x_{p-1} x_p + \epsilon \qquad (1)$$

When the strong heredity holds, Choi et al. (2010) proposed to re-parameterize the coefficient $\alpha_{jj'}, (j < j'; j, j' = 1, \dots, p)$ as $\alpha_{jj'} = \gamma_{jj'} \beta_j \beta_{j'}$.

For the purpose of variable selection, we can consider the following penalized least squares criterion:

$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_j \beta_j x_{ij} - \sum_{j<j'} \alpha_{jj'} x_j x_{j'}{}^2 + \lambda_\beta (|\beta_1| + \ \dots + \ |\beta_p|) + \lambda_\gamma (|\gamma_{12}| + \ \dots$$

$$+ \ |\gamma_{p-1,p}|). \quad (2)$$

As pointed out by Choi et al. (2010), by writing the coefficient as a product, the coefficient of an interaction term does not equal zero only if both of its parent terms are not equal to zero. Thus, the strong heredity constraint is built into this model. Similarly, when the coefficient of the interaction term $\alpha_{jj'}, (j < j'; j, j' = 1, \dots, p)$is reparameterized as $\alpha_{jj'} = \gamma_{jj'}(|\beta_j| + |\beta_{j'}|)$. the weak heredity constraint holds (Li and Zhu, 2014).

Finally, we develop a simple algorithm that specifically leverages the properties of DSDs and compared it with the above mentioned variable selection methods. DSDs have a particular structure that allows them to project in full response surfaces if no more than

3 main-effects are active. This property, together with the orthogonality of main-effects, suggests a two-stage procedure for conducting variable selection:

1. Select the active main-effects in the model, denoted by g.

2. If g ≤ 3, evaluate all possible models in the full quadratic matrix that comprises interactions and quadratic terms with the selected MEs. Otherwise, if g > 3, for each group of 3 out of the selected MEs, perform the same analysis. Finally select the final model according to a defined criterion.

For each method, we considered the available model selection criteria from existing literature. According to the *principle of parsimony*, a more parsimonious model with fewer variables should be preferred as long as it explains the data well. There must be a balance between data fitting and predictions. In fact, a model that fits the data too well may give poor predictions, as the variance of the estimated regression coefficients depends on the number of parameters in the model. Model selection criteria should then reward model fitting but penalize the increase in model complexity (which is the increase of number of terms in the model).

One commonly used criterion is the Akaike Information Criterion AIC (Akaike, 1973), defined as

$$AIC = N \log\frac{SSE}{N} + 2p, \qquad (1)$$

This criterion tends to over-fit the model when the sample size is small. Modifications of the AIC criterion have been reported in literature. Hurvich and Tsai, (1989) developed the *AICc* criterion, where:

$$AICc = AIC + \frac{2(p+1)(p+2)}{N-p-2}, \qquad (2)$$

Phoa et al., (2009), proposed instead the following modification:

$$mAIC = N \log\frac{SSE}{N} + 2p^2. \qquad (3)$$

Performance measures

In order to evaluate and compare the efficacy of variable selection methods in the different cases, we use the measures of sensitivity and specificity, as described in Choi et al. (2010). Moreover, we add two additional performance measures. All the measures are defined as follows:

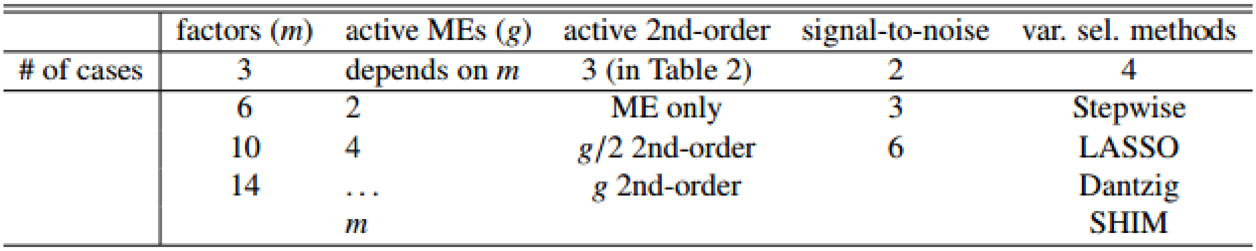
- Sensitivity is the number of correctly selected terms over the total number of active terms;
- Specificity is the number of correctly unselected terms over the total number of non-active terms;
- Percentage correct is the sum of the number of active terms selected and non-active terms unselected, over the total number of parameters in the full quadratic model;
- Root mean square error of predictions is a measure that we evaluate over the selected model and a random sample of a thousand design points.

Together these measures give us a way to evaluate the screening capacity of the methods, which is the ability to select the correct active terms and to not select the non-active ones, and the ability to make predictions with the selected model.

Simulation algorithms

We conduct two simulation studies. In the first study, we consider the models following the strong heredity principle. Table 8 summarizes the four factors, as well as their levels, in the simulation study.

Table 8: Factors considered in the simulation



	factors (m)	active MEs (g)	active 2nd-order	signal-to-noise	var. sel. methods
# of cases	3	depends on m	3 (in Table 2)	2	4
	6	2	ME only	3	Stepwise
	10	4	$g/2$ 2nd-order	6	LASSO
	14	...	g 2nd-order		Dantzig
		m			SHIM

In total, there are (3 + 5 + 7) × 3 × 2 × 4 = 360 cases.

The simulation procedure then works as follows:

- For each m, consider each number of active MEs (g) shown in Table 8.
- For each number of active 2nd-order effects shown in Table 8, randomly generate nm = 10 models that follow strong heredity.
- Obtain the β coefficients by using the two signal-to-noise ratios shown in Table 8, and then simulate the responses.
- For each generated model and data, perform the model selection procedure ns = 10 times, using each of the four variable selection methods considered.

In total, there are (3 + 5 + 7) × 3 × 2 × nm × 4 × ns = 36, 000 simulations performed.

In a second simulation setting we allow the true model to be constructed randomly, according to the conditional probabilities that a second-order effect is active. We still explore different cases of numbers of active MEs.

In this second setting, the simulation runs 675, 000 times. Since the actual number of second-order effects active terms is a random variable, we can calculate its expected value. For example, in a DSDs with 6 factors and 4 active main-effects, there are $\binom{4}{2}$ interactions effects and 4 quadratic effects that follow strong heredity; $\binom{2}{2}$ interactions plus 2 quadratic terms that do not follow heredity; and the remaining ones follow the weak heredity principle. These effects are randomly picked with probabilities 0.25 for strong heredity effects; 0.1 for weak heredity effects; 0.01 for no heredity effects. According to these probabilities then the expected number of active second-order effects is 3.33.

Results and discussion

In this section we discuss the results of our simulation for the design with 6 factors. We conducted the same simulation for 10 and 14 factors, and we saw that all findings hold in these cases as well.

Moreover, the following refer to the case of signal-to-noise ratio equal to 3. Again, we did not see a substantial relative difference when imposing a different signal-to-noise value. We initially did some preliminary testing on separate variable selection methods. We wanted to make sure that each method could be compared to the others in the best setting of simulation variables and tuning parameters. Separate studies led us to set up all tuning parameters of the different variable selection methods, as for instance in LASSO, Dantzig selector, and SHIM. In the very beginning of our study, we simulated the Best Subset procedure and we soon noticed that the extremely long computing time was not worth any substantial improvement in performance compared to other model selections.

In including stepwise regression in our comparison with the more sophisticated methods, we first explored in which setting this method offered the best performance. We initially tested forward stepwise regression based on probability to enter level, α, equal to 0.05, 0.10, and 0.20. We found that a higher value of α results in higher sensitivity, and we retained $\alpha = 0.2$ for further comparison. Through the test of different model selection criteria in conjunction with the other methods, like LASSO, Dantzig, and our DSD-specific method, we determined that *AIC* was giving us slightly better results; for this reason, we decided to maintain the use of this criterion in all methods of our simulation and consequently also apply this criterion in stepwise regression.

Focusing on the measure of sensitivity, it is possible to compare the variable selection methods in the primary purpose of screening, which is to correctly identify active terms. In case of only *MEs* in the active terms, all methods reach the value of sensitivity equal to 1. This is true both in the case of sparse effects and in the extreme case where all *ME* are active. This is a very important finding that shows the high efficacy that is related to very efficient designs such the DSDs. The designs prove to be robust in the correct identification of active main-effects, regardless the variable selection method used in the analysis. The performance of all methods obviously decreases when the number of active second-order terms increases. It appears that overall there is no clear winner, especially when the tradeoff between sensitivity and specificity is considered. However, if again we focus on sensitivity as our main goal in screening, then the best performance is the modified SHIM algorithm. Our proposed DSD-specific method that considers the design projections in 3 or less *MEs*, results by construction penalized when more than 3 factors are active. This happens mainly because of the way we construct the true model without restricting the number of factors that are involved in second-order terms. If one believes that in reality, or in a particular case, no more than 3 factors are involved in higher-order terms, regardless of the number of active linear terms, then this method is a very interesting and valid alternative to the others and specifically leverages the properties of the DSDs. In any case, it may also be worth considering the implementation a model averaging procedure.

We conducted our main simulation under the assumption of strong heredity, and we investigated different levels of sparsity. To make our analysis more complete, we ran the same simulation with a modification that allows the true model to be constructed following the probabilities found in the empirical investigation of Li et al. (2006). We did not enforce the number of active *MEs* to be determined by the empirical probability, and we kept our initial choice to investigate different numbers of active *MEs*; if we would have followed that probability of 40%, we would have focused on the case with 2 active *MEs* in the design with 6 factors. This case may be considered as the most representative of a real case. According to the above mentioned probabilities that a second-order effect is active conditionally to the fact that the *MEs* involved in that term are active or not (strong, weak and no-heredity terms), the expected number of active second-order terms is 1.47 when the active *MEs* are 2; 3.47 when the active MEs are 4; 6.93 when the active *MEs* are 6. Running our simulation with 50 different randomly constructed models that obey these probabilities, again, we did not see a clear winner, especially in consideration of the sensitivity/specificity tradeoff. However, it is interesting to note that still, in terms of sensitivity, SHIM performs better than the other methods, in addition to the fact that this method enforces strong heredity and the true model does not follow this principle.

Acknowledgments

This research has been funded by the Juran Research Center through the Juran Fellowship Award 2013.

References

- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected Papers of Hirotugu Akaike (pp. 199-213). Springer New York.
- Albrecht, M. C., Nachtsheim C. J., Albrecht T. A., and Cook R. D. (2013). "Experimental Design for Engineering Dimensional Analysis". Technometrics 55 (3), 257–270.
- Beattie, S. D., Fong, D. K. H., and Lin, D. K. J. (2002). A two-stage Bayesian model selection strategy for supersaturated designs. *Technometrics*, 44(1), 55-63.
- Box, G.E.P., Meyer, R.D., (1986). An analysis for unreplicated fractional factorials. *Technometrics*, 28, 1118.
- Burnham, K. P., and Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach. Springer.
- Candes, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when *p* is much larger than *n*. *The Annals of Statistics*, 2313-2351.

- Chipman, H., Hamada, M., and Wu, C. F. J. (1997). A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, 39(4), 372-381.
- Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489), 354-364.
- Dragulji, D., Woods, D. C., Dean, A. M., Lewis, S. M., and Vine, A. J. E. (2014). Screening strategies in the presence of interactions. *Technometrics*, 56:1, 1-16.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32, 407499.
- Hamada, M. and Wu, C.F.J., (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24, 130137.
- Holcomb, D. R., Montgomery, D. C., and Carlyle, W. M. (2003). Analysis of supersaturated designs. *Journal of Quality Technology*, 35(1), 13-27.
- Hotelling, H. (1944). “Some Improvements in weighing and Other Experimental Techniques”. The Annals of Mathematical Statistics, 15 (3), 297–306.
- Hurvich, C. M., and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297-307.

- Jones, B. (2013). “Comment: Enhancing the Search for Compromise Designs”. Technometrics 55, pp. 278–280.
- Jones, B. and Nachtsheim, C. J. (2011a). “Efficient Designs With Minimal Aliasing”. Technometrics 53, pp. 62–71.
- Jones, B. and Nachtsheim, C. J. (2011b). “A Class of Three-Level Designs for Definitive Screening the Presence of Second-Order Effects”. *Journal of Quality Technology* 43, pp. 1–15.
- Jones, B. and Nachtsheim, C. J. (2013). “Definitive Screening Designs with Added two-Level Categorical Factors”. Journal of Quality Technology 45, pp. 121–129.
- Li, R., and Lin, D. K. (2002). Data analysis in supersaturated designs. *Statistics and probability letters*, 59(2), 135-144.
- Li, R., and Lin, D. K. (2003). Analysis methods for supersaturated design: some comparisons. *Journal of Data Science*, 1(3), 249-260.
- Li, W., and Zhu, J. (2014). Comment: Model Selection With Strong and Weak Heredity Constraints. *Technometrics*, 56:1, 21-22
- Li, X., Sudarsanam, N., and Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, 11(5), 32-45.
- Lin, D. K. (1993). A new class of supersaturated designs. *Technometrics*, 35(1), 28-31.
- Lu, X., and Wu, X. (2004). A strategy of searching active factors in supersaturated screening experiments. *Journal of Quality Technology*, 36(4), 392-399.

- Nachtsheim, C., & Jones, B. (2003). A powerful analytical tool. *ASQ.org*
- Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4), 661-675.
- Margolin, B. H. (1969). “Results on Factorial Designs of resolution IV for the 2^n and 2^n3^m Series”. Technometrics 11(3), pp. 431–444
- Marley, C. J., and Woods, D. C. (2010). A comparison of design and model selection methods for supersaturated experiments. *Computational Statistics and Data Analysis*, 54(12), 3158- 3167.
- Meyer, R. K. and Nachtsheim, C. J. (1995). “The Coordinate-Exchange for Algorithm Exact Constructing Optimal Experimental Designs”. Technometrics 37(1), pp. 60–69.
- Miller, A. (2002). Subset selection in regression. CRC Press.
- Miller, A. and Sitter, R. R. (2001). “Using Folded-Over 12-Run Plackett-Burman Designs to Consider Interactions”. Technometrics 43 (1), pp. 44–55.
- Miller, A. and Sitter, R. R. (2005). “Using Folded-Over Nonorthogonal Designs”. Technometrics 47 (4), pp. 502–513.
- Phoa, F. K., Pan, Y. H., and Xu, H. (2009). Analysis of supersaturated designs via the Dantzig selector. *Journal of Statistical Planning and Inference*, 139(7), 2362-2372.

- Sun, D. X., Li, W. and Ye, K. (2008). “Algorithmic construction of catalogs of non-isomorphic two-level designs for economic run sizes”. Statistics and Applications 6 (1 & 2), New Series, 123–140.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, Series B, 58, 267288.
- Westfall, P. H., Young, S. S., and Lin, D. K. (1998). Forward selection error control in the analysis of supersaturated designs. *Statistica Sinica*, 8(1), 101-117.
- Wu, C. F. J. (1993). Construction of supersaturated designs through partially aliased interactions. *Biometrika*, 80(3), 661-669.
- Wu, C. F. J. and Hamada M. J. (2008). Experiments: Planning, Analysis, and Optimization.
- Xiao, L., Lin, D. K., and Bai, F. (2012). Constructing definitive screening designs using conference matrices. *Journal of Quality Technology* 44(1), 2–8.
- Yuan, M., Joseph, V. R., and Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49(4), 430-439.
- Zhang, Q. Z., Zhang, R. C., and Liu, M. Q. (2007). A method for screening active effects in supersaturated designs. *Journal of Statistical Planning and Inference*, 137(6), 2068-2079.

Managerial and statistical tools for Six Sigma and marketing research: conjoint analysis and discrete choice experiments

Introduction

In the framework of the five steps of the DMAIC (Define, Measure, Analyze, Improve, Control) cycle a fundamental importance and careful attention has to be addressed in understanding the voice of the customer in order to achieve radical quality improvements and customer satisfaction. It helps define which characteristics are critical to quality from the customer perspective and it accordingly establishes a strong relationship between Statistics and Marketing Research.

The following paragraphs are arranged in such a way to first introduce how marketing research is related to Six Sigma, then the most popular techniques used in order to translate the voice of the customer in product specifications are presented and discussed in terms of evolution of the methods and recent steps forward, opening points and research opportunities for further studies.

In this perspective Quality Function Deployment (QFD) and Conjoint Analysis (CA) will be discussed. In particular great attention will be dedicated to *discrete choice experiments*, models, design criteria, and possible fields of applications.

Finally some possible path for further research studies will be highlighted from what is actually missing or still not well defined or established in the current literature.

Six Sigma and marketing research

Six Sigma is a long-term, forward-thinking initiative designed to fundamentally change the way organizations work. It is first and foremost "a business process that enables companies to increase profits dramatically by streamlining operations, improving quality, and eliminating defects or mistakes in everything a company does." (Rylander & Provost, 2006). While traditional quality programs have focused on detecting and correcting defects, Six Sigma encompasses something broader: it provides specific methods to recreate the process so that defects are never produced in the first place (Harry and Schroeder, 2000).

