Aalto University
School of Science
Master's Programme in Industrial Engineering and Management

Julius Hokka

Controlled Experiments for Data-driven Retail Optimization

Master's Thesis
Helsinki, June 29, 2023

Supervisor: Professor Lauri Saarinen
Advisor: Erkka Saarinen, M.SSc.

Aalto University
School of Science

Abstract of the Master's Thesis

Author Julius Hokka

Title Controlled Experiments for Data-driven Retail Optimization

Programme Master's Programme in Industrial Engineering and Management

Major Strategy

Supervisor Professor Lauri Saarinen

Advisor Erkka Saarinen, M.SSc.

Date June 29, 2023 **Pages** 75 + 1 **Language** English

Abstract

Retailing is an industry shaped by low margins, large volumes and high fixed costs. Recently, consumers have become growingly demanding and price sensitive while competition has accelerated. Retailers must optimize their processes and generate revenue while cutting costs more than ever. Still, even the most advanced retailers often resort to intuitive decision-making, repeating old patterns that are often far from optimal. While modern analytics help optimize areas such as supply chain, merchandising and space planning, understanding the causal effects of decisions is difficult.

The scientific world has embraced experimenting as a causal inference methodology for centuries, but in business, its uses mostly limit to marketing and user interface development, because these offer an easy context to divide consumers into comparable groups. In retailing, the logic of a controlled experiment (CE) could be to apply a new business idea to only a group of stores, allowing a causal comparison to be made with stores where the idea was not applied.

This thesis synthesizes how retailers can benefit from CEs. Prior literature has not focused on a practical approach to the process that retailers should follow in order to successfully utilize CEs as a means for optimization. The study is conducted with a design science approach, where a literature review and interviews are used in designing a methodology framework for CEs in a retail context. The framework is further applied in two fabricated case examples to illustrate its applicability.

Running successful CEs needs vast understanding of the business opportunity that is being tested and thorough planning to ensure that the experiment produces credible results directed to the problem. In a retail context, a critical part of CEs is addressing the extensive implications of a CE outside the focal area of business. Complying with the framework steps ensures that the CE process is logical, but each case still requires adjusting its details accordingly. Future research could apply the framework in a pilot case and compare controlled experimenting to other causal methods in a retailing context.

Keywords retail optimization, controlled experiment, A/B testing, design science

Aalto-yliopisto
Perustieteiden
korkeakoulu

Diplomityön tiivistelmä

Tekijä Julius Hokka

Otsikko Kontrolloidut kokeet vähittäiskaupan dataohjautuvassa optimoinnissa

Koulutusohjelma Tuotantotalous

Pääaine Strategia

Valvoja Professori Lauri Saarinen

Ohjaaja Erkka Saarinen, VTM

Päiväys 29. kesäkuuta 2023 **Sivumäärä** 75 + 1 **Kieli** Englanti

Tiivistelmä

Vähittäiskauppaa luonnehtivat pienet katteet, suuret myyntimäärät ja korkeat kiinteät kustannukset. Viime aikoina kuluttajien vaativuus ja hintatietoisuus on kasvanut, ja kilpailu on koventunut. Vähittäiskauppojen on optimoitava prosessejaan, kasvatettava liikevaihtojaan ja pienennettävä kustannuksiaan enemmän kuin koskaan. Tästä huolimatta jopa kehittyneiden ketjujen päätöksenteko on tyypillisesti intuitiivista ja vanhojen tapojen toistamista. Modernit analytiikkasovellukset auttavat optimoinnissa läpi arvoketjun, mutta päätösten kausaalisten riippuvuussuhteiden ymmärtäminen on hankalaa.

Koeasetelmia on käytetty laajasti eri tieteenaloilla jo pitkään, mutta liiketoiminnan päätöksenteon tukena se on yleinen lähinnä markkinoinnissa ja käyttöliittymien kehityksessä, joissa kuluttajat voidaan helposti jakaa verrattaviin ryhmiin. Vähittäiskauppaketjuissa uutta hanketta tai ideaa voitaisiin testata osassa myymälöitä, mahdollistaen kausaalisen vertailun ulkopuolelle jätettyjen ryhmien kanssa.