Tom McCarty, vice president of Motorola University Six Sigma Services and co-author of the book “The New Six Sigma”, describes this process as looking for opportunities for customer engagement. Generalized customer information may be nice to have, but targeted research that seeks answers to specific questions is more likely to yield actionable results (Colby, 2003).

One of the most popular way to obtain information from the customer is nowadays obviously the Internet. Online research delivers superior results over traditional methods because it leverages the unique strengths of the Internet by:

- eliminating group bias and dominant personalities;
- getting unrushed and thoughtful answers sample;
- having the ability to test, change, and retest on the fly (Sang, 2003).

With the Internet finding a place in the business world, online reporting is taking on a new role.

Information users are now accessing survey data from their desktops and slicing and dicing it over the Internet in ways that suit their particular needs (Hogg, 2001). Using a real-time data collection method produces actionable information virtually overnight and generally within minutes of completing a survey. Furthermore, adding the principles and elements of Six Sigma to the data makes the information dynamic, robust, and easily measured.

Many conventional approaches to tracking customer satisfaction take too long to get information to the right people or focus on less-efficient measures. Future advances should address the mass customization of the approach for multiple industries, allowing technology to work with Six Sigma practices to improve customer satisfaction (Rylander & Provost, 2006).

In the United States service organizations are now taking dissatisfaction seriously. Recent applications of the primarily manufacturing-oriented Six Sigma philosophy to services (Kim 2000; Pande, Neuman and Cavanagh 2000) point to the potential economic value of refining relevant business processes to deliver absolutely predictable, defect free, service products - that is, the creation of perfect technical quality (Woodall, 2001).

Conjoint analysis and quality improvement: understanding the voice of the customer

Well established research in the management of technology suggests that cooperation and communication among marketing, manufacturing, engineering and R&D leads to greater new-product success and more profitable products. Quality Function Deployment (QFD) is one methodology that improves communication among these functions by linking the voice of the customer to engineering, manufacturing and R&D decisions (Griffin & Hauser, 1993). QFD uses perceptions of customer needs as a lens by which to understand how product characteristics and service policies affect customer preference, satisfaction, and ultimately, sales. The methodology’s steps are addressed to the construction of the so-called House of Quality (HOQ) (Figure **4**4).

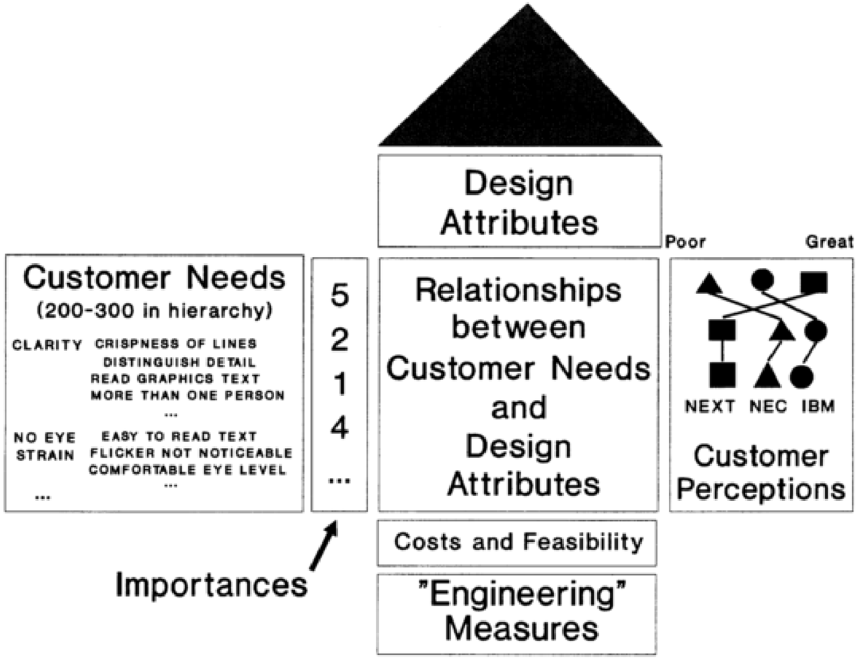


Figure 4: House of Quality. Griffin (1993)

Despite its popularity (Chan & Wu, 2002) the methodology has some limitations:

- identification of customer requirements is subjective and limited by the customer's knowledge of product design alternatives
- it evaluates and rank customer requirements with a one-requirement-at-a-time approach that ignores trade-offs and interactions among alternatives
- generally it ignores segmentation of the customers
- it is complex and time consuming

Recently, some authors have suggested the use of Discrete Choice Experiments (DCE) to elicit the voice of the customer when constructing the House of Quality (Gustafsson, Ekdahl, & Bergman, 1999; Kazemzadeh, Behzadian, Aghdasi, & Albadvi, 2008). A detailed comparison between a QFD based on HOQ and one based on DCE resulted on the conclusion that DCE perform better (Katz, 2004; Pullman, Moore, & Wardell, 2002).

Conjoint analysis, developed in fields like marketing and economics, is a useful tool for understanding the voice of the customer in order to guide quality improvement efforts, is emerging as a strategic tool, providing actionable intelligence businesses can use to go beyond product optimization to support organic growth. In other words, conjoint analysis has become a new source of insight into customer segments (Meer, 2011).

Its ultimate goal is to find what attributes and what levels of those attributes are important to the customers and give him satisfaction of the product or influence his willing to pay. This method has so far received little attention in the quality area.

Conjoint analysis has its origins in the work of Duncan Luce, mathematical psychologist, and John Tukey, statistician. According to their findings ranking of objects described by a set of attributes, measured on ordinary scales could be used to create interval attributes and response scale (Luce, 1964). Simultaneously Kruskal developed a numerical method to create such a scale, MONANOVA (Kruskal, 1964). The introduction of conjoint analysis is due to Green and Rao (Green & Rao, 1971) who showed the possible application to marketing research problems, then in 1974 Green described how to use Design of Experiments to create product profiles and choice sets (Green, 1974).

Discrete Choice Experiments

Discrete Choice Experimentation (DCE) is a marketing research methodology that uses design and analysis of experiments to find the relationship between product attributes and consumer preferences. Instead of asking each customer the importance of each attribute in a one-factor-a-time approach, a discrete choice experiment uses treatment combinations to build product prototypes to be evaluated by the respondent.

Using a full factorial design given M attributes each of those having l_i levels, the complete design implies a number of profiles to evaluate equal to $\prod_{i=1}^{M} l_i$. The respondent usually ranks the prototypes or chooses one from a set. If the profiles are rated, the methodology is referred to as *metric conjoint analysis*; when the respondent identifies his *first choice*, the method is called *choice-based conjoint* or *discrete choice experiment*.

Generally the number of combinations of attribute levels is too high and fractional factorial designs or orthogonal arrays are needed. Further approaches tend to reduce the size of choice sets, such as blocked designs, and particularly balanced incomplete block designs (Green, 1974).

In DCE each respondent selects only his favorite profile from the choice set. If J profiles are evaluated by N customers, the joint probability that N_1 customers select the profile 1, N_2 select the profile 2 and so on, is denoted $P(N_1, N_2, ..., N_j)$ and it is obtained from a multinomial distribution with parameters N, P_1, P_2, P_j, where P_j is the probability that a customer select the j-th profile.

The non-linear regression model called Multinomial Logit (MNL) is used to relate customer preferences to product attributes (Train, 2003). In DCE several issues arise, making them different and more complicated than usual Design of Experiments (DOE):

- The use of *non-linear models*: so that generalized linear approach is required for the analysis because the response model is not linear in the parameters (Kutner & al. 2005)
- *Subject differences*: the subject-to-subject variation might be important to take into account, it can results from gender, age, socio-economic or other demographical variables. In standard DOE non-homogeneity of the experimental units is controlled by blocking. Here instead identifying and describing differences among customers may be one research objective, as for instance for market segmentation.
- *Subject fatigue and missing observations*: any subject can tolerate a limited number of evaluating tasks so that the number of profiles need to be limited. It's common that the respondent pays attention to the first choice sets, and then the fatigue arises and influences his answers which became quicker and inaccurate (Batsell & Louviere, 1991). A reasonable design ranges usually from 2 to 4 profile per choice set, for a total of 8 to 16 choice sets.
- *Identification of interactions*: most of DCE employ main effects design, stressing the assumption of additivity and disregard the effects of interactions. On the other hand however, there is the need to limit fatigue that motivates the use of small designs.

Some advantages of DCE enforce the use of such method instead of metric conjoint or other popular methodology which are used to study the voice of the customer, such as Quality Function Deployment (Katz, 2004; Pullman et al., 2002).

Among those:

- *External validity*: choosing the preferred profile and not evaluating the single attributes, the customer actually simulates what he actually does when he made a purchase decision or any other decision task.
- *Ability to assess tradeoffs*: in evaluating the desirability of a set of attributes jointly, it is possible to assess interactions among them.
- *Ability to explore the entire space*: a well design experiment can explore any possible product alternative as combination of attributes' levels.
- *Customer segmentation*: through the analysis of the DCE results and sophisticated clustering methods.

Multinomial Logit Model

The basis and theoretical argumentations regarding the MNL model will be now detailed and discussed.

The model is based on the idea the respondent, in a given choice set, will choose the profile having for him the maximum utility, defined generally as "the net benefit derived from taking some actions" (Train, 2003).

A decision maker *n*, faces *J* alternatives in the choice set *t*. The utility that the decision maker obtains from alternative *j* can be written as:

$$U_{njt} = V_{njt} + \varepsilon_{njt}$$

where V_{njt} is known by the researcher up to some parameters, and an unknown part ε_{njt} is treated by the researcher as random. So V_{njt} is function of the profile's attributes:

$$U_{njt} = \boldsymbol{f}^{\boldsymbol{T}}(\boldsymbol{x}_{njt})\boldsymbol{\beta} + \varepsilon_{njt}$$

The logit model is obtained by assuming that each ε_{nj} is independently, identically distributed with an extreme value distribution, also called Gumbel.

The density for each unobserved component of utility is

$$f\ (\varepsilon_{njt}\ )\ =\ e^{-\varepsilon_{njt}}\, e^{-e^{-\varepsilon_{njt}}}$$

and the cumulative distribution is $F(\varepsilon_{njt}\ )\ =\ e^{-e^{-\varepsilon_{njt}}}$

If ε_{nj} and ε_{ni} are i.i.d. extreme value, then $\varepsilon_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$ follows a *logistic* distribution.

The extreme value distribution gives slightly fatter tails than a Normal, which means that it allows for slightly more aberrant behavior than the Normal. Usually, however, the difference between extreme value and independent Normal errors is indistinguishable empirically.

The key assumption is not so much the shape of the distribution as that the errors are independent of each other. This independence means that the unobserved portion of utility for one alternative is unrelated to the unobserved portion of utility for another alternative. It is a fairly restrictive assumption. However, it is important to realize that the independence assumption is not as restrictive as it might at first seem, and in fact can be interpreted as a natural outcome of a well-specified model.

Under the assumption of independence, the error for one alternative provides no information to the researcher about the error for another alternative. Stated equivalently, the researcher has specified V_{njt} sufficiently that the remaining, unobserved portion of utility is essentially "white noise."

If the researcher thinks that the unobserved portion of utility is correlated over alternatives given his specification of representative utility, then he has three options:

- use a different model that allows for correlated errors
- representative utility so that the source of the correlation is captured explicitly and thus the remaining errors are independent
- use the logit model under the current specification of representative utility, considering the model to be an approximation.

The logit choice probability that decision maker n chooses alternative i is (McFadden, 1974):

$$P_{njt} = Prob\{V_{njt} + \varepsilon_{njt} > V_{nit} + \varepsilon_{nit}\}\ \forall i \neq j$$

$$= Prob\{\varepsilon_{nit} < \varepsilon_{njt} + V_{njt} - V_{nit}\}\ \forall i \neq j$$

Some algebraic manipulation results in a simple closed form expression:

$$P_{njt} = \frac{e^{V_{njt}}}{\sum_i e^{V_{nit}}}$$

which is the logit choice probability.

Representative utility is usually specified to be linear in parameters: $V_{njt} = \ \beta X_{njt}$

So the logit probability can be written as:

$$P_{njt} = \frac{e^{\beta X_{njt}}}{\sum_i e^{\beta X_{nit}}}$$

The logit probabilities exhibit several desirable properties. First, P_{nit} is necessarily between 0 and 1, as required for a probability. When V_{nit} rises, reflecting an improvement in the observed attributes of the alternative, with V_{njt} $\forall j \neq i$ held constant, P_{nit} approaches to 1. Second, the choice probabilities for all alternatives sum to one.

The relation of the logit probability to representative utility is sigmoidal, or S-shaped. This shape has implications for the impact of changes in explanatory variables. If the representative utility of an alternative is very low compared with other alternatives, a small increase in the utility of the alternative has little effect on the probability of its being chosen: the other alternatives are still sufficiently better such that this small improvement doesn't help much. Similarly, if one alternative is far superior to the others in observed attributes, a further increase in its representative utility has little effect on the choice probability. The point at which the increase in representative utility has the greatest effect on the probability of its being chosen is when the probability is close to 0.5, meaning a 50-50 chance of the alternative being chosen. In this case, a small improvement tips the balance in people's choices, inducing a large change in probability. The sigmoidal shape of logit probabilities is shared by most discrete choice models and has important implications for policy makers (Figure **5**5).

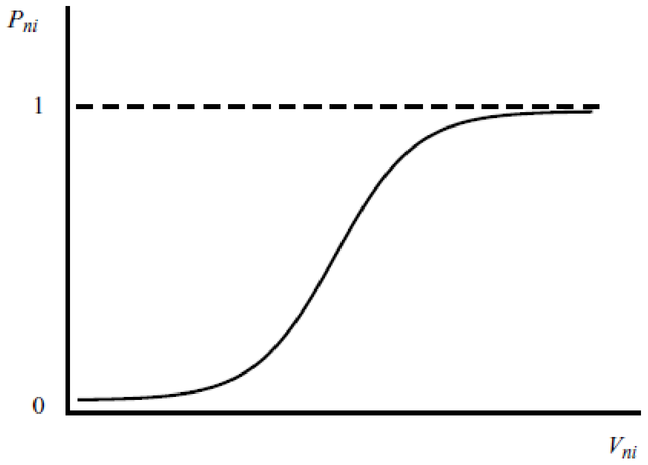


Figure 5: Logistic distribution. Train (2003)

The parameters estimation can be achieved using the maximum likelihood function:

$$L(\beta) = \prod_{n}\prod_{j}\prod_{t} {P_{njt}}^{y_{njt}}$$

where $y_{njt} = \begin{cases} 1\ if\ respondent\ n\ choose\ profile\ j \\ 0\ otherwise \end{cases}$

taking the logarithm the log-likelihood function is:

$$LL(\beta) = \sum_{n}\sum_{j}\sum_{t} y_{njt} \ln(\, P_{njt})$$

McFadden demonstrated that the log-likelihood function with these choice probabilities is globally concave in parameters β, which helps in the numerical maximization procedures. Numerous computer packages, such as Sawtooth, SAS and JMP contain routines for estimation of logit models with linear-in-parameters representative utility.

Test of hypothesis on the estimated parameters can be made with the likelihood ratio test (Wollen, 1963).

Other models

Some limitations of the Logit model gave opportunities of improvements and the rise of other similar non-linear model.

The logit model's limitations are:

- it's not possible to capture random taste variation non connected to observed characteristics
- substitution patterns are not taken into account (the Independence from Irrelevant Alternatives property holds, meaning that the relative odds of subject *n* choosing alternative *j* over *i* does not depend on the presence of other alternatives (Train, 2003))
- it's not possible to handle panel data in which unobserved factors are correlated over time

The most popular alternative models are:

1. Generalized Extreme Value (GEV)

   ± The error terms of all alternatives are jointly distributed as generalized extreme value. The most widely spread model of this family is the *nested logit* (Train, 2003). Anyway this model can't represent preference variation and panel data

2. Multinomial Probit (MNP)

   ± Assumes that the error terms $\varepsilon_{nt} = \left(\varepsilon_{n1t}, \dots, \varepsilon_{nJt}\right)^T$, where t represent the choice set, follow a multivariate normal distribution. This allows great flexibility in modeling the substitution patterns by the specification of the variance-covariance matrix of ε_{nt}; but the non-close form of the normal pdf (probability distribution function) implies computationally intensive numerical methods to calculate the probability.

3. Mixed Logit (MMNL)

± Permits heterogeneity among respondents assuming that the parameter vector β varies from customer to customer, so that $\beta_n \sim N(\beta_0, \sum_0)$. A restriction of this model is that assumes that each response is an independent realization of β_n. More commonly the experiments require repeated choices, so that slight modification of the model are necessary.

Design of a DCE

Design issues are related to the non-linearity of the MNL model, so that standard design criteria such as determinant of the information matrix are function of unknown parameter vector β.

It’s required to use locally optimal designs (Atkinson, Donev & Tobias, 2007) or Bayesian designs (Chaloner & Verdinelli, 1995).

Assuming the DCE design *d* has to be constructed with *T* choice sets, each consisting of *J* profiles, each profile defined by the level combination of *M* attributes, each having *l_i* levels; the information matrix corresponding to the model **f** is:

$$\boldsymbol{I_f}(d, \boldsymbol{\beta}) = \sum_{n=1}^{N} \sum_{t=1}^{T} \boldsymbol{X}_{nt}^{T} \, (diag[\boldsymbol{P}_{nt}] - \boldsymbol{P}_{nt}\boldsymbol{P}_{nt}^{T})\boldsymbol{X}_{nt}$$

Where $\boldsymbol{X}_{nt}$ is the *J x k* model matrix for the *i_{th}* choice set shown to respondent *n*, $\boldsymbol{X}_{nt}^{T} = (\mathbf{f}(\mathbf{x}_{n1t}), \dots, \mathbf{f}(\mathbf{x}_{n1t}))$ and $\boldsymbol{P}_{nt} = (P_{n1t}, \dots, P_{nJt})^{T}$.

Since *β* is unknown this information matrix needs an a priori estimation of ***β*** , ***β₀*** (Chernoff, 1953). A design *d** is locally D-optimal if it maximize the determinant of $\boldsymbol{I_f}(d, \boldsymbol{\beta_0})$. Under this formulation the effectiveness of the design depends on how close is the prior guess $\boldsymbol{\beta_0}$ to $\boldsymbol{\beta}$.

A common approach is $\boldsymbol{\beta_0} = \boldsymbol{0}$, so that the information matrix becomes:

$$\boldsymbol{I_f}(d, \boldsymbol{0}) = J^{-2} \sum_{n=1}^{N} \sum_{t=1}^{T} \boldsymbol{X}_t^T \, \boldsymbol{W}_J \boldsymbol{X}_t$$

Where $\boldsymbol{W}_J$ is a *JxJ* matrix with diagonal elements equal to *J-1* and off-diagonal elements -1. This method was used by many early DCE but also criticized (Huber & Zwerina, 1996).

Bayesian methods are used to reduce the dependence on a single a priori guess (Box & Lucas, 1959). A proposed criterion for nonlinear design is the following (Chaloner & Verdinelli, 1995):

$$\varphi_1(d, f) = \int log \left[det\left(\boldsymbol{I_f}(d, \boldsymbol{\beta})\right)\right] \pi(\boldsymbol{\beta}) \boldsymbol{d\beta}$$

Where $\pi(\boldsymbol{\beta})$ is the prior distribution of parameters vector.

A more frequently employed criterion is the D_B-error criterion (Sandor & Wedel, 2002):

$$\varphi_2(d, f) = \int \left[\det\left(\boldsymbol{I_f}(d, \boldsymbol{\beta})\right)\right]^{-\frac{1}{k}} \pi(\boldsymbol{\beta}) \boldsymbol{d\beta}$$

Where *k* is the number of model terms.

This criterion to be minimized is implemented in the JMP statistical software. Several optimality criteria, such as *D, A, G, V* criteria can be compared to this one (Kessels, Goos, & Vandebroek, 2006).

An alternative is the Expected Normalized Information (ENI):

$$\varphi_3(d,f) = \int \left[\det\left(\boldsymbol{I_f}(d,\boldsymbol{\beta})\right)\right]^{\frac{1}{k}} \pi(\boldsymbol{\beta})\boldsymbol{d\beta}$$

Which is to be maximized.

The very critical issues remain in the choice of prior distribution. Usually attributes are ordinal and bear monotone relationships to the probability of selection, where the direction of monotonicity is known in advance. For example for the attribute price is known a priori that the customers' utility decreases with increasing price level, so the prior distribution must be selected to reflect such information (Arora, Ginter, & Allenby, 1995).

A good and reasonable approach could be a small pilot study to obtain a prior distribution.

Then for any design optimization criteria the design can be constructed using an algorithmic approach, as the coordinate-exchange algorithm (Kessels et al., 2006; Meyer & Nachstheim, 1995).

Use of DCE for market segmentation

Segmentation is one the most established marketing technique that allows firms to better satisfy customers' needs. It has also been considered one of the main purposes for carrying out DCE (Vriens, Wedel, & Wilms, 1996; Wittink & Cattin, 1989).

Various approaches can be used for market segmentation:

1. *A priori approach*. It assigns customers into groups based on demographic variables, then a separate DCE is fitted for each identified segment (Green & Krieger, 1991; Wind, 1978)
2. *Interaction approach*. It's an extension of the previous one where the segment-level models are obtained by an estimation of interactions between model parameters and demographic variables
3. *Two-stage approach*. It runs one DCE analysis on all subjects, then cluster the respondents on the basis of estimated β-vectors (Green & Krieger, 1991)
4. *Bayesian approach*. Respondent-level estimates of β are obtained computing the posterior distribution of $[\beta|Y]$ as the product of the likelihood of the respondent's observed choice $L(Y)$, and the population level distribution of β (prior distribution). This approach requires a DCE analysis on all subjects and numerical simulation to compute the product of the likelihood and the prior distribution. Then clustering techniques are applied to identify similarities in estimated coefficients.
5. *Latent Class Method*. The market segments and regression parameters are estimated simultaneously using finite mixtures of MNL models

(Desarbo, Ramaswamy, & Cohen, 1995; Desarbo, Wedel, Vriens, & Ramaswamy, 1992)

The mixed logit model (McFadden & Train, 2000) offers a good alternative for market segmentation when subject-level information are not available and Bayesian approach is needed to determine individual level parameters estimates.

Given an estimated distribution of $\beta \sim (\beta_0, \sum_0)$, that can be collectively called θ, $\boldsymbol{X}_n$ denotes the *T* choice sets for respondent *n*, and $\boldsymbol{Y}_n = (y_{n1}, \dots, y_{nt})$ the choices made,

$$P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \boldsymbol{\beta}) = \prod_t P(y_{nt}|\boldsymbol{X}_n, \boldsymbol{\beta})$$

$$P(y_{nt}|\boldsymbol{X}_n, \beta) = \frac{e^{\mathrm{f}^T(x_{n(y_{nt})t})\boldsymbol{\beta}}}{\sum_j e^{f^T(x_{njt})\boldsymbol{\beta}}}$$

Since $\boldsymbol{\beta}$ is unknown

$$P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \boldsymbol{\theta}) = \int P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \boldsymbol{\beta})\, \boldsymbol{g}(\boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta}$$

Applying Bayes rule for the posterior distribution of $\boldsymbol{\beta}$

$$h(\boldsymbol{\beta}|\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{\theta}) = \frac{P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{\beta}|\boldsymbol{\theta})}{\int P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \boldsymbol{\beta})\, \boldsymbol{g}(\boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta}}$$

and the posterior mean of $\boldsymbol{\beta_n}$ is

$$\overline{\boldsymbol{\beta_n}} = \frac{\int \boldsymbol{\beta} P(\boldsymbol{Y_n}|\boldsymbol{X_n}, \boldsymbol{\beta})\boldsymbol{g}(\boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta}}{\int P(\boldsymbol{Y_n}|\boldsymbol{X_n}, \boldsymbol{\beta})\, \boldsymbol{g}(\boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta}}$$

This integral is computed by Monte Carlo simulation, let $\boldsymbol{\beta^r}$ be the *r*-th random draw from $\boldsymbol{g}(\boldsymbol{\beta}|\boldsymbol{\theta})$, the simulated posterior mean is

$$\widehat{\boldsymbol{\beta_n}} = \sum_{r=1}^{R} w_r \boldsymbol{\beta}^{\boldsymbol{r}}$$

Where

$$w_r = \frac{P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \boldsymbol{\beta}^r)}{\sum_r P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \boldsymbol{\beta}^{\boldsymbol{r}})}$$

Then multivariate clustering techniques such as k-means clustering can be applied on $\widehat{\boldsymbol{\beta_n}}$ (Train, 2003)**.**

Reflections about the use of response latency in DCE

Choice experiments have become prevalent as a mode of data collection in conjoint analysis. In conjoint choice experiments, respondents make choices from several sets of alternatives. These choices are analyzed with discrete choice models, which produce measurements of preferences for the attributes including the choice alternatives (Louviere and Woodworth 1983). In those discrete choice models, only the observed choices and the design of profiles and choice sets are taken into account. However, with modern computer-assisted data collection information on the time taken by the respondents to make choice decisions is readily available.

Although response latencies are collected automatically, without additional cost to the researcher, burden for the respondent, or interference with the choice task, they seem to have been overlooked in the conjoint literature. When response latencies are used to scale the covariance matrix of the MNP choice model, better estimates are obtained for the parameters of interest, which leads to better fit and predictions of holdout choices. (Haaijer et al., 2000).

Prospective research topic of interest

As a result from the presented literature review, some interesting open issues can be taken into account as prospective possibility for research:

- conjoint analysis is a really widespread and established research method in marketing and customer satisfaction studies, but, surprisingly, very little attention has so far been dedicated to a fundamental phase of such method that is the design of experiment behind the planning and analysis of a conjoint study, especially in the optimization of the experimental design;
- one of the big limitation of choice based experiments is the discrete (generally binary) response that from one hand complicates the models underlying this kind of studies, and from the other hand limits the possibility to express slightly fuzzy response which perhaps in some situations could better reflect the customer opinion;
- simplifying the survey tools in such a way that they become less annoying for the respondent and reliable for the researcher is always a challenge for researchers in any field who wish to obtain the maximum possible information from the customers, employing sometimes very completed models which however need to be translated in friendly tools to present to the subject for any kind of interview.

From all these inputs some research questions arise:

1. How the design of a DCE can be improved in terms of effects estimation in consideration of all the constraints and differences with standard DOE? How efficient is a design which is dependent from an a priori estimation of the parameters? How can we let it be robust?

2. Can we use the degree of preference (lying between 0 and 1) instead of the choice (0 or 1) to represent the decision of the respondent? How this changes the analysis? Does it improve the amount of information we can obtain? Could the response latency be the way to calculate and express this degree of preference? How much this is helpful in improving the analysis? Is it a reliable methods for these purposes?

3. Are the current methods of this kind fully taken advantage from the modern tools that the web-based marketing research can use? The computationally intense more complicated methods are perhaps now easier to implement in order to combine complex analysis with simple but precise survey tools?

A behavioral perspective

Some initial research effort has been devoted to the interesting challenge of optimizing DCEs in a combined perspective which could account for both statistical and behavioral efficiency.

The ultimate goal of an optimal DCE is to efficiently and accurately acquire information about the respondents' preferences. The correct acquisition of such information helps the researcher in understanding the respondents' decision making rules and quantifying their preferences. The objective to gain the biggest amount of information can be viewed both in terms of quantity and quality of data. In order to do so both analytical and behavioral implications of a certain design need to be taken into account when making decision about the experimental design of a DCE.

For such purpose, both statistical properties and behavioral consequences of DCEs are to be investigated. Research hypothesis and methodology for conducting future steps of this research project are briefly exposed in the last part of this chapter. This work is to be considered a work in progress. Computer simulations and behavioral experiments will be used to test some research hypothesis concerning the effects of design complexity on behavioral outcomes. Ultimately insights from this research and previous studies will be used for better inform the analytical optimization for the construction of optimal DCEs. This optimization needs to combine the maximization of statistical properties with minimization of negative behavioral outcomes; to a certain extent, the aim of taking into account both perspectives highlights the need to accept some sort of trade-off.

When a researcher wants to conduct a DCE he/she has to make a certain number of decision regarding the way the experiment will be carried out. Along with decisions such as the survey mode, the subject pool, et cetera, a key decision that will influence the final outcome of the study regards the experimental design. Note that there are a number of judgmental calls about this particular aspect of the experiment, which require a careful set of decisions to make for the researcher: How many attributes? How many and which levels for each attribute? What Experimental design? How many choice sets? How many alternatives per choice set? What profiles to compare in each alternatives?

Like in other experimental settings the ideal objective of the experimentation would be to gain the maximum information about the phenomenon under study at the minimum cost.

The objective to gain the biggest amount of information can be viewed in terms of both quantity and quality of data. In order to do so both analytical and behavioral implications of a certain design need to be taken into account when making decision about the experimental design of a DCE. Simply stated one can imagine, for instance, that the more choice sets are included in the experiment the better the information about the respondents' preferences is collected. This may be perfectly true and intuitive if the subjects are perfectly rational, certainly aware of their preferences, and insensible to behavioral consequences such as fatigue or learning. In case this phenomena occur, the quantity of information acquired may not have the desired quality, meaning that does not reflect the real preferences of the respondent, because he/she is getting tired of the survey and his/her answers start to be inconsistent with his/her real preferences.

The construction of optimal DCEs needs then to take into account both quantity and quality of data. In standard Design of Experiment (DOE) approaches, optimal designs can be created relatively easily according to specified optimization criteria. In most of the case for linear models applications of DOE efficient experimental designs are available in literature and easy to implement and also easy to analyze.

Unfortunately, as seen in the paragraphs above, there are several reasons why the optimization of a DCE is not as immediate and it involves several and correlated issues, even just from the analytical standpoint. As discussed, an active research area in optimal experimental designs is dedicated to the analytical solution of the optimization of DCEs.

Behavioral implications related to the experimental design add a further complexity to the solution of this problem.

Along with the analytical complexity of solving this problem and optimizing the experiment, additional complexity comes from the need to combine the statistical efficiency with the account of behavioral implications. Choice experiments usually involve human subjects (in contrast to other experiments in industry). For this reason, in a series of choices, the outcomes (subsequent choices) are not completely independent to each other, the subject may show fatigue, inconsistency of responses, boredom.

It seems intuitive that, for instance, is not feasible to ask to one respondent to evaluate a too large number of choice sets (characteristic of the entire design) with very similar alternatives in each choice set (characteristic of the choice set) without incurring in phenomenon such as fatigue and incoherence of responses. These implications may ultimately affect the usefulness of the final information gained with the experiment. So again, in that case the ultimate goal of the experiment is not accomplished. It seems important to guarantee the respondent’s coherence by maintaining a limited number of profiles to evaluate. This implies that the complexity of the experiment may be lower, and overall fatigue as well (note some terms used here will be better defined as variables in the subsequent sections of this paper).

If we were able to construct optimal DCEs taking into account in the analytical optimization the related behavioral implication of each design’ features, the optimal designs that we would construct were not only more suitable for the accomplish the objective of the experiment and gather the right information from the respondents, but also easier to analyze for the researcher in the analysis stage.

A research proposal builds on the following simple research questions: How can we also take into account the behavioral outcomes? Are statistical and ‘behavioral’ optimization going in the same direction? If not, what’s the trade off? Can behavioral experiments help us exploring the behavioral outcomes? Can we incorporate some behavioral parameter into our analytical model for optimization?

This research has some potential theoretical and practical contributions. First there isn’t currently any comprehensive analysis of all design’s characteristics impacts on behavioral outcomes; some effects have been studied but there isn’t a clear recommendation/list of recommendations to help the researcher to optimally design a DCE; second, a clearer identification of tradeoff between desirable properties (statistical and behavioral) is needed in order to assess when and how both statistical and practical efficiency can be accomplished; third there is a need for an analytical model for construction of optimal DCE that incorporate behavioral factors in the design stage; finally there isn’t any procedure in the analysis stage that helps mitigating or at least identifying these behavioral effects.

A huge collection of works in the marketing literature uses DCEs and conjoint analysis to study consumers’ behavior. In order to simply have an overview about the popularity of conjoint analysis in marketing research and commercial/industrial applications, please refer to Cuttin and Wittink (1982), Wittink and Cattin (1989), Green, Krieger and Wind (2001). Marketing literature very often takes not so much attention on the statistical optimization of the design, and most applications use standard DOE approaches even though non optimal for DCEs.

In the more statistical and mathematical literature on this topic there are a number of papers that focus on statistical efficiency, analytical properties of the DCE design and techniques for the analysis. This papers very often assumes rational choices, not accounting for the human subjects involved in the experiment. Very often the discussion section of such papers call for more research on the behavioral impacts of the experimental designs of a DCE. On the other hand some behavioral economists have already highlighted several behavioral issues related to these kind of experiments.

There hasn't been, to the best of our knowledge, any research addressed to jointly optimize the experiment from the analytical standpoint, which explicitly accounts for the behavioral outcomes.

There has been only limited research on how changes in the structure of the choice set (experimental design settings) change choice outcomes and the occurrence of behavioral issues. Even in these cases there isn't a subsequent optimization that takes those effects into account.