Diplomityössä tutkitaan kontrolloitujen kokeiden hyödynnettävyyttä vähittäiskaupassa. Aikaisemmassa kirjallisuudessa ei ole käsitelty käytännöllistä prosessia, jolla vähittäiskaupat voisivat hyötyä kontrolloiduista kokeista. Työ hyödyntää suunnittelutiedemetodia, jossa kirjallisuuskatsauksen ja haastattelujen perusteella suunnitellaan metodologiaviitekehys kontrolloiduille kokeille vähittäiskaupan kontekstissa. Viitekehystä arvioidaan ja havainnollistetaan kahden kuvitteellisen esimerkin kautta.

Kontrolloitujen kokeilujen toteuttaminen edellyttää laajaa ymmärrystä tutkittavasta liiketoimintamahdollisuudesta sekä huolellista suunnittelua, jotta koeasetelma tuottaa uskottavia ja soveltamiskelpoisia tuloksia. Vähittäiskaupoille erityisen kriittistä on hahmottaa kokeen vaikutukset läpi organisaation. Viitekehyksen vaiheiden noudattaminen auttaa tekemään koeasetelmasta johdonmukaisen, mutta yksittäiset tapaukset vaativat prosessin mukauttamista. Jatkotutkimuksen tulisi keskittyä viitekehyksen soveltamiseen todellisessa tapauksessa sekä kontrolloitujen kokeiden vertaamiseen muiden kausaalisten menetelmien kanssa vähittäiskaupan kontekstissa.

Avainsanat vähittäiskaupan optimointi, kontrolloitu koe, A/B-testaus, suunnittelutiede

Preface

This thesis concluded my journey in Aalto University. Five years went by quickly, but the memories will never fade. It has felt a privilege to be a part of such an inspiring community and programme. Thank you all my fellow students for sharing this ride.

For this thesis, I'm extremely grateful for all the advice I received. Thank you Lauri for supervising me and giving me the directions to make this thesis academically valid and coherent. Thank you Erkka for providing a fascinating topic and guiding me through inspiring conversations. Thank you Anu for your inputs, they helped me become a better writer and sharpen the message.

Thank you Jussi for giving me the opportunity to dedicate time for this project. Thank you Chethana and all other team members for supporting me and making me feel like home every day.

Finally, thank you Mom and Dad, Joonas and Niklas for your continuous support and helping me grow to the person I proudly am.

Helsinki, 29.6.2023

Julius Hokka

Abbreviations

AI	Artificial intelligence
ARP	Automated replenishment program
CE	Controlled experiment
DC	Distribution center
KPI	Key performance indicator
ML	Machine learning
POC	Proof-of-concept
RTC	Randomized controlled trial
SKU	Stock-keeping unit
UI	User interface

Table of contents

1 Introduction 1
1.1 Scope and research questions 3
1.2 Structure 4
2 Background 5
2.1 On the shifting retail landscape 5
2.2 Data-driven decision making in retailing 7
2.3 Application areas of data-driven optimization 8
2.4 Controlled experiments 10
2.5 Retail applications of controlled experiments 15
2.6 Summary and theoretical synthesis 18
3 Methodology 21
3.1 Research design 21
3.2 Data collection and analysis 23
3.3 Methodology validation 26
4 Findings 29
4.1 Interview analysis 29
4.2 Controlled experiments' methodology in retail 42
4.3 Illustrative review 55
5 Discussion 64
5.1 Theoretical contributions 64
5.2 Managerial implications 67
5.3 Limitations and further research areas 69
References 71
Appendices 76
Appendix 1 – Interview template 76

1 Introduction

It is safe to say that retail industry has been under massive changes (Hänninen et al., 2021). Novel technological developments have shifted entire business models (Sorescu et al., 2011) and given access to unforeseen quantities and qualities of data (Dekimpe, 2020). The recent pandemic and geopolitical conflicts (e.g., Roggeveen & Sethuraman, 2020; Ngoc et al., 2022) have further made the world growingly uncertain, putting pressure on retailing as one of its most important sectors. Under last decades, new giants such as Amazon and Alibaba have stepped up (Hänninen et al., 2021), and several ones have faced demise with a magnitude that has coined the term *retail apocalypse* (Helm et al., 2018). But an apocalypse is perhaps a too fierce metaphor for what is happening. Retailing remains as one of the largest and most diversified and dynamic industries in the world (Dekimpe, 2020) and while changes happen and boundaries shift, the need to improve existing processes and make decisions based on facts will only be amplified in the upcoming decades.