This gap in the literature persists not only because economists tend to assume that consumers are omnipotent optimizers, but also because they have not been able to measure differences in the structure of choice (DeShazo & Fermo, 2002).

In addition to what already discussed in previous paragraphs of this chapter, we should cite that, still from the analytical/statistical perspective, Huber and Zwerina (1996) identify four desirable properties for an efficient design of a discrete choice experiment. Two of these, *level balance* and *orthogonality*, also characterize linear designs. The third, *minimal overlap*, becomes relevant for choice designs, because each attribute level is only meaningful in comparison to others within a choice set. The fourth property is *utility* *balance*, which occurs when alternatives within each choice set have more equal choice probabilities. This paper advocates the use of designs constructed according to these efficiency measures, however in the discussion the authors point out that these results are relevant for ideal customers and call for further research needed on the impact of human factors on choice designs; also Zwerina & Kuhfeld (1996) in their analytical paper on the construction of efficient DCE point out the need for future research on the behavioral impact of different choice designs.

In fact in the conduction of a DCE experiment several behavioral phenomena occur and overlap while the subject goes through the experiment from one choice set to the other: on one hand learning and increasing awareness of preferences, on the other hand fatigue and incoherence/inconsistency of the answers. Some papers investigates some of these effects and the relationship with certain design's characteristics (for instance Brouwer, Dekker, Rolfe, & Windle, 2009; Savage & Waldman, 2008).

Some studies on DCEs investigate the complexity of the experimental design on the responses consistency (DeShazo & Fermo, 2002; Hensher, Stopher, & Louviere, 2001). In terms of design complexity, rational choice theory would assume that consumers are able to evaluate, compare, and rank-order the alternatives in the choice sets, and also that they can do so optimally and costless, regardless of the size and correlation structure of the information in their choice set, which increase complexity. In contrast, behavioral economists state that an increase in choice set complexity compromises choice consistency.

Simon (1955) suggested that consumers develop some sort of approximate satisficing decision rules that avoid the full cognitive cost of complexity by considering only a portion of the information available in the choice set. Thus, as complexity increased, the respondents tend to use simplifying decision rules (March, 1978). Heiner (1983) argued that increasing choice complexity would increase the gap between a respondent's cognitive ability and the cognitive demands of the decision. As this gap grows, consumers will make choices restricting the range of decision rules they consider. Although this behavior produces increasingly predictable outcomes, it is not utility maximizing. Thus, for both Simon and Heiner an increase in choice set complexity will affect the analysis of the results of the experiment because it adds noise to the error term of the utility function.

De Palma et al. (1994) consider the choice processes used by individuals with an imperfect ability to choose and predict that as choice complexity increases, so does the magnitude of sub-optimal mistakes, as manifested in lower consistency. Building on this work, Swait and Adamovitz (2001) suggest to incorporate the choice complexity into the analytical modeling of choices. This authors propose a synthetic measure of choice complexity to add in the MNL model.

This whole previous research in the different research area will be used in the following sections for the formulation of our research hypothesis, the definition of the variables that we will manipulate and the outcomes that we will measure in the simulation studies and behavioral experiments that will be exposed in Section 4.

Proposed Methodology

Given the difficulty that raises when constructing a DCE and the large number of judgmental calls that the experimenters has to consider when choosing a design of a DCE, the objective of this initial exploratory research would be to first investigate some of the issues found in the related literature. Exploring how different DCE designs perform in terms of both statistical properties and behavioral implications has the aim to better inform the optimization of such models. Following steps of this research will use this findings and insights for the implementation of a numerical optimization of DCEs that incorporates some parameters related to the behavioral outputs in the analytical formulation of the model.

First a computer simulation and then an experimental study should be conducted in order to assess how different DCEs (different in terms of design, number of choice sets, attribute settings, similarity of profiles) compare in terms of efficiency (which is related to statistical properties) and learning, fatigue and coherence (behavioral consequences of the human decision making process). Controlled experiments can potentially help in explaining the effect of design choices in the effectiveness of the analysis and optimality of the design.

Variables under study

Our objective is to study statistical properties and behavioral implications of a design for DCEs. In order to do so we need to define this properties and the variables we are going to measure. For this purpose we will define some variables representing the design's complexity. Statistical efficiency of the design can be measured as D-efficiency, a measure related to the D-optimality of the design. Finally the outputs that we are interested in are: the actual choices, parameter vectors estimations (beta vectors), and the variables describing behavioral phenomena involved in conducting a DCE, such as fatigue, learning, consistency.

- Design Complexity

The variables representing the design complexity are variable that we can measure in a given design taken from the literature. Those are also variables that we could ultimately manipulate in the subsequent step of the numerical optimization and construction of new designs.

Unfortunately, there is no generally accepted definition of ‘‘complex’’ and there is little empirical evidence available to guide researchers wishing to design experiments but also not wanting to design overly complex experiments (Hensher et al., 2001).

Complexity can be defined as a mix of subcomponents. The first and most intuitive of these components is the total number of choice sets.

Hensher et al. (2001) investigate the effects of different numbers of choice sets (4, 8, 12, 24 and 32) on response variability and model parameters in designed choice experiments. They find that fewer choice sets produce very similar mean elasticity estimates as more choice sets. This is encouraging, because it suggests that the empirical gains from more choice sets are marginal, at least for commonly used applications. These authors’ work provides interesting insights for the purpose of this paper, even though the designs that they compare are simply orthogonal fractional factorial designs for linear models, they have not used optimal DCEs.

DeShazo and Fermo (2002) provide the most comprehensive analysis of the effects of task complexity on responses to and outcomes of designed choice experiments. The authors vary the complexity of choice sets to evaluate its impact on choice consistency. For doing so they first define five measures of complexity that capture either the amount of information or the correlational structure of information in a choice set; then they investigate on the variation in these measures to changes in the distribution of the error term by parameterizing the scale factor of a heteroskedastic random utility model, which is considered a proxy of the consistency.

These measures of complexity are grouped in two categories:

- Variables related to the amount of information:

1. Number of alternatives per choice set,
2. Number of attributes per each alternative,

- Variables related to correlational structure of information:

3. Number of attributes whose levels differ across alternatives (NADA),
4. Mean standard deviation of attribute levels within each alternative (S.D. of Attributes), computed as:

$$SD_j = \sqrt{\left[\sum_{i=1}^{M}\left(x_{ij} - \overline{x_j}\right)^2\right]/J},$$

so that:

$$Average\ SD_k = \left[\sum_{j=1}^{J} SD_j\right]/J,$$

5. Dispersion of the S.D. of attribute levels across alternatives (Dispersion of the S.D):

$$Dispersion\ SD_j = \sqrt{\left[\sum_{j=1}^{J}(SD_j - average\ SD_k\ )^2\right]/J}$$

DeShazo & Fermo (2002) find that all these measures of choice set complexity affect choice consistency, and in particular that changes in the correlational structure of information have the largest impact on choice consistency; finally their work shows that choice complexity significantly distorts utility estimates.

Another interesting and synthetic variable that represent the design complexity can be found in Swait and Adamovitz (2001). They call it *Entropy of Choice*, it refers to each choice set *t* and they define it as:

$$H(X_t) = -\sum_{j=1}^{J} p_{jt} * \log(p_{jt})$$

H(*X*) is nonnegative for all values of its arguments. In a case with *J* alternatives in a choice set, entropy reaches its maximum if each of the *J* are equally likely. If the number of equally likely alternatives increases, entropy also increases, at all levels of choice probability. Entropy is minimized if there is one dominant alternative in the choice set.

- **Learning, fatigue and consistency**

DeShazo & Fermo (2002) utilize as observable proxy for choice consistency the variance of the conditional distribution of the random error term. They explain the systematic portion of the error terms by parameterizing the scale factor in a standard random utility model. By viewing this dispersion measure as their implicit dependent variable the authors are able to characterize how the *relative noise* contained in consumers' actual choices varies as the characteristics of the choice set vary, *ceteris paribus*.

Savage & Waldman (2008) investigates the effect of survey mode on respondent learning and fatigue during repeated choice experiments. Objective of their analysis is to compare different survey modes, mail and online surveys. Their way to measure learning and/or fatigue can be used even in other experiments that aim to test for instance the effect of different designs on these behavioral outcomes. These authors consider the ratio of the variance of the error components in the first half of the choice experiment over the variance of the error components in the second half of the experiment. If this ratio is approximately 1 the subjects are not showing any learning or fatigue. If the ratio is bigger than 1, subjects are becoming more proficient at the choice task as they move through more question occasions, the quality of the data improves. In other words respondents are showing learning towards the experiment. Contrary if the ratio is smaller than 1, subjects are becoming tired or bored as they move through the repeated choice questions, the quality of the data deteriorates.

Research Hypothesis

The research questions that we aim to address in this research project are quite complex. During the conduction of following steps the research hypothesis may be refined and extended. For the purpose of this paper this section will simply delineate some initial hypothesis and suggest the following steps to carry out in the next months.

Starting from what has been found in literature we can develop a set of research hypothesis on the relationships of the design's characteristics and resulting properties and behavioral implications.

In order to meet the objective of this paper we would need to carefully and comprehensively analyze how different designs characteristics impact on behavioral outcomes.

A first test would be needed to explore simpler hypothesis, confirming some intuitive insights or available findings. Also, a comprehensive analysis of distinctive effects may induce new insights when testing for interactions. Then tradeoffs and thresholds need to be identified for less intuitive hypothesis.

H1: As the number of alternatives in the choice sets increases the respondents experience more cognitive complexity and fatigue increases.

In other words the more alternatives have to be compared the more difficult the task is.

H2: As the correlation between alternatives in the same choice set increases the respondents experience more cognitive complexity and fatigue increases (and time).

In other words the more similar (in terms of probability) are the alternatives, the more difficult is to choose. To test this hypothesis we could use the entropy of choice as complexity measure.

This latter hypothesis also relates to another related topic that, for the purpose of more clarity and simplicity of this paper, has not been introduced so far: the response time. It is currently under discussion if and how we should incorporate the response time in this entire argumentation. Some interesting works on this topic are Barone et al. (2007, 2012).

In respect to the entire experimental design a research hypothesis that can be conjectured relates to the total number of choice sets.

H3: The effect of number of choice sets is U shaped.

H3a: As the number of choice sets increases, the initial effect is learning, and it has a positive effect, the quality of data improves.

H3b: After a certain threshold of number of choice sets, fatigue increases and the quality of data decreases.

In other words, first the respondent becomes more aware of his/her preferences and learns in the repeated choice tasks. But then he/she starts to get bored and the responses are less consistent and quality of the data deteriorates.

We can test the performance of two different DCE designs through the use of simulation tools and behavioural laboratory experiments. A first exploratory approach may be to compare two DCE designs available in literature.

It’s necessary to pick two designs that are different in terms of complexity, in particular, following some measures suggested by Swait and Adamovitz (2001) and Hensher et al. (2001) we can find designs which are different in terms of:

1. Number of alternatives per choice set
2. Number of choice sets
3. Entropy of choice sets

Through simulation experiments we can assess the D-efficiency of the designs and robustness from parameters prior.

Running an experiment in the behavioral lab we can confirm and investigate the actual performance of the designs we are comparing. Moreover a behavioral experiment would allow us to test the behavioral implications associated with each design under study, so that we could infer what characteristics of the design better avoid or at least attenuate fatigue and inconsistency.

Dependent variables to measure during the experiment, or to compute in the analysis stage, could be:

- Choices
- Time (in case recorded during the experiment)
    - Total time taken for the survey
    - Time response patterns
- Fatigue/learning (as defined in previous research in literature)
- Consistency (as defined in previous research in literature)

We will ultimately evaluate the designs under study in terms of all the properties and behavioral consequences in order to test our research hypothesis and use the knowledge gained to better inform the construction of an optimal DCE design.

Subsequent steps, as said above, would be to use the knowledge acquired in this first step to incorporate behavioral implications effects in the analytical model in order to numerically construct new DCEs and next evaluate them in the same manner.

Studying the relationship between design characteristics and choice's behavioral implications we can use this knowledge to inform the construction of optimal DCEs, which ideally would meet both statistical and practical efficiency.

We could be able to better identify methods for mitigating negative behavioral effects:

- at the source, in the design stage; optimizing the design taking somehow these effects into account

- in the analysis; attempting to identify these effects and de-bias the data

DeShazo and Fermo (2002) for instance show that the negative impacts of design choices may be mitigated if precautions are taken at the design and estimation stages of stated preference methods: first, minimizing the complexity of choice sets at the survey design stage by choosing the optimal number of alternatives and carefully selecting attributes and correlation structures; second, at the estimation stage, economists can identify, parameterize, and properly control for complexity using a heteroskedastic logit model to mitigate the impacts on estimates.

We want to accomplish an even more ambitious objective to numerically construct new designs that are accounting for the behavioral impacts.

This research could find its theoretical and practical contribution, reconciling the need for statistical efficiency with the accounting of behavioral impacts of the experimental design. We may find that taking into account both kind of needs in designing the experiments, and accepting some sort of trade off, can make these kind of experiments more feasible and pleasant for the respondents and (more important) more useful for the researchers.

References

- Atkinson A.C., Donev A., Tobias R., (2007). Optimum Experimental Designs with SAS. Oxford University Press.

- Barone, S., Lombardo, A., & Tarantino, P. (2007). A Weighted Logistic Regression for Conjoint Analysis and Kansei Engineering. Quality and Reliability Engineering International.
- Barone, S., Lombardo, A., & Tarantino, P. (2012). A heuristic method for estimating attribute importance by measuring choice time in a ranking task. Risk and Decision Analysis, 3, 225-237.
- Batsell, R. R., & Louviere, J. J. (1991). Experimental analysis of choice. Marketing Letters, 2(3), 199-214.
- Box, G. E. P., & Lucas, H. L. (1959). Design of Experiments in Non-Linear Situations. Biometrika, 46(1), 77-90.
- Brouwer, R., Dekker, T., Rolfe, J., & Windle, J. (2009). Choice Certainty and Consistency in Repeated Choice Experiments. *Environmental and Resource Economics*, 46(1), 93-109.
- Cattin, P., & Wittink, D. R. (1982). Commercial use of conjoint analysis: A survey. *The Journal of Marketing*, 44-53.
- Chaloner, K., & Verdinelli, I. (1995). Bayesian Experimental Design: A Review. Statistical Science, 10(3), 273-304.
- Chan, L.-K., & Wu, M.-L. (2002). Quality function deployment: A literature review. European Journal of Operational Research, 143, 463-497.
- Chernoff, H. (1953). Locally Optimal Design for estimating parameters. Ann. Statist., 24, 586-602.

- Chrzan K., Orme B. (2000). An overview and comparison of design strategies for choice-based conjoint analysis. *Proceedings of the Sawtooth Software Conference*
- Colby, D. S. (2003, May). Using marketing research: Views from a CFO. *Quirk's Marketing Research Review*. 44-47
- De Palma, A., Myers, G. M., & Papageorgiou, Y. Y. (1994). Rational Choice under an Imperfect Ability to Choose. *The American Economic Review*, 84(3), 419-440.
- Desarbo, W. S., Ramaswamy, V., & Cohen, S. H. (1995). Market segmentation with choice-based conjoint analysis. Marketing Letters, 6(2), 137-147.
- Desarbo, W. S., Wedel, M., Vriens, M., & Ramaswamy, V. (1992). Latent class metric conjoint analysis. Marketing Letters, 3(3), 273-288.
- DeShazo, J. R., & Fermo, G. (2002). Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. *Journal of Environmental Economics and Management*, 44(1), 123-143.
- Fischer, G. W., Luce, M. F., Jia, J., Frances, M., Jianmin, L., & Carolina, N. (2000). Attribute Conflict Time and on Preference Judgment Uncertainty : Effects Error. Management Science, 46(1), 88-103.

- Green, P. E. (1974). On the Design of Choice Experiments Involving Multifactor Alternatives. Journal of Consumer Research, 1(2), 61-68.
- Green, P. E., & Krieger, A. M. (1991). Segmenting Markets with Conjoint Analysis. Journal of Marketing, 55(4), 20-31.
- Green, P. E., & Rao, V. R. (1971). Conjoint Measurement for Data Quantifying Judgmental. Journal of Marketing Research, 8(3), 355-363.
- Griffin, A., & Hauser, J. R. (1993). The Voice of the Customer. Marketing Science, 12(1), 1-27.
- Gustafsson, A., Ekdahl, F., & Bergman, B. (1999). Conjoint Analysis: a useful tool in the design process. Total Quality Management, 10(3), 327-343.
- Haaijer, R., Kamakura, W., & Wedel, M. (2000). Response Latencies in the Analysis of Conjoint Choice Experiments. Journal of Marketing Research, 37(3), 376-382.
- Harry, M., Schroeder, R. (2001) Six Sigma. The Breakthrough Management Strategy Revolutionizing The World's Top Corporations.
- Heiner, R. A. (1983). The Origin of Predictable Behavior. *The American Economic Review*, 73(4), 560-595.
- Hensher, D. A., Stopher, P. R., & Louviere, J. J. (2001). An exploratory analysis of the effect of numbers of choice sets in

designed choice experiments : an airline choice application. *Journal of Air Transport Management*, 7(2001), 373-379.

- Hogg, A. (2001). Conducting online research. *American Marketing Association*.
- Huber, J., & Zwerina, K. (1996). The Importance of Utility Balance in Efficient Choice Designs. Journal of Marketing Research, 33(3), 307-317.
- Katz, G. M. (2004). PRACTITIONER NOTE: A Response to Pullman et al.'s (2002) Comparison of Quality Function Deployment versus Conjoint Analysis. Journal of Product Innovation Management, 21(1), 61-63.
- Kazemzadeh, R. B., Behzadian, M., Aghdasi, M., & Albadvi, A. (2008). Integration of marketing research techniques into house of quality and product family design. The International Journal of Advanced Manufacturing Technology, 41(9-10), 1019-1033.
- Kessels, R., Goos, P., & Vandebroek, M. (2006). A Comparison of Criteria to Design Efficient Choice Experiments. Journal of Marketing Research, 43(3), 409-419.
- Kim S. (2000). Reduce closed store passive suppression unsellable at American Express. *ASQ's 54th Annual Quality Congress Proceedings*, May 8-10, Indiana Convention center and RCA Dome, Indianapolis, Indiana. pp. 225-226.

- Kruskal, J. B. (1964). Nonmetric Multidimensional Scaling: a numerical method. Psychometrika, 29(2), 115-129.
- Kutner M. H., Nachtsheim C. J., Neter J., Li W., (2005). Applied Linear Statistical Models. McGraw-Hill.
- Li, W., Nachtsheim, C. J., Wang, K. E., & Reul, R. (2013). Conjoint Analysis and Discrete Choice Experiments for Quality Improvement. *Journal of Quality Technology*, 45(1), 74-99.
- Luce, D. R. (1964). Simultaneous Conjoint Measurement : A New Type of Fundamental Measurement. Journal of Mathematical Psycology, 1, 1-27.
- March, J. G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *The Bell Journal of Economics*, 9(2), 587-608.
- McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior (p. chapter 4).
- McFadden, D., & Train, K. E. (2000). Mixed MNL Models for Discrete Response. Journal of Applied Econometrics, 15, 447-470.
- Meer D. (2011), A new way to gain customer insights. Strategy and Business, Booz & Company
- Meyer R. K., Nachstheim C. J., (1995). The Coordinate-Exchange Algorithm for Constructing Exact OptimalExperimental Designs. *Technometrics 37*, 60-69

- Otter, T., Allenby, G. M., & Zandt, T. V. A. N. (2008). An Integrated Model of Discrete Choice and Response Time. Journal of Marketing Research, 45(October), 593-607.
- Pande, P. S., Neuman, R. R., and Cavanagh, R. R. (2000) The Six-sigma Way: How GE, Motorola, and other top companies are honing their performance. London. McGraw-Hill.
- Pullman, M. e., Moore, W. L., & Wardell, D. g. (2002). A comparison of quality function deployment and conjoint analysis in new product design. Product Innovation Management, 19, 354-364.
- Rylander, D. H., & Provost, T. (2006). Improving the Odds : Combining Six Sigma and Online Market Research for Better Customer Service. SAM advanced Management Journal, 13-20.
- Sandor, Z., & Wedel, M. (2002). Profile Construction in Experimental Choice Designs for Mixed Logit Models. Marketing Science, 21(4), 455-475.
- Sang, K. (2003). Qualitative, quantitative methods combine for best online research.
- Savage, S. J., & Waldman, D. M. (2008). Learning And Fatigue During Choice Experiments : A Comparison Of Online And Mail Survey Modes. *Journal of Applied Econometrics*, 371(February), 351-371.
- Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1), 99-118.

- Smith, J. E., & Winterfeldt, D. V. (2004). Decision Analysis in Management Science. Management Science, 50(5), 561-574.
- Swait, J., & Adamowicz, W. (2001). Choice environment, market complexity, and consumer behavior: a theoretical and empirical approach for incorporating decision complexity into models of consumer choice. *Organizational Behavior and Human Decision Processes*, 86(2), 141-167.
- Train, K. E. (2003). Discrete Choice Methods with Simulation. Cambridge: Cambridge University Press.
- Vriens, M., Wedel, M., & Wilms, T. (1996). Metric Conjoint Segmentation Methods: A Monte Carlo Comparison. Journal of Marketing Research, 33(1), 73-85.
- Wind, Y. (1978). Issues and Advances in Segmentation Research. Journal of Marketing Research, 15(3), 317-337.
- Wittink, D. R., & Cattin, P. (1989). Commercial Use of Conjoint Analysis : An Update. Journal of Marketing, 53(3), 91-96.
- Wollen, K. A. (1963). Relationships between choice time and frequency during discrimination training and generalization tests. Journal of Experimental Psycology, 66(5), 474-484.
- Woodall, T. (2001). Six Sigma and Service Quality : Christian Gronroos Revisited. Journal of Marketing Management, 17, 595-607.

- Zwerina, K., & Kuhfeld, W. F. (1996). A General Method for Constructing Efficient Choice Designs. Durham, NC: Fuqua School of Business, Duke University, (September).

Conclusions

All research studies presented and discussed in this document are the results of the three years doctoral program spent between the University of Palermo and the Carlson School of Management.

Very simply, our work in these years and our current effort to continue working on high quality research projects addressing relevant problems, can be summarized in the following way:

WHAT we have tried to accomplish: to advance the methodology of Six Sigma by extending and adapting its applicability to different instances of applications, especially tailored for cases of limited resources availability

HOW we did so: investigating critical issues in the field with case study research methodology and developing and refining both managerial and statistical tools

WHY we did this all: is to show how it is possible to efficiently and effectively apply the methodology and its advanced tools in various situations, in different ways, with different tools and approaches.

WHO, WHEN, and WHERE our work is applicable: we humbly hope to have contributed both within and beyond the Six Sigma framework.

In viewing this work as a whole, our objective has been both academic and practitioners oriented. From an academic standpoint we aimed to contribute to different research streams involving Six Sigma as methodology, and its philosophy and its tools in a broader sense. In this perspective Six Sigma is a topic of interest for both top academic and practitioners' journals in operations management and applied statistics.

In our initial work, we had the empirical approach of the case study methodology, together with some review of practitioners' journals. Comparing several experiences in the implementation of Six Sigma in small and medium companies, both in Italy and abroad, we had the opportunity do get closer to the phenomenon we wanted to investigate, and to understand better how we could contribute to the literature and especially to practical applications with further developments of our work.

To do so, we turned completely our method into specific studies addressing one or more managerial and/or statistical tools of the Six Sigma toolbox. Small and medium enterprises, and, in a broader sense all instances with resources availability constraints, can have incredible financial and operational gains from applying such methodology and the correct statistical thinking. It is not needed to change the methodology, nor either the approach, in order to tailor Six Sigma to such kinds of environments. Even with structural and infrastructural constraints, what is really needed is a broader and more flexible set of tools, which could allow for a more efficient and effective use of the available resources and the same relative, or even bigger gains.

Moreover, we worked on decision making and/or statistical tools that could advance the discipline both theoretically and practically, not just within the framework of Six Sigma in quality improvement, but also in explorative innovation activities and broadly in other applications of such tools in operations management, marketing, decision sciences.

Our work that combines the response latency model in the AHP procedure, was a way to leverage on existing and emerging analytical models and methods in order to accomplish the same task in a faster, more intuitive, reliable, and more efficient way than before.

In the same spirit our entire research agenda on experimental design, all our work on design planning and analysis, both for linear and non-linear models, can be viewed as an efficiency oriented approach that aims to gain more information at a smallest cost and in a more reliable and robust way.

We have worked with all these concepts in the background of our different works partly described here, and we are strongly oriented to pursue this approach in continuing our work, in completing what is currently in progress and in embracing new projects in the near future.

Appendix

Case study: a Six Sigma project at Structo Hydraulics

The company and general description of the project

Structo Hydraulics AB is a Swedish manufacturing company producing steel tubes mainly for hydraulic applications. Its production facility is located in Storfors, in the Swedish county of Värmland. Structo Hydraulics is among the European leading suppliers of tubes for the hydraulic industry and it has more than 400 years' experience in iron and steel processing. The product portfolio includes cold-drawn seamless tubes, cold-drawn welded tubes, roller-burnished cylinder tubes, cold-formed tubes and components. The product variety covers a wide and comprehensive range of geometrical dimensions. Today, Structo Hydraulics counts around 100 employees and a production capacity of 30,000 tons of cold-drawn tubes, 18,000 of which are skived and roller-burnished, and 6,000 are components. The company has an annual turnover of 450 million SEK (approximately 70 million dollars) and covers a wide market in Europe. It is owned by ISMT Ltd., a leading Indian global firm that supplies precision seamless tubes and steels to the bearing, automotive, mining, OCTG (Oil Country Tubular Goods) and energy industries worldwide.

According to the European definition of SME (European Commission 2003), Structo Hydraulics (henceforth only Structo for simplicity) is a medium-sized enterprise.

Nowadays suppliers have to face many challenges imposed by customers. Customers want to buy products and services at a price viewed as good value for money (Malliga & Srinivasan 2007). The interest in Six Sigma at Structo derived by one of its biggest customers. This is common in many medium-sized companies, especially those operating at the highest levels of the supply chain for a certain industry. A large company with consolidated Six Sigma infrastructure and years of experience in Six Sigma often pushes its suppliers to use the same approach and sometimes offers Green/Black Belt training. The drawback in this widespread approach is that such training programs are often expensive and highly company-oriented.

During the last decade, Chalmers University of Technology in Gothenburg, Sweden has established a Six Sigma Black Belt education within the Master Program in Quality and Operations Management. The Black Belt course involves both master students and industrial participants, grouped in small teams to carry out Black Belt projects while attending the course. Black Belt education at academic level is a good formula. Rao and Girija Rao (2007) have suggested introducing the subject of Six Sigma as a full subject at the Academic level. They believe that all students of management should leave the University as certified Black Belts.

Taking the opportunity offered by Chalmers University of Technology, Structo decided to take part to the Black Belt course in 2011. The Managing Director of Structo, in agreement with the Quality Manager, selected a suitable project idea and a Quality Engineer to attend the course. She was selected because of her role and her skills were the right prerequisites for a Black Belt candidate. The choice of the right people to educate in Six Sigma is of crucial importance for an organization that wants to successfully implement the methodology (Black& McGlashan 2006).

The project idea was to reduce waste in the cutting processes of the warehouse department. The company had already dedicated previous efforts to solve this issue, without significant effects. The management recognized that Six Sigma could help solve this big issue. According to the standard practice of the Black Belt course at Chalmers, the team was composed by two master students and the company's Quality Engineer. During the project, there was continuous interaction with company Managers to get feedback and advice, and the team involved also other people in the company, directly or indirectly related to the project; as among them the executive managing director, the process owner, the production manager, the quality manager, the production leader of the warehouse, and some operators.

It was soon clear that the company was seriously committed to devote some resources (human and financial) to the implementation of the Six Sigma methodology. The attractiveness of the potential savings related to successful Six Sigma projects is definitely a further incentive for the management commitment to these initiatives, along with the pressure from the suppliers. In approaching Six Sigma, Structo's management was pursuing the objective of gaining benefit out of the methodology, although aware of the limit of not being capable to build a solid infrastructure, at least initially.

The DEFINE phase

As mentioned above, the project idea was selected before the course start, by the Managing Director together with the Quality Manager. Although many previous efforts to reduce the yield losses in the warehouse (this is a very common problem in Six Sigma projects, see for instance similar issues in Sarkar, 2007), none of the past projects had been completely successful, and scrap remained a major issue. **Error! Reference source not found.**.1 illustrates the different causes of scrap for the company and their relative percentage contribution to the total amount.

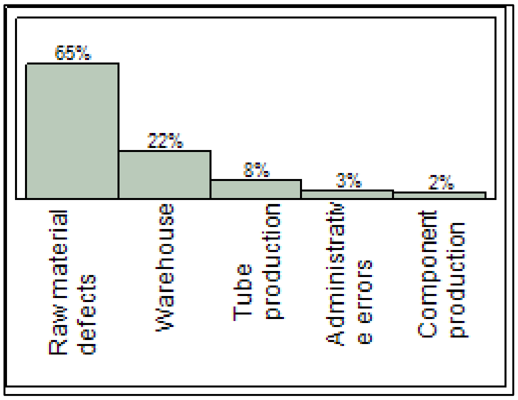
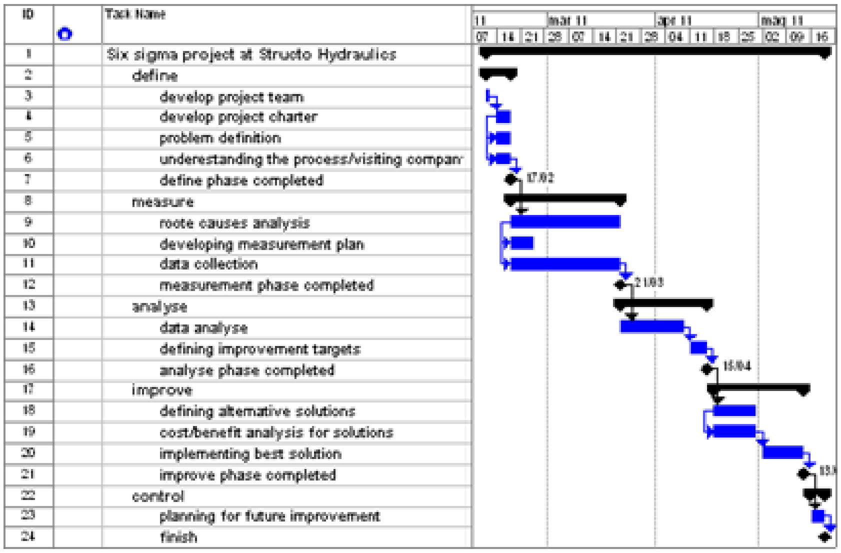


Figure 1.1: Sources of scrap

The sources of scrap could be grouped into five main categories: raw material defects, which means scrap due to non-compliance of suppliers' materials with required specifications; scrap in warehouse, meaning the remaining material in cutting processes; scrap in tube production, because of defects caused in the production of cold-drawn tubes; scrap in component production, because of defects caused in the components department; scrap due to administrative errors.

Raw material defects could be due to several causes. There was a dedicated database for this type of data. The main supplier of Structo was also the owner of the company and, as mother-company, it had several ongoing projects aiming to reduce the scrap rate of the raw material supplied to Structo. For this reason, the focus of this project was the second-largest source of waste, the warehouse activities. Waste arises when the last part of a long tube, cut according to customer's specifications, remains after the cutting process and it is not long enough to be used to fulfill other orders. This is pure waste; the material complies with quality standards, but not with the length requirements of the customers. It's important to notice here that, as previously discussed, the size of potential gains from a Six Sigma project can be even more substantial for a SME than for a large company. Simply consider that the waste in the warehouse is the largest issue of internally produced scrap and amounts to 2.6% of the company's annual turnover.

As one the most important initial activities in the define phase, the project charter was written at the project start, as the main official document for the Six Sigma project (hereinafter bold font will highlight the Six Sigma techniques used in the project). The **project charter** framed the project, its main scope and goals, the potential savings, the team and all involved people, the milestones for each phase and a detailed project plan. The project was planned to be completed in five months, i.e. the time horizon of the Six Sigma course. The DMAIC phases and milestones were planned as shown in the **Gantt chart** of **Error! Reference source not found.**. However, due to the limited time span and the ambitious goal, it was clear since the beginning that the phases of improve and control could not be completed by the end of the project, but they were going to have a follow-up plan.



ID	○	Task Name
1		Six sigma project at Structo Hydraulics
2		define
3		develop project team
4		develop project charter
5		problem definition
6		underestanding the process/visiting compan
7		define phase completed
8		measure
9		roote causes analysis
10		developing measurement plan
11		data collection
12		measurement phase completed
13		analyse
14		data analyse
15		defining improvement targets
16		analyse phase completed
17		improve
18		defining alternative solutions
19		cost/benefit analysis for solutions
20		implementing best solution
21		improve phase completed
22		control
23		planning for future improvement
24		finish

Figure 1.2: Gantt chart of the project plan

The main goal of the project was to increase the profitability of the warehouse department, and consequently of the company, by better utilizing materials and thereby reducing the cost. Increasing the yield by 1% – a reasonable project goal – would have represented savings of approximately 600,000 SEK (about 95.000 $)per year.