Sorescu et al. (2011) portray the recent stretch in retail industry:

> *Retailers today can no longer be accurately characterized as "merchant intermediaries" that buy from suppliers and sell to customers. Rather, they are best described as orchestrators or conductors of two-sided platforms that serve as ecosystems in which value is created and delivered to customers and, subsequently, appropriated by the retailer and its business partners.*

From a pure research aspect, Dekimpe (2020) discusses what makes retailing an attractive domain, concluding that in addition to its size and multi-faceted, dynamic nature, retailing poses an interesting field to contrast observative matters with academic thinking and business analysis through an abundance of data. In fact, the volume of data created by retailing is massive. Already a decade ago, Walmart collected estimatedly 2,5 petabytes of data every hour through customer transactions (McAfee & Brynjolfsson, 2012). In addition to the buzzword-esque *big data* revolution, Bradlow et al. (2017) declare the phenomenon as *better data* revolution, implying that new data sources and tools to analyze it have also increased the quality of retail data.

But by no means is this advent of data-driven retailing solved. Big data can conventionally only provide conclusions of the past, and true innovations that are often against intuition of even experienced executives, are thus hard to come by with only descriptive analysis (Thomke & Manzi, 2014). Even with adequate data, managers predominantly make decisions instinctively and with high confidence despite a clear room for improvement and optimization (Anderson & Simester, 2011). Already in the mid-20th century, alternatives to historical data studies or sample surveys as approaches to retail decision making have been discussed, namely experiments that isolate the effect of given variables to allow true causal inferences (Brunk, 1953). Retail applications of controlled experiments (CE) as a decision-making tool have gotten fairly modest attention in academia despite clear success stories. These existing implementations, discussed in e.g., Thomke & Manzi (2014) are however more or less in-house built schemes. As the retailing sector is heavily multi-faceted and in a constant flux (Dekimpe, 2020), a standardized methodology for conducting CEs would have both managerial and technological relevance.

The retail industry is characterized by minimal net margins and high fixed costs (Fisher & Raman, 2018). This, combined with massive product portfolios and chains operating hundreds or even thousands of outlets (Dekimpe, 2020) makes the sector extremely favorable to even small incremental improvements having major impact in revenue and margins. Therefore in retail, strategic, tactical and operational planning levels are more closely knit together. Individual decisions made by retail managers responsible for e.g., merchandizing or supply chain planning have a straightforward impact on profit and loss, working capital and the overall impression by consumers. Any means by which more justifiable decisions can be made, are thus desired. Experimentation can provide a transformative way of conducting business when done right (Thomke, 2020).

This thesis investigates the retail sector as an application area for CEs. A design science approach for creating a conceptual methodology framework is taken, and the research includes interviews with retail solution software vendor experts to highlight the practical issues with current processes and CEs as a methodology. The aim of the thesis is to bridge the gap between a well-known research method in CEs, and a practical setting where such rigor is yet rarely used. After all, data is only as good as the tools and methods used to analyze it.

1.1 Scope and research questions

As the terms *retail* and *retailing* encompass an extremely wide area of activities in modern society, the perimeter of the study has to be unambiguous. Retailing is generally defined as the activity of selling goods or certain services to the consumer public.[1] In addition to the vast array of industry sectors such as grocery stores, convenience stores, specialty retailers, drug stores and department stores, the industry today serves multiple overlapping business models and strategies. Sorescu et al. (2011) mention e.g., the propagation of e-commerce, mass customization, streamlined operations in fast fashion, innovative customer interfaces and different retailing formats offering essentially same products, for instance food that can be bought from a grocery store or a mass merchandiser differing in size, pricing and assortment. Nowadays the word omnichannel is used to describe the seamless customer experience regardless of the channel used to make purchasing decisions, evolving from the term multi-channel that just emphasizes several parallel channels for operating business (Verhoef et al., 2015).

As the topic of CEs in retail applications is fairly overlooked and the fundamental principles of retail remain the same regardless of industry sectors and business models, this thesis does not make too heavy restrictions on which branches or business types to study. But from an implementation perspective, altering any features that hope to bring notable differences has to be tangible in order to be able to distinguish control and treatment groups and make aggregate level decisions. In e-commerce and social media, this is rather commonly and easily practiced through A/B testing where users are blindly divided into two (or more) groups that receive e.g., alternate user interfaces (UI), leading to distinct conversion rates and subsequent conclusions on what consumers prefer (Kohavi & Longbotham, 2017). In more traditional brick-and-mortar retail, such experiments dividing consumers into sections would be more difficult and, in many instances unethical. But experiments can for instance be run between store clusters where e.g., accumulated point-of-sales (POS) data speaks for the worth of the treatment in comparison to the control group during a certain time frame.