Indirectly, the project goal included a better planning for late deliveries to the final customers, when those were due to shortage of material. This issue was indirectly caused by a non-optimal cutting planning which, if completely or at least partially solved, would reduce the risk of losing orders due to late deliveries.

In order to better frame the scope of the project, note that Structo sells tubes in *random* and *cut* lengths. Random length means that a customer orders a total length, and each single tube can be delivered in a length ranging between 4 and 12 m. However, most of the large customers place orders in cut lengths, which means specific required lengths. The warehouse is the department responsible for the cutting process according to requests from customers. In this scenario, the **big Y** that the project aims to improve is the yield of the cutting processes, defined as:

$$\text{yield}= \frac{\text{output (Kg)}}{\text{input (Kg)}}\%$$

Where *output* refers to the total amount of cut tubes obtained in the process, and *input* refers to the amount of long tubes that have been cut (measured in kilos). The yearly yield for the year 2010 was 90.9%, which represented a scrap cost of about 5.3 MSEK (about 838,000 $). Due to technical limitations of the cutting machines (because a part of the long tube is always needed just to fix the tube on the cutting machine), the yearly yield cannot exceed 96÷97%. The goal of the project was set by calculating the potential savings for different levels of yield improvements, where the monetary value of the losses based on the average cost of the product (cost is related not only to the weight in kilos, but also to refinement processes). According to the previous reasoning, the goal set by Structo for this project was a yield of 92%, representing potential savings of 100,000$ per year. On a long-term basis, this goal would become more ambitious and the aimwas to further improve the yield to 94%, by developing successful improvement solutions arising from the project.

To complete the Define phase, it was necessary to fully understand the process under study and its related activities. The planning of the cutting process was usually based on a two-week time-frame (see the **process flow** of **Error! Reference source not found.**). The orders were loaded in the warehouse information system at the time the y were received from the Marketing department. The process had different lead times depending on the required final product; in some cases the required final product was already available in stock (input tubes ready to be cut) or it might need to be supplied by starting a new production order or a purchase order; in some other cases, customers' special requirements for a new product or new characteristics could require a specific product design.

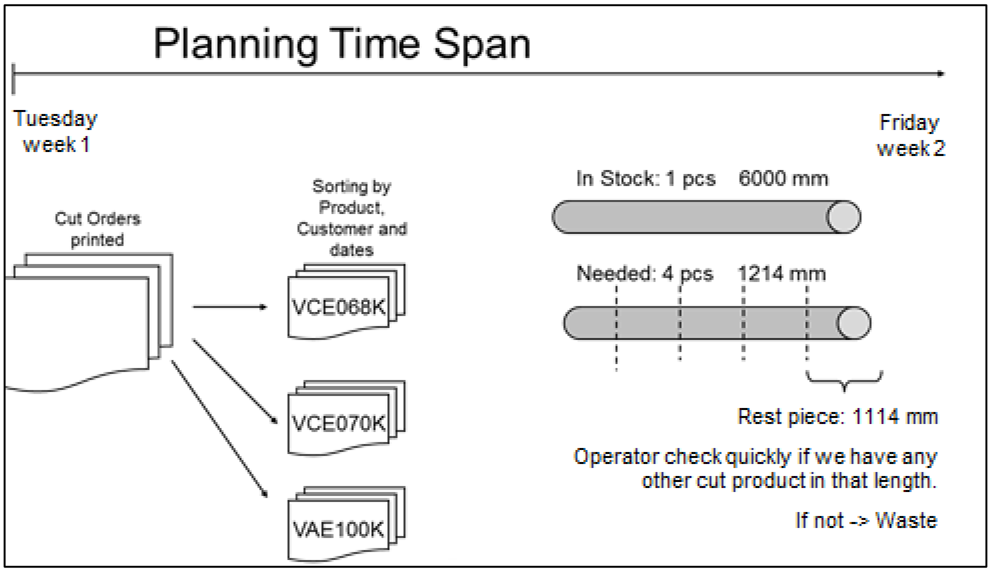


Figure 1.3: Planning for cutting - process flow

The cutting orders to execute in the warehouse were printed out and sorted by date, customer, and product (input product). The planning was not made according to a well-structured method or facilitated by software. The operators had a list with commonly sold lengths per product, and – after executing the order –with that list they used to check whether or not the remaining pieces could be suitable for smaller cut lengths. This check was done in a non-structured way.

Finally a **SIPOC diagram** (**Error! Reference source not found.**.1) was used by the project team to better understand how the process was currently working. This tool helped identifying the process' suppliers, input, output, and customers, including the requirements for input and output and the process start and end. The SIPOC was divided into two key processes – administrative planning and actual production – in order to capture both the flow of information and the actual flow of materials. The two processes often run in parallel. The two customers in the process were the internal customer (operator) and the external customer (final customer).

SUPPLIERS	INPUT	PROCESS	OUTPUT	CUSTOMERS
Marketing Check material in stock; enter the customer order in the system. If there is not enough material in stock, start the process of buying or producing tubes	Customer order (Fix cut products) Part number Length Delivery date Quantity	Delivery Date between Tues week1 and Fri week2 ↓ Planning of Cutting When and what to cut ↓	Cutting schedule	Warehouse operators
Tube suppliers/Production	Long tubes (material)	Start of cutting process	Cut products	End customers

Table 1.1: SIPOC

MEASURE phase

The second phase of the DMAIC cycle had the main purpose to deeply understand the process and to develop a **measurement plan** for collecting quantitative data.

A detailed **process map** was made with the aim of better understanding the process' inputs in each step. They were then used in a **cause-and-effect matrix** to link the inputs to the outputs of the process and to show how much they were relevant in the fulfillment of the key process output for the customer.

Then a measurement plan was designed for the data collection, it was considered a time frame covering the entire year 2010. This period was chosen because it was long enough to represent truthful data for the analysis. In fact, in shorter periods, some periodical changes in the orders might influence the data and hide the general trends. The dataset was an Excel© worksheet of 3290 orders.

The most important information collected for the analysis included: part numbers (the output part number was a code identifying one product with specific physical and process characteristics like inner and outer diameter, wall thickness, surface characteristics, etc.; the input part number identified the long tubes used to cut the final products);

customers, identified by name and code,(only customer codes will be shown for confidentiality reasons); number of pieces required (when the sales unit was meters rather than pieces, the quantity was calculated by dividing the total length by the length of each product); length, as specified by the customer; input and output weights and lengths, used to calculate the yield and the amount of waste related to the order after execution; date of the order and due date.

The dataset was mainly collected from weekly and monthly company's internal reports. It was not possible to verify the data accuracy. Therefore, the first issue to consider was the trustworthiness of the data. From a first look at the data sheet, it was clear that there were some orders with a yield higher than 100%. This evidence gave rise to four questions:

1. What was the problem when this data was recorded? What was causing the yield to be a clearly wrong number?
2. Were these the only unreliable data or this was a symptom of a more severe problem in the system, or in the way data were recorded?
3. If that was the case, how was it possible to recognize other unreliable data with yield lower than 100%? In other words, was the data inaccuracy identifiable?
4. What conclusions could be drawn from such dataset?

Since the purpose of the data analysis was to better understand the problem, the first analysis was exploratory. The team decided to proceed in the best possible way with other analysis, but the issue raised in this phase showed a non-trivial problem that needed to be addressed.

In order to successfully apply a data driven approach as the Six Sigma methodology, the quality and the reliability of the data and the information system are fundamental. This issue represents an important contribution to the discussion on the applicability of Six Sigma in SMEs that must be grounded on reliability of the data in order to assure a ‘decision-by-fact’ approach.

ANALYZE phase

An **Affinity Diagram** session was held with four operators and the Production leader of the warehouse. The main goal of such session was to ascertain their contribution in understanding the causes of waste. One of the team members attended the session and an Affinity Diagram was then created. The session started with the simple question: “What is causing waste?”. All participants wrote their ideas on notes, independently from each other. The second step was to present the ideas to the group and to discuss the ideas related to each other, then to group them in categories. Another step was to evaluate the different ideas by scoring the five most affecting causes in order to gain a better picture of which ideas to focus on in the next phases of the project. The Affinity Diagram session helped identifying the reasons behind cutting problems that the Managers had not initially mentioned, for instance unreadable handwriting and communication problems. Some findings were later considered in the Improve phase, like establishing a systematic, computer-based way of planning the cutting processes.

In order to consider all possible causes of waste in the cutting processes, a **Brainstorming** session was arranged and a **Fishbone Diagram** was created (**Error! Reference source not found.**). The causes of the problem were classified into seven categories: management, marketing, planning, machinery, stock, operators and material.

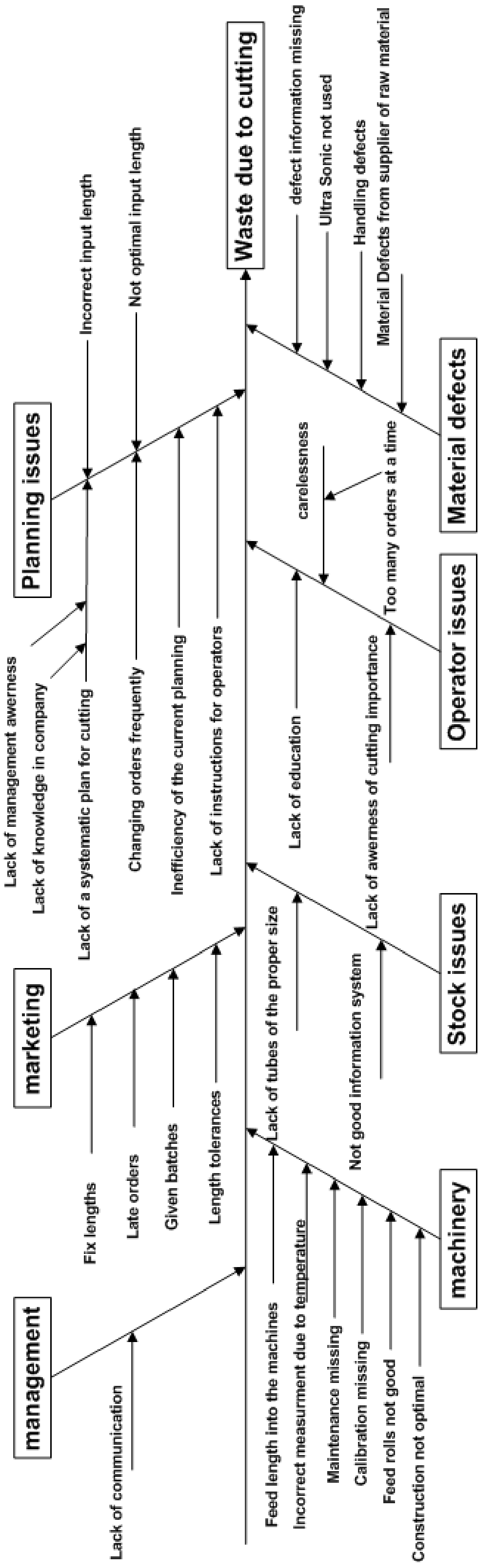

Waste due to cutting
management
Lack of communication
marketing
Fix lengths
Late orders
Given batches
Length tolerances
Planning issues
Lack of management awerness
Lack of knowledge in company
Lack of a systematic plan for cutting
Incorrect input length
Not optimal input length
Changing orders frequently
Inefficiency of the current planning
Lack of instructions for operators
machinery
Feed length into the machines
Incorrect measurment due to temperature
Maintenance missing
Calibration missing
Feed rolls not good
Construction not optimal
Stock issues
Lack of tubes of the proper size
Not good information system
Operator issues
Lack of education
Lack of awerness of cutting importance
carelessness
Too many orders at a time
Material defects
defect information missing
Ultra Sonic not used
Handling defects
Material Defects from supplier of raw material

Figure 1.4: Fishbone diagram

Another tool, the **Cause-and-Effect Matrix**, was used based on the process map and customer ranking of the process inputs. The Fishbone Diagram identified all potential causes of the problem, but did not recognize the most important and critical ones. The Cause-and-Effect Matrix was used to prioritize potential causes according to what was considered critical to customers. The Cause-and-Effect Matrix, in which the input is sorted according to its importance, showed that the first four process steps related to the most affecting input were related to: getting customer orders, planning for cutting, prioritizing the orders, and checking the availability of the tubes in stock. The critical inputs were: customer order's specifications, tolerances, and stock data available in the information system.

At this point, having a better qualitative picture of the process, a **quantitative data analysis** aimed to quantify some aspects of the problem that would help identify the potential improvements. This data analysis distinguished the products that were often sold in the same lengths and to the same customers (in other words, the products whose demand could be forecasted to some extent) from other products which were requested more rarely and most likely with particular and occasional specifications. The products sold in 2010 were identified in the data collection by a code, 216 codes in total, identifying a specific type of tube with certain characteristics and dimensions (inner and outer diameter, wall thickness). Each customer order reported the required length of the cut products.

The orders were analyzed according to different criteria, unfortunately the number of products was huge and the relative frequencies were always small. The team first searched for products which were the best sellers in terms of number of orders, then in terms of the total amount of pieces. For example, some orders required few pieces of one specific product and other orders required several hundred pieces; the number of pieces ranged from 1 to 1386. Another criterion for grouping the products was based on how many different lengths were found in the orders of the previous year. The reasoning behind this analysis was that some products were almost always required in the same length or few variants. For other products, the range of length and the number was huge. The former products were also mainly the best sellers. An optimization study for these products could be more complex but reasonably effective for the whole business of Structo. Finally, the best-selling products in a specific length and in terms of total amount of required pieces were identified.

Customers were analyzed in terms of number of orders as well (**Error! Reference source not found.**). Other analysis on customers could have considered total tubes ordered in terms of weight or the quantity of cut products in order to figure out the importance of each customer for the company. **Error! Reference source not found.** shows that there were two most important customers in terms of orders. It was found that 80% of the company's revenue was related to its top eight customers, representing around the 20% of the company's total number of customers.

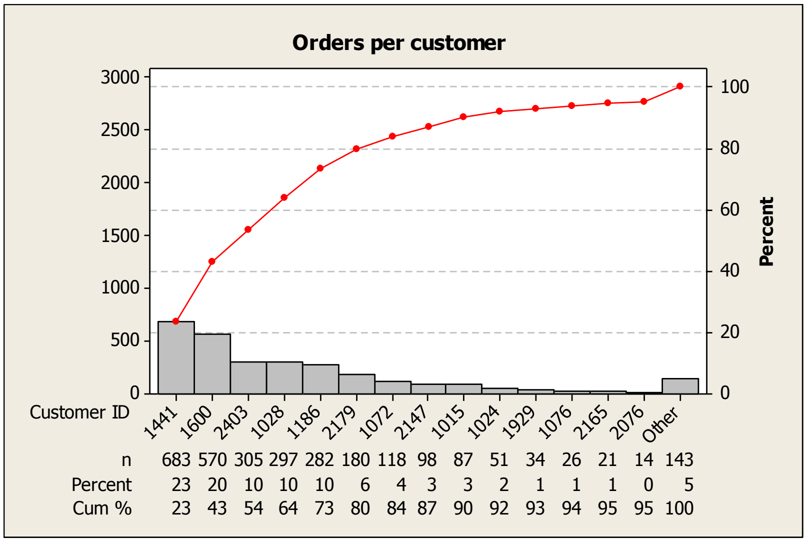
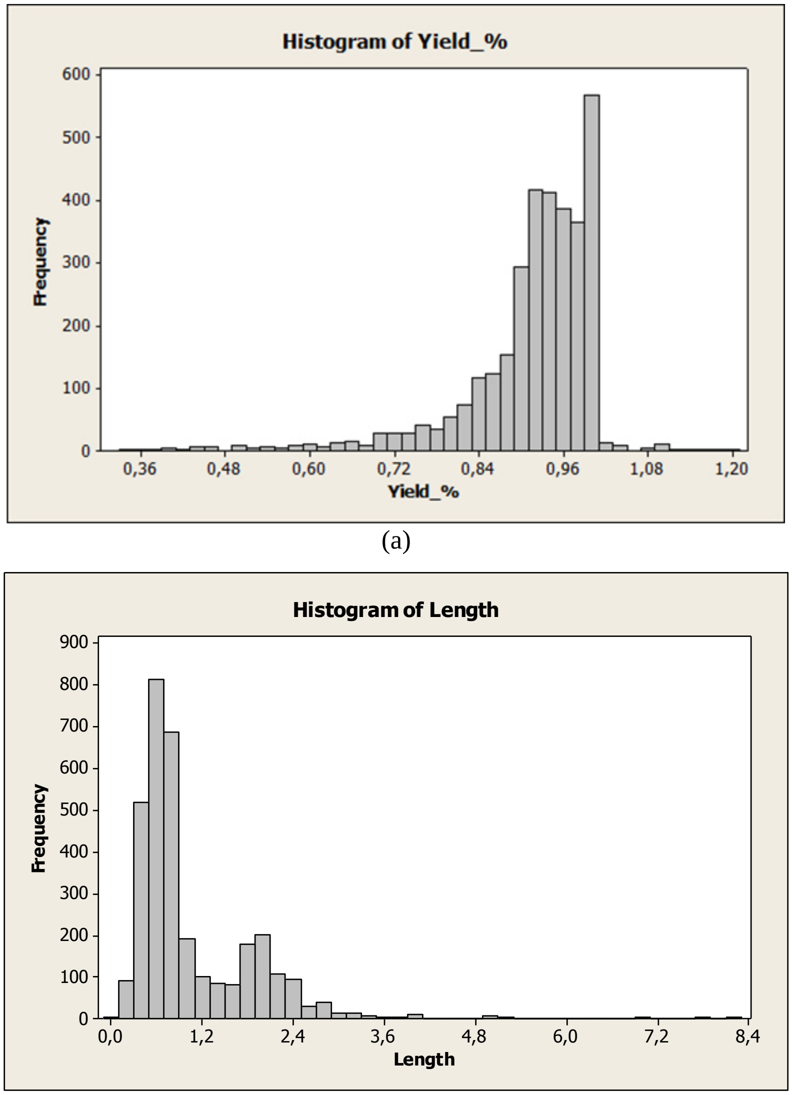


Figure 1.5: Pareto chart, customers – total number of orders

Then the attention was shifted to the bestsellers. Further analyses of these products were performed in order to study the distributions of these products' orders.

As noted before, the data set consisted of 3290 observations, each one representing a specific order. The yield (The big Y) for the whole time period provided the following statistics: average = 0.92; median = 0.94; mode = 1.00; standard deviation = 0.099; range = (0.33; 1.20).

The **histogram** of the yield data (**Error! Reference source not found.**.a) was negatively skewed as a truncated distribution (note: some data were higher than 1 as discussed above) and the histogram of sold lengths (**Error! Reference source not found.**.b) was positively skewed and bimodal.



(a)

(b)

Figure 1.6: Two aspects of the exploratory data analysis

In the Analyze phase it was very important to use a variety of tools for qualitative and quantitative analysis. On one hand, for this phase it had a crucial role the knowledge acquired during the Six Sigma training, along with the support of the Academic tutors of the course. On the other hand the perspective and contribution of the operators during the Affinity Diagram session was fundamental for a deep understanding of the problem. Some of the nuances of the process could have been highlighted with only the view of the Managers.

This remark doesn't simply acknowledge the support of all people involved in the project but also strengthen the point made in the initial discussion, that consult of academics and people in the field may make the difference, especially in initial steps into Six Sigma projects in a SME.

IMPROVE phase

The most important aims of the Improve phase of the DMAIC cycle were: find improvement solutions; test the solutions; validate and implement the best ones.

All possible solutions, discussed below, needed to be evaluated and classified. In this project, some were relatively simple to implement and to be routinely applied to the process in the future; others were radical improvement solutions. Ratings were given after reflections and discussions within the team.**Error! Reference source not found.** shows the solutions in a **Cartesian graph** on two dimensions: "Difficulty of Implementation" and "Effectiveness". This simple tool helped prioritize the improvement solutions.

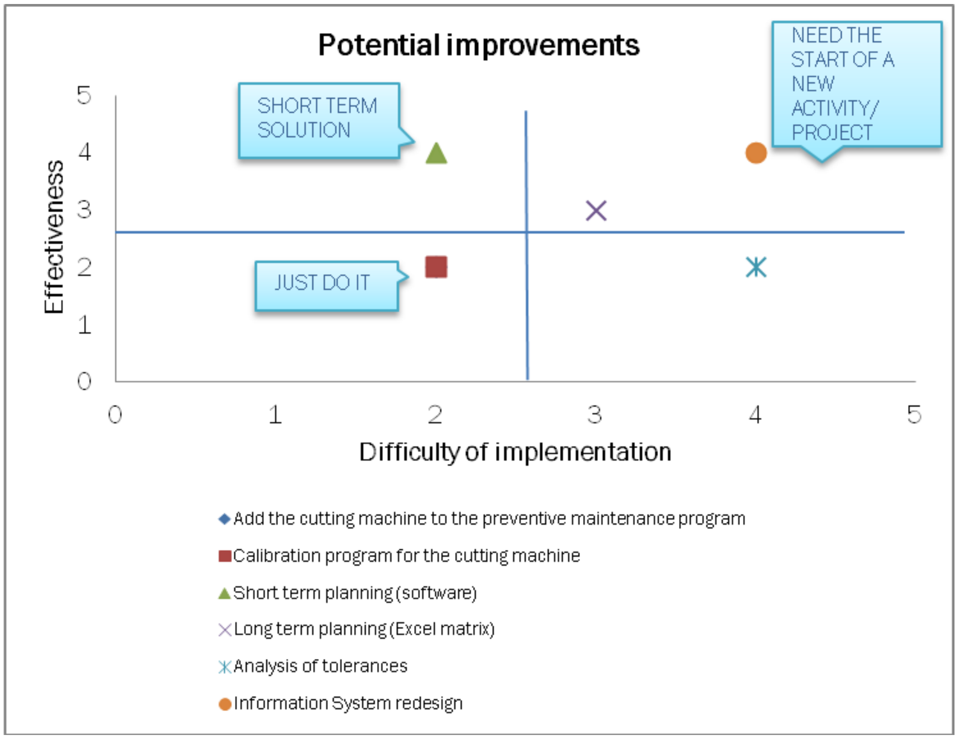


Figure 1.7: Rating of potential improvements

Improvements that needed to be done as soon as possible were those with a low difficulty and a high effectiveness (top left in the graph). Improvements with a high difficulty and lower effectiveness could be postponed. Here it is important to note that the long term versus short term perspective in prioritizing improvements is another crucial aspect in the implementation of Six Sigma in a small company. In fact, on one hand, these kind of projects very often discover urgent and critical needs for improvements that require prompt attention and resources, along with the more substantial and radical changes in the process. On the other hand, all this needs that appear to be indispensable have to face the limits of the available human and financial resources at disposal. If this argument is true in general, it becomes extremely important for an SME.

The Improve phase aimed to address the issues emerged in the previous phases and to suggest solutions. One of the problems that emerged in the Affinity session was a non-systematic plan for maintenance and calibration of the cutting machines. This resulted in problems during the cutting process and, consequently, some waste of material. If a tool does not measure correctly, a piece can be cut too short and not meet specifications. The proposed solution was to include this machine in the existing general maintenance plan and periodically check its effectiveness.

Another important issue that emerged from the Affinity Diagram concerned the tolerances. The standard cutting tolerance was +2/-0 mm. Some customers gave their own requirements of tolerances. When no requirement was given, the standard tolerance was used. When the tolerances were not met, this resulted in a waste of material. However, that material could be good for the customer, which could avoid the waste to some extent. This issue could be easily overcome by reviewing customer tolerances regarding each particular order.

An issue concerned the company’s information system. On the one hand, the data that were found to be incorrect showed that there was something wrong in the system used for registration of data or in the way it was used. On the other hand, a practical problem with the information system concerned a difference in the quantity of available details regarding the customer orders compared to the same necessary details about the available stock, which consists of long tubes produced in the company or bought from other external suppliers.

The issue was how the system showed this material in stock, regarding the input to the cutting process. The system showed only a total quantity per batch; for instance, 12 pieces with a total length of 100 m. It did not show the length of each piece. It was possible to reflect upon the implementation of an efficient solution for the biggest problem (the cutting schedule) only with complete information on the available stock and orders. A redesign of the information system was needed. This was seen as an opportunity for another Black Belt project.

The most important issue emerged during this project was the lack of a structured way of planning and cutting the long tubes in fixed lengths according to customers' specifications.

Both at strategic and operational level, the cutting plan was essentially based on the operators' experience and knowledge. Therefore, the input to the cutting plan was mostly their knowledge and experience, with guidance from the warehouse's production leader.

Two types of improvement were considered here: a short-term and a long-term option. The short-term improvement tried to answer the question of how to cut the tubes that were already in stock, and to generate the least possible waste. This improvement is short-term because it does not consider what tube length should be bought or produced according to the orders to minimize the waste, but how to cut the existing tubes to reduce the waste. The long-term improvement idea was to forecast the orders before receiving them from the customers so that it could be possible to purchase the right lengths of tubes, so reducing waste at the cutting operations.

The costs of the various solutions were split into direct and indirect costs. The direct costs are for instance purchasing software and hourly costs of educating personnel and maintaining the solutions on a yearly basis. A rough **Cost-Benefit Analysis** (see **Table 1.9**) helped the evaluation.

Table 1.9: Cost-benefit analysis

Improvement	Direct Cost	Indirect Cost	Total Cost	Benefits
Add Cutting Machines to the Preventive Maintenance Program	Maintenance Cost	Machine Downtime	€500	Less intervention to failure
Calibration Program for the cutting machines	Calibration Cost	Machine Downtime	€500	More accuracy and precision
Short Term Planning (Software no 1)	€75	€1000/year Education and Maintenance	€1075	Using the software, the goal of increasing by 1% will definitely be achieved.
Long Term Planning (Excel Matrix)	€0	Education and Maintenance	€750 €	
Analysis of tolerances	This is a suggestion for improvement activities/projects.			
Information System	This is a suggestion for improvement activities/projects.			

One of the options that could be considered for optimizing the cutting planning was to implement an optimization software, which could be easily found on the Internet. A trial version of *Cutting Optimization Pro* v.4.9.1.8 was chosen for this purpose and some preliminary runs were performed to test its applicability and effectiveness. The different aspects to consider were: cost of the software, integration of the software into the company's information system, extent of changes in the process needed for a practical implementation, ease of use and training for people who had to use it, potential increase in yield.

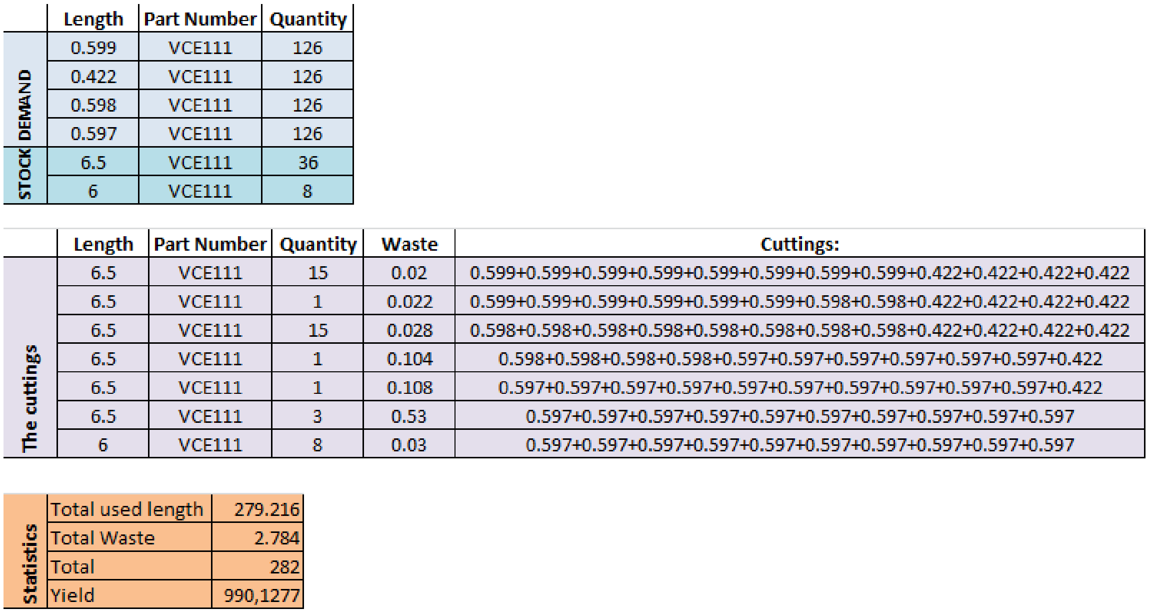
In order to test the software, data about customer orders and stock on hand were needed. For customer orders, it was possible to use both historical data from previous data collection and current orders extracted from the company database. For stock on hand, data in the system had a certain structure. When looking at the system, in order to check the feasibility to execute a customer order, the operator saw a table with the following information for each lot in stock: part number (code identifying the product) and description; location in the stock; length (total per batch); number of pieces in the lot; lot number and remarks.

The variable *length* represented the total length (in meters) of that product in that specific lot, but it was not possible to obtain information about the length of a particular tube in that lot. Obviously, the operator could physically go to the warehouse and check the length of the tubes. Sometimes in the remarks column of the database it was possible to find records on whether a specific lot had been already cut into pieces of a certain length.

Before doing **simulations**, it was considered that the data about stock on hand was continuously updated during the daily work of the company. The only chance to use this data for the simulation was to assume a picture of the stock at a certain time and use the data only for the simulation purpose.

Keeping in mind other simplifications, it was possible to run some simulations of eventually optimal cutting patterns using data extracted from the company system, including orders up to June 2011 of the customer with code 1600, stock-on-hand information about the most popular input part numbers.

Error! Reference source not found. shows the output of one simulation on the part number VCE111. It was found that the waste could be reduced and the yield increased when considering, for instance several orders of different lengths at the same time and not executing each order one at a time.



	Length	Part Number	Quantity
DEMAND	0.599	VCE111	126
	0.422	VCE111	126
	0.598	VCE111	126
	0.597	VCE111	126
STOCK	6.5	VCE111	36
	6	VCE111	8

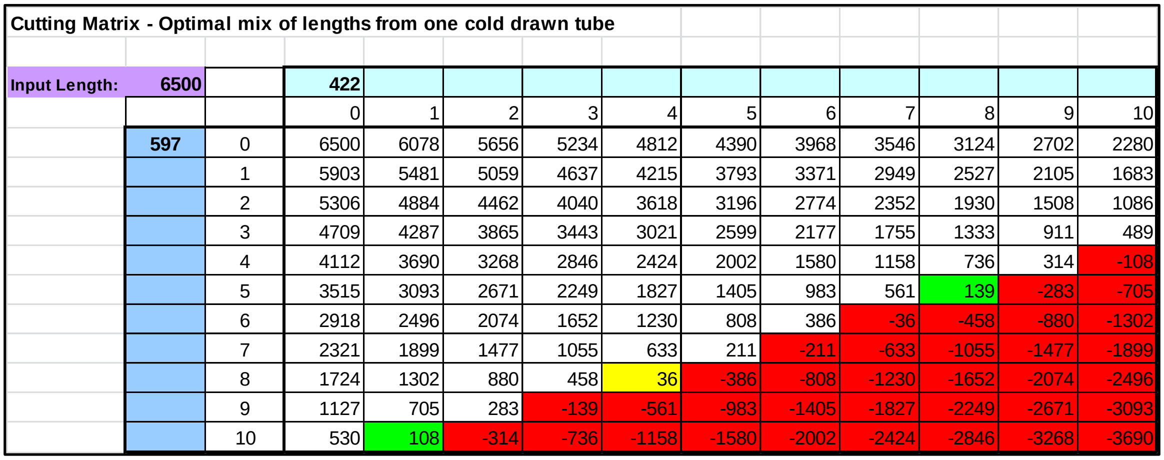
	Length	Part Number	Quantity	Waste	Cuttings:
The cuttings	6.5	VCE111	15	0.02	0.599+0.599+0.599+0.599+0.599+0.599+0.599+0.599+0.422+0.422+0.422+0.422
	6.5	VCE111	1	0.022	0.599+0.599+0.599+0.599+0.599+0.599+0.598+0.598+0.422+0.422+0.422+0.422
	6.5	VCE111	15	0.028	0.598+0.598+0.598+0.598+0.598+0.598+0.598+0.598+0.422+0.422+0.422+0.422
	6.5	VCE111	1	0.104	0.598+0.598+0.598+0.598+0.597+0.597+0.597+0.597+0.597+0.597+0.422
	6.5	VCE111	1	0.108	0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.422
	6.5	VCE111	3	0.53	0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597
	6	VCE111	8	0.03	0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597+0.597

Statistics	Total used length	279.216
	Total Waste	2.784
	Total	282
	Yield	990,1277

Figure 1. 8: Optimal pattern from the software optimization

These results needed to be verified and tested; the yield increase had to be interpreted as a hypothetical possibility of improvement using optimization tools, which showed that better cutting patterns could be found and that the waste could be reduced.

Another possible way to improve the cutting schedule was related to the use of different tools developed in a previous project at Structo Hydraulics. This could be considered as a solution that was more suitable in a long-term planning perspective. The previous project had the aim of reducing waste when considering two orders at the same time.An Excel spreadsheet was created as a tool to be used every time the customer specified two orders of different lengths for the same part number. This sheet (**Error! Reference source not found.**) was further developed during the project to make it more user-friendly. The cells of the Excel sheet gave the length of the remaining piece after cutting in all possible combinations. The colors indicated what was most effective. This tool could be easily used for long-term planning to decide the optimal input length using forecasts of orders.