[1] See common dictionary definitions by e.g., Cambridge or Britannica.

This thesis will focus on experimentation that touches the physical operations of retailing and leave e-commerce A/B testing out of the scope. Therefore, the scope is bound by retailing where brick-and-mortar plays significant enough of a role to allow such experimentation between stores or channels. The terminology will remain as *retail* instead of *brick-and-mortar retail* to avoid confusion. An omnichannel retailer could well conduct experiments that include both physical and online channels. The practical areas of application will be discussed in later chapters.

With this definition, the objective of the thesis is to answer the following research questions:

> *1: How do modern analytics help retailers make better decisions and optimize their processes?*
>
> *2: How do controlled experiments fit the landscape of data-driven retail analytics?*
>
> *3: How should a systematic controlled experiment solution be designed in order to help retail decision makers in improving their business?*

The first two research questions aim to comprehend the analytics environment that retailers occupy, and the solutions currently in use. The first has a more overarching intent, while the second focuses on understanding how CEs situate in this overall context. To answer these questions, this study will conduct a literature review and interviews with subject matter experts. These interviews will also provide primary data for the third research question, which seeks to synthesize and validate a practical methodology framework for CEs in a retail context through a design science approach.

1.2 Structure

Chapter 2 presents a literature review on data-driven retail decision making and optimization, followed by CEs generally and in retail applications. The third chapter discusses the study methodology including research design, data collection and methodology validation. Results are presented and analyzed in chapter 4 and finally, chapter 5 provides discussion including managerial and theoretical contributions, limitations and suggestions for further research.

2 Background

This chapter presents a literature review on controlled experiments (CE) for retail decision making. First, the retail industry and its recent shift in nature is discussed. Then, in sections 2.2 and 2.3 retail decision making through data is explored. After this, CEs are introduced as a general concept in section 2.4, followed by how CEs are used in retail applications in section 2.5. Finally, section 2.6 synthesizes the literature review into a theoretical framework that serves as the context for the empirical part described in chapter 3.

2.1 On the shifting retail landscape

While humans have practiced trade for as long as we have evidence of civilizations, the fundamentals of retail as we know it have not been out there for particularly long. The oldest marketplaces have been operating for several hundred years, but McArthur et al. (2016) point out that the arrival of department stores not until the mid 19th century was a pivotal point in how the industry has shaped to what it currently is. The *raison d'être* for retailing is to function as the intermediary between suppliers and consumers, which according to Zentes et al. (2017, p.4) is modernly understood from a transaction cost perspective; it is more favorable for a supplier to use a middleman if it minimizes transaction costs. Naturally, from a consumer perspective, this applies as well, since consumers can often consolidate their needs into only a handful of transactions by e.g., shopping at a department store or supermarket.

The COVID-19 pandemic certainly boosted the recent developments of retail processes and provided both opportunities and challenges for retailers in a landscape that is inherently difficult to predict (Roggeveen & Sethuraman, 2020). Regardless of the pandemic, the sector is shifting from physical retail to online shopping, and consumer experience is evergrowingly important in an omnichannel world (Helm et al., 2018). The quote by Sorescu et al. (2011) in chapter 1 stating that retailers should be viewed as orchestrators of ecosystem platforms for consumer value creation instead of just being intermediaries for customer products, is definitely an interesting one from the perspective of omnichannel business models and increased customer-centricity seen lately in many of the traditional incumbents (Hänninen et al., 2021).

But as can easily be seen anywhere humans are settled in, classical retailing has its place and will likely remain a vital part of our societies for generations. New business models arise (Sorescu et al., 2011), but the basic challenges remain and the general role and function of retailers stays practically the same, as discussed in Zentes et al. (2017, p.3-10). Retailing is all about bringing the right products to the right place at the right time (Fernie & Sparks, 2009, p.5) to allow transactions to happen. At a store level, several operational problems have to be solved on a daily basis. Mou et al. (2018) list the most common parts in retail store operations where continuous decisions must be made: *demand forecasting, in-store logistics, inventory management, assortment and display, product promotion, checkout operations* and *employee management*.

At a higher level, the entire retailing value chain is full of issues to address. Fernie & Sparks (2009, p.15) present the extended value chain that retailers have to operate in order to create value for consumers while minimizing costs. Figure 1 portrays this concept, highlighting the intermediary role of retailers. Rooderkerk et al. (2022) find nine decision areas where most of the retail analytics literature focuses: *inventory management, product promotions, distribution and delivery, demand planning, assortment planning, returns handling, customer service operations, employee management* and *warehousing*. While these heavily overlap with the aforementioned operational store problems by Mou et al. (2018), Rooderkerk er al. (2022) point out that most of them also dwell on strategic and tactical decision levels instead of just operational. Dekimpe (2020) has similar findings, that retailers can benefit from big data throughout the value chain, namely in *marketing, merchandising, operations, supply chain* and *new business models*.

All in all, an argument can be made that new ways of conducting the same fundamental activities are what is shaping the industry, rather than a rudimentary change in what retailing is in the first place.

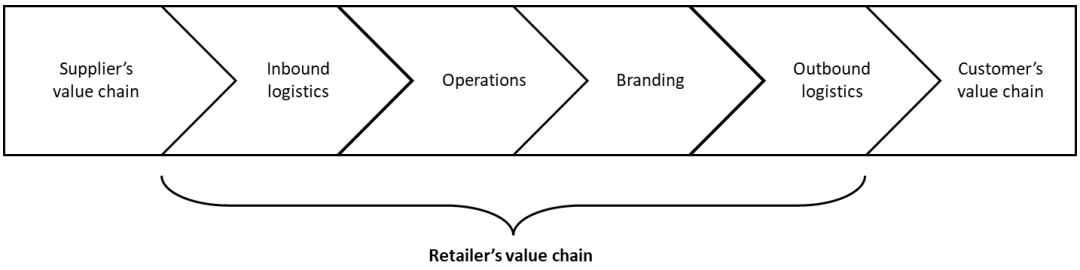

Figure 1: Extended retail value chain, modified from Fernie & Sparks (2009, p.15)

2.2 Data-driven decision making in retailing

As mentioned earlier, retailing is an arena for proper quantities and qualities of data. Dekimpe (2020) goes as far as calling it "almost by definition a big data industry". Bradlow et al. (2017) discuss big data dimensions in retailing, namely *customer, product, time, location* and *channel*. Nowadays data can be collected granularly in all these dimensions, with for instance POS data linked to the specific customer, basket, time and location through rather simple means such as a loyalty card. Analyzing this multidimensional data can give valuable insights through not just descriptive and diagnostic analytics, but also predictive, prescriptive and autonomous analytics (Rooderkerk et al., 2022).

Analytics solutions can range all the way from simple spreadsheet calculations to complex artificial intelligence (AI) and machine learning (ML) applications that breathe big data. Guha et al. (2021) classify all AI solutions within the retail domain to customer-facing and non-customer-facing applications. This division essentially denotes whether the AI is being communicated to the customer, or that the calculations and analysis happens behind the scenes. This thesis will focus on the non-customer-facing applications in not just AI solutions but all data-related tools that retailers can use in aiding business decisions. Even if a notable change happens in a retail store, the underlying decision making process is not explicit from the consumer perspective.

Data is not of much use if the analytics built around it are poorly suited with the data itself or the business problem. Bradlow et al. (2017) mention that the sheer volume of data, e.g., several years behind, rarely adds value compared to examining only recent years of data. With predictive models, understanding the theoretical foundations is crucial, as Bradlow et al. (2017)

continue with stating that despite having advanced predictive algorithms and machine learning tools, retail managers' decision making will be *far from being fully automated*. In their highly popular Harvard Business Review article on big data revolution, McAfee & Brynjolfsson (2012) assert that in order to succeed with data, leadership teams need to ask the relevant business questions and include human insight into the processes. In retail context, this is even more crucial since it is a low margin, high fixed-cost industry and thus even small improvements in operations can pan out heavily in net profits and market capitalization (Fisher & Raman, 2018).

Hence, there is a risk of shifting retail managers' thinking to tactical decisions at the cost of strategic considerations (Dekimpe, 2020). Data analysis and AI should not substitute executive and managerial decision making, but rather augment it through valuable discoveries that are interpretable (Elgendy & Elgaral, 2016). Paradoxically, whereas big data and advanced analytics can shift the managerial judgement process towards tactical and even operational questions, these tactical and operational questions shift into a direction that is growingly relevant from a strategic perspective. Data-driven decision making is becoming equally important for executives and directors, too, as they have to carefully navigate the potential and risks that novel analytics solutions hold and make sure their businesses remain competitive.