Cutting Matrix - Optimal mix of lengths from one cold drawn tube													
Input Length:	6500		422										
			0	1	2	3	4	5	6	7	8	9	10
	597	0	6500	6078	5656	5234	4812	4390	3968	3546	3124	2702	2280
		1	5903	5481	5059	4637	4215	3793	3371	2949	2527	2105	1683
		2	5306	4884	4462	4040	3618	3196	2774	2352	1930	1508	1086
		3	4709	4287	3865	3443	3021	2599	2177	1755	1333	911	489
		4	4112	3690	3268	2846	2424	2002	1580	1158	736	314	-108
		5	3515	3093	2671	2249	1827	1405	983	561	139	-283	-705
		6	2918	2496	2074	1652	1230	808	386	-36	-458	-880	-1302
		7	2321	1899	1477	1055	633	211	-211	-633	-1055	-1477	-1899
		8	1724	1302	880	458	36	-386	-808	-1230	-1652	-2074	-2496
		9	1127	705	283	-139	-561	-983	-1405	-1827	-2249	-2671	-3093
		10	530	108	-314	-736	-1158	-1580	-2002	-2424	-2846	-3268	-3690

Figure 1. 9: Excel spreadsheet for cutting schedule improvement

The aim of the project team was to implement these two tools to optimize the cutting process in two different perspectives: short-term and long-term. For the most popular products, the best sellers, found in the analyze phase, it was important to study the orders in more depth, to identify some patterns on the data that can enable the construction of demand forecast for this products. Combining the use of the tools found with this forecast, could make it possible to buy or produce the optimal input length. On the other hand, the analysis showed that some products were rarely bought, with some lengths being required by few and smaller customers. For these products, it was more convenient to have a short-term perspective for the optimization, trying to schedule the cuts in a way that could minimize the amount of waste as much as possible, simply by utilizing the software tool.

CONTROL phase

It was important to make plans for further activities after a full implementation of the improvement solutions in a precise and easily understandable way. Having a person responsible for the follow up and for the report to the management and the operators, was another important aspect in the control phase. A **Control Plan** established both the time schedule for the implementations and the **Control Plan Schedule**. **Error! Reference source not found.** shows the control frequencies and the responsibilities. The person responsible for control was also responsible for implementation.

Solution	Control Frequency	Person responsible
Maintenance Program	Once a month	Production manager
Calibration Program	Once a month	Production manager
Short Term Planning	Once a week	Project Black Belt
Long Term Planning	Once a week	Project Black Belt

Table 1. 3: Control frequency and responsibilities

The project team designed a **Control Chart** in order to keep attention on the yield related to the customer orders. In particular, the control chart was supposed to focus on the yield data of the top five customers in terms of their financial importance for the company.

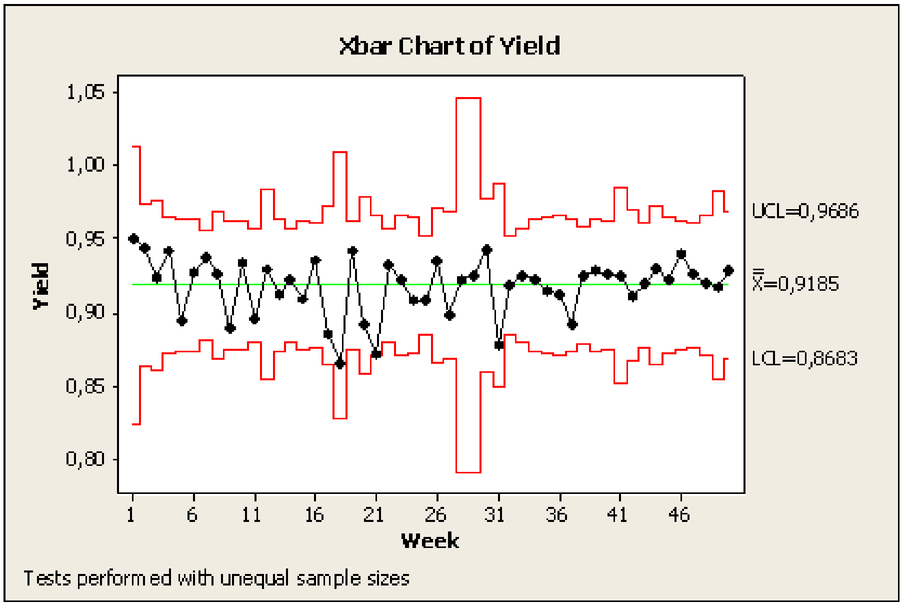


Figure 1. 10: Control chart for top five customers' orders of 2010

Error! Reference source not found. shows the control chart. The control frequency had to be weekly, which means that each point in the chart had to be calculated on a different sample size according to the number of orders of that specific week. Therefore, this chart has variable control limits. The control chart of **Error! Reference source not found.** was made on the basis of the yield data for the orders of 2010. However, an increase of the average value of the yield (more or less abrupt, depending on the efforts of the company) should also be taken into account due to the implementation of the recommended improvement solutions.

LEARN phase

Although not included in the classic DMAIC framework, another phase which was important, was the so-called Learn phase (Magnusson et al. 2003). At the end of the course, the project was not fully completed. The Improve and Control phases were planned but not implemented and some important issues had to be discussed with the management and other people involved in the project and particularly in the process itself. This is an important step for validating the solutions found and properly achieving the planned improvements.

The project showed the need to have more accurate data in some cases and in other cases data that were not yet available; these issues were related to the need to redesign the process in a more structured way, which was essentially one of the largest struggles during the project. One of the key lessons of this project was from the previous experiences of the company. For example, the matrix presented as one of the improvement solutions had previously been used in the company for a short period of time and seemed to be a practical tool for at least short-term improvements. However, this also showed the importance of fully completing the phases of improve and control.

Further analysis would have been necessary to achieve the real improvement of the process results and the increasing of yield in the cutting warehouse. After the end of the Black Belt course and project presentation, the company continued the discussions, meetings, and reflections. Meetings with the operators and involvement of them in the project's improvements suggestions made them more aware of the importance of the cutting waste. Thereinafter, even a little amount of scrap was reported in the system with the proper code, and this avoided miscalculations on the causes of scrap. What the Six Sigma philosophy tended to introduce in the daily business of the company was starting to produce results even if they must be more incisively supported by the top management.

An application of AHP + response latency model

Eight agri-tourism service attributes have been initially selected, based on a preliminary sector study (**Error! Reference source not found.**.1). In the early development of an agri-tourism service, one has to prioritize the efforts in terms of investments of time and money on the selected attributes, all potentially very attractive and important for customers. For this purpose a web survey was created with the platform Qualtrics® and the investigation was carried out through internet. The invitation to respond to the survey was sent to users of international agri-tourism blogs or to webmasters of farms located in various places of the world, and it was circulated on social networks. After two weeks of promotion, 155 respondents had undergone the survey. Figure 2.1 shows the distribution of respondents' country of origin.

With the eight attributes of **Error! Reference source not found.**.1, a total of 28 pairwise comparisons were submitted to each respondent in the survey. The attributes were visually displayed through clip art (see e.g. **Error! Reference source not found.**.1 for the comparison of the attribute 2 and attribute 5).

Reports for each respondent were easily obtained conducting the interviews online, with real time data logging from the server of the Qualtrics® web-site. After a first step of data download on a PC, a data cleaning operation was carried out. In fact, before going ahead with the calculations and application of the whole procedure, an accurate data filtering was necessary to eliminate the surveys which showed some major problems, as for example, too many clicks in the same page or too long response times on one question (meaning that the respondent was surely distracted by other tasks). This was a pre-filtering in which the response time was used, before any deeper calculation, as a way to get a good quality dataset, since it helped to identify cases which could negatively bias the analysis.

Table 2. 1: A list of agri-tourism attractive attributes

Attribute	***Description***
1. Swimming pool	A swimming pool accessible to the guests
2. Transportation services	Public transportation easily accessible, or shuttle bus to airport/city center
3. Typical architecture and nature	The structure is an historical building and/or the location is within a natural reserve
4. Comfort	Facilities typical of a hotel service
5. Sport facilities	Availability of sport facilities on site (e.g. gym, tennis court)
6. Gastronomy and typical food	Dining available, typical food, drinks.
7. Natural excursions	Organization of excursions in the surroundings
8. Cultural and traditional activities	Organization of visits to museums and places of cultural interest

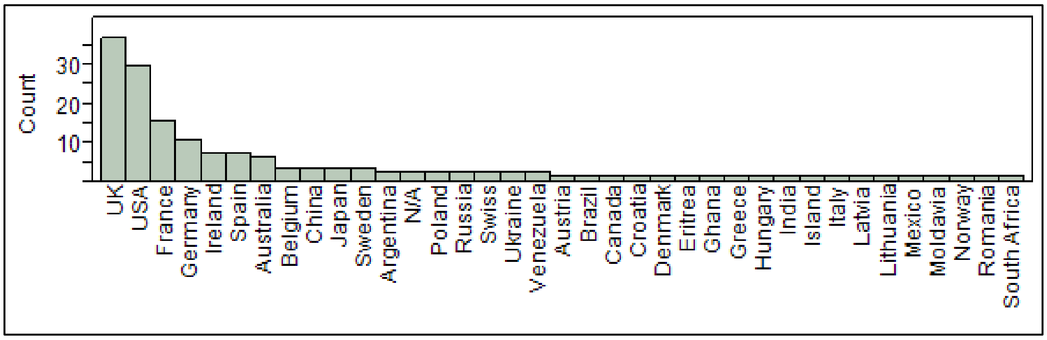
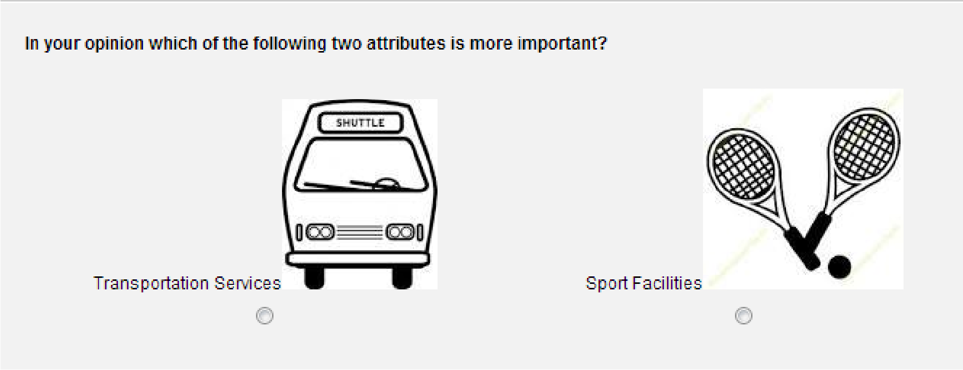


Figure 2. 1:Empirical distribution of the respondents' country of origin



In your opinion which of the following two attributes is more important?

Transportation Services ○

Sport Facilities ○

Figure 2. 2: A pairwise comparison visually presented to the respondent

After the pre-filtering a final sample of 102 surveys was used to implement the procedure and to calculate the importance weights of the attributes. Firstly the relative importance weights of each pair of attributes were computed according to the response latency model described above. Then a pairwise comparison matrix was created for each respondent.

The average consistency ratio of the pairwise comparison matrices so obtained is equal to 0.075. The respondents with a consistency ratio strictly below the suggested limit 0.1 are 86, which is a pretty satisfying percentage of the entire sample (average CR for this subgroup of respondents is 0.045). Two respondents were excluded from further analysis because of high CR of their pairwise comparison matrices (we used here a threshold of the 98th percentile of the entire distribution of CR).

The final rating of the eight attributes for each respondent was calculated through the implementation of the AHP procedure. In particular the weights were computed by the formulas in the model previously described. Figure 2.3 shows the empirical distributions of the calculated weights per each attribute.

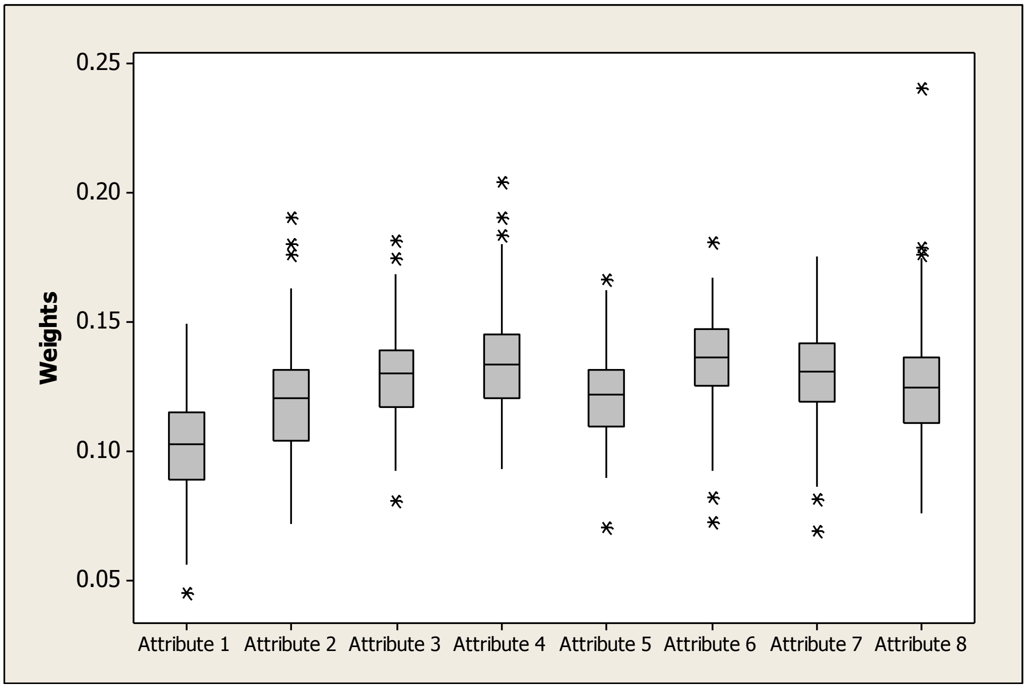


Figure 2. 3: Empirical distribution of the relative importance weights of the agri-tourism service attributes (see Error! Reference source not found.**.1)**

The prioritization of the agri-tourism attributes given by this sample of respondents provided a useful indication for the development of the service. For example the attributes 4 (Comfort) and 6 (Gastronomy and typical food) are considered the most important. Attributes 3 (Typical architecture and nature) and 7 (Natural excursions) follow, and so forth. A manager in charge of developing a the agri-tourism service or in charge of making new investments in the existing service may know in advance which priority it is best to give to the several aspects involved.

A good result of the case study is that the time spent by the respondents to complete the entire task is quite small, despite the 28 pairwise comparisons to undergo. **Error! Reference source not found.** shows the distribution of the total time spent by the respondents on the AHP part of the survey. The average time is 137 seconds with a standard deviation of 55seconds. It is also worth noting that the distribution of the time taken is not symmetrical.

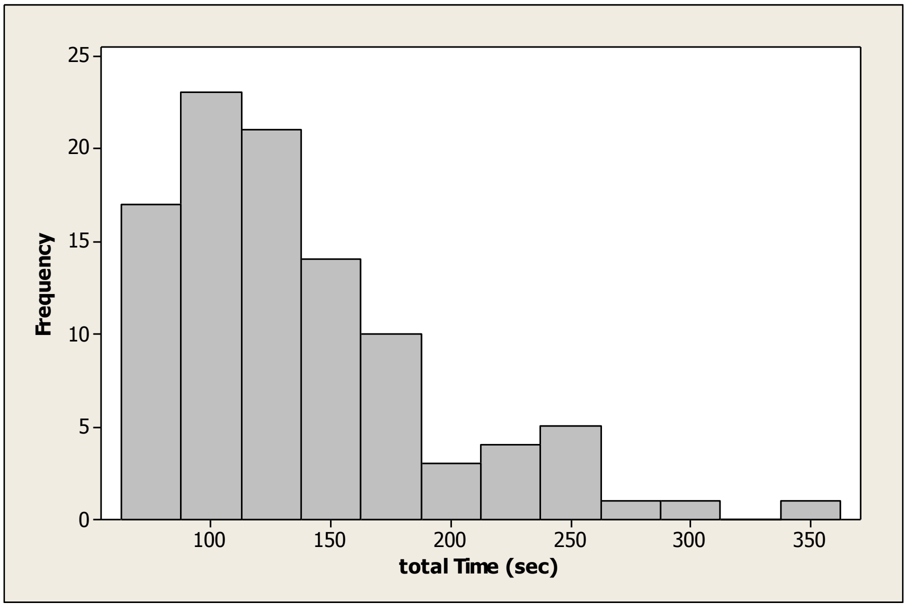


Figure 2.4: Empirical distribution of the total time (in seconds) spent on the interview