2.3 Application areas of data-driven optimization

Since retailing weaves around material flows, there is plenty of optimization in the supply chains, both upstream and downstream. The fundamental aspect of value generation in supply chains is accurate demand forecasting, that also encompasses coordinating this information between all stakeholders (Syntetos et al., 2016).

Poor demand forecasts reflect to all other parts of operations, including store fulfillment and availability, inventory management and product perishing (Mou et al., 2018). At best, demand forecast calculation doesn't happen in a black box, but rather it has the possibilities to intake human judgments as part of decision making (Syntetos et al., 2016). While the theoretical and statistical understanding of different forecasting methods are well understood (see e.g., the textbook by Hyndman & Athanasopoulos, 2021) and there is general availability of commercial software purposed for complex demand forecasting (e.g., ML models), retailer adoption of such services has not yet been dramatic due to challenges in convincing that complex models

outperform traditional ones (Fildes et al., 2022). ML models are not always necessarily better than more traditional statistical counterparts, but as the amounts and types of data is continuingly growing, ML solutions will be able to take more factors into account.

In supply chains, forecasts only hold as much value as the processes that utilize them. Syntetos et al., (2016) note that with higher complexities of data, a major challenge for enterprise resource planning (ERP) systems is accessing and combining all crucial data for whatever happens after a forecast is calculated. Namely, the next sequential part is to utilize the demand forecasts to optimize the replenishment process of all stock-keeping units (SKU) in all outlets and channels. Sillanpää & Liesiö (2018) discuss the way of managing store replenishment through *planned orders* which use e.g., material requirements planning (MRP) models to suggest ordering dates for each SKU from a central distribution center (DC). ERP systems often contain an automated replenishment program (ARP) that allows the supply chain planning personnel to oversee the planned orders integratedly within stores and DCs (Kiil et al., 2018).

Simplified, this cycle of demand forecasting and replenishment is the backbone of retail supply chain operations. But as discussed in literature (e.g., Bradlow et al., 2017; Fisher & Raman, 2018; Mou et al., 2018; Dekimpe, 2020; Guha et al., 2021; Rooderkerk et al., 2022), retailing involves several other aspects that need to be taken care of to remain competitive and profitable. In a 2020 report, McKinsey argue that the biggest impact by data-driven automation in retail can be achieved in merchandising planning, pricing, promotions, markdowns and inventory (McKinsey, 2020). The same areas were already recognized in the consultancy's 2011 report on big data opportunities (McKinsey, 2011), indicating that the general understanding of the possibilities is on point, while the advancement and acceptance of planning and optimization solutions addressing this potential is fulfilling slowly.

Similarly to these managerially targeted discussions, academic literature recognizes alike opportunity areas in retail backend operations. For example the textbook by Kerkhove (2022) focuses foremost on optimizing pricing, promotion and markdown optimization and inventory through data, which is in line with the McKinsey reports. Fisher & Raman (2018) discuss the role of data in helping to execute retailer's extant business models more optimally, including assortment optimization, dynamic pricing and decisions to open and close new locations. Rooderkerk et al. (2022) conduct a thorough

operations management (OM) journal search to distinguish the nine overarching retail analytics decision areas mentioned in section 2.1.

But even if the strategic, tactical and operational areas are well known and there are solutions available to support the management of them, rigorous analysis and interpretation conducted by looking at past data usually requires complex technical skills (Anderson & Simester, 2011). One can easily make descriptive analysis on what has happened in the past, and analysis on correlating phenomena. But as Bradlow et al. (2017) put it, predictive analyses that help sense what is ahead need more delicate tools. ML forecasts, ARPs, product attribute-driven assortment optimization (Fisher & Raman, 2018) and Bayesian estimation for household purchase timing and brand sensitivity (Mou et al., 2018) are good examples of analytical triumphs that give valuable insights to retail managers that struggle to make sense what creates value.

But in addition to understanding *what* creates value, these methods do not capture the true *how much* and *why* in a causal manner. Bradlow et al. (2017) find that in all empirical research a significant challenge is to ensure that whatever is being done gets isolated in order to allow causal inferences, continuing with stating that controlled tests are necessitated to measure an effect of any decision. Rooderkerk et al. (2022) find that typically diagnostic analytics include an empirical part, often in the form of a field experiment. Next section discusses controlled experiments (CE) as a methodology that allows examining the effects of actions that instead of correlation indicate causality and isolate the effect.

2.4 Controlled experiments

As far as is known, the first written documentation of a CE can be found in the Old Testament, in the first chapter of the Book of Daniel[2]. In the tale, Daniel – with three other young men – asks a chief of staff for permission to only eat vegetables and drink water for ten days, while another group of young men eat the same food as the king and drink wine. After the trial, Daniel and friends were comparatively much healthier and better nourished than the other group, which led to a decision of them getting served produce and water ever since. While scriptures should by no means be considered of

[2] (Daniel 1:5-16). Bible available at e.g., bible.com.

scientific proof any more than fairy tales or modern television drama, it is intriguing that already a few millenia ago, scientific approaches to commonplace questions have been applied. In this thesis, the selected literature on experimentation focuses on the era in which retailing started to find its current characteristics.

During the past century, plenty of literature has been written on the topics of experiments and experimental design. While the classic books by Fisher ([1935] 1971) and Campbell & Stanley (1963) did not coin experimenting or the different designs to conduct them, they are perhaps the most renowned pieces of work within the area of experimental design. These pieces of work have paved the way for the versatility and rigor that experiments today allow researchers to practice. More recently, textbooks such as Shadish et al., (2002) and Montgomery (2013) have become well distinguished academic reference works in the field. However, academic articles that exhibit results from experiments rarely touch on the fundamental rationale and characteristics of experimenting. For example, Fisher et al. (2019) merely just explain that their study leverages a quasi-experiment with a difference-in-differences analysis on the experiment data, leaving it to the reader to grasp what that exactly means and why it makes sense in that particular study. This indicates that on a general level, experimenting as a methodology is fairly mature, and its advantages as a means to making valid inference of an intervention are both well understood and agreed on within the research world. This section presents the essentials of experimenting, and particularly CEs.

Shadish et al. (2002, p.3) assert that the fundamental characteristic of experiments is to purposely vary a specific object in order to later disclose an effect on something else. Pandey & Pandey (2015, p.89) define the experimental research method through the relationships between manipulated and measured variables. Montgomery (2013, p.1) aligns with these, continuing that experimenting should produce a model for *process or system improvement or other decision making*. These are reasonable general definitions of experimentation as a method, but do not per se capture the essence of control. In the Bible example, a proper decision could only be made because the group that did not switch to vegetables and water existed, allowing people to see a clear contrast. Indeed, having control is what gives the methodology its strength. Throughout, this thesis refers to a CE as a trial setup in which the sample is divided into treatment and control groups where the treatment is being applied in only the indicative subset, and these groups

are then compared to one another, similar to the popular A/B testing scheme mentioned in e.g., Kohavi & Longbotham (2017). Besides A/B testing, other common rubrics for the topic are *split test* and *randomized controlled trial* (RCT), of which the latter has the notion that it necessitates randomization. CEs, in this thesis, do not necessarily stipulate randomization, but an important remark is that the mathematical models used to analyze experiments are generally simpler when samples are randomized.

Randomized or not, CEs need to have a clear protocol in order to satisfy principles of the scientific approach. Even in practical settings that are far from the scientific world, rigor is needed for an experiment to be reliable and purposeful. Montgomery (2013, p.14) defines seven guidelines for designing experiments. These are summarized in Table 1. Most notably, the first item *Recognition of and statement of the problem* is in line with what was discussed in section 2.2 about management needing to ask the right questions before using complex data analytics in solving managerial challenges.

Interestingly, the guideline doesn't state a clear need to form a hypothesis at the problem definition stage. This is in contrast to e.g., Kerkhove (2022, p.247), who emphasizes the importance of a specific and measurable hypothesis in order to design a valuable experiment, going as far as stating that the main difference between experiments and data science is that experimening requires a hypothesis to be tested whereas data science is about diving into large amounts of data with an unknown objective. But even if problem framing and hypothesis formulation are on point, experimentation can only bring as sensible results as the underlying issue examined is. This requires practicioners to have expertise in the characteristics of the field. In business settings, understanding of what creates value and impacts key performance indicators (KPI) in the particular context is vital.

Table 1: Guidelines for designing experiments, adapted from Montgomery (2013, p.14)

Guideline	Description
1: Recognition of and statement of the problem	Clear description of the overall objective of the examined problem. Recognition of whether experimenting can answer the question and if, how.
2: Selection of the response variable	Choosing the output that the experiment hopes to alter. The measured characteristic of it, e.g., average, standard deviation, summation.
3: Choice of factors, levels and ranges	Which factors are varied, which are purposely held constant and which are allowed to vary naturally. How noise is controlled. Selecting the magnitude and scale of the tested variation.
4: Choice of experimental design	Defining the experiment setup and sample selection (incl. randomization). Run order and periods of analysis. Selection of empirical (mathematical) model.
5: Performing the experiment	Thorough monitoring and ensuring factor settings beforehand. Consideration of pilots or test runs to validate steps 1-4.
6: Statistical analysis of the data	Hypothesis-testing, graphical presentation of data, results of the empirical model, model adequacy testing.
7: Conclusions and recommendations	Practically oriented analysis of the results. Concrete action recommendations. Decisions on follow-up runs and iterations.

Item 4 of the Montgomery (2013) guidelines in Table 1 necessitates an experimental design to be chosen. Experimental design, or *design of experiments*, is defined by Fisher ([1935] 1971) as the *logical structure* of the experiment. Campbell & Stanley (1963) distinguish different experimental designs, of which the *pretest-posttest control group design* is the most recommended and commonly used, and practically easy to implement. It also allows a straightforward way of conducting data analysis in practical, dynamic settings. The design, adapted from Campbell & Stanley (1963) is formulated as;

$$(R) \quad O_1 \quad X \quad O_2$$

$$(R) \quad O_3 \qquad O_4$$

where the first row [(R) O_1 X O_2] indicates the workflow for the treatment group and the second row [(R) O_3 O_4] the control group. R indicates randomizing the test population into the corresponding groups, and is in parenthesis because sometimes randomization might not be practically or ethically possible, making such experiment a quasi-experiment. O_1 and O_3 denote pretest observations, and O_2 and O_4 posttest observations respectively for the corresponding groups. The treatment group ($O_1 \rightarrow O_2$) is exposed (X) to a change in an experimental variable or event, which is bypassed with the control group ($O_3 \rightarrow O_4$). (Campbell & Stanley, 1963). Perhaps the most commonly known example of this experimental design is the classic clinical trial, in which a group of randomly selected patients are given a real medicine or treatment while another group gets a placebo. Pretest and posttest results are then compared by statistical analysis to determine whether there is a difference between the groups i.e., what is the true causal effect of the treatment.

As to causal effects, it is a commonplace research adage that correlation does not indicate causation. Typically, in our everyday lives we mistake them for one another, while they are clearly two distinct relationship types. Shadish et al. (2002, p.7) give a classic example of income and education, which more often than not correlate, but one can still not make the inference that one causes the other. The chance of a third variable – a confounder – affecting both must be studied. In this example, IQ could be a valid variable to study as a confounder to both income and education. The authors continue with pointing out that experimenters need to identify the confounders when making conclusions. Randomization is a good way of eliminating confounding.

Moreover, CEs are valid in describing causal descriptions (does A cause B?), but often lack in causal explanations (why does A cause B?) (Shadish et al. (2002, p.7). In many practical settings, such as retail, causal description is already a satisfactory outcome, but it goes beyond the experimentation to truly understand the underlying mechanisms of why exactly a certain type of intervention such as an altered store layout causes the observed effect of

more or less sales[3]. Whether or not it is of the experimenter's best interest to truly understand the underlying relationships and micro-level fundamentals depends on the context, but discovering this kind of information might be helpful in planning of activities. For instance, a retailer might learn that upon a holiday season, placing toothbrushes on a promotional display shelf might accelerate their sales, but the same doesn't hold for toothpaste. Experimenting can only help finding this causal description, but not the causal explanation that people tend to buy new toothbrushes on special occasions, typically for the whole family, whereas toothpaste demand is more constant and thus a worse fit for such display promotions.

2.5 Retail applications of controlled experiments

As mentioned, retailing is characterized by low margins, high volumes and plenty of data opportunities. It is a highly competitive field where customer satisfaction has a growing importance, but the means by which this is pursued are more or less heuristic and intuitive. Rooderkerk et al. (2022) find that risk aversion and inertia are among the largest cultural reasons why retailers have been reluctant to adopt analytics solutions, and that many practicioners dread analytics as a threat to their *art* of retail, when in reality modern solutions should be approached as help to execute it better. From a cultural perspective, Rooderkerk et al. (2022) proceed with discovering that the firms getting the most out of analytics are the ones that have a culture of experimentation that supports scaling new ideas based on analytically processed experiment results. This is well in line with Thomke (2020), who in a Harvard Business Rewiew article calls for companies to adopt a cultural change where business decisions are actively experimented with to give insights.

The most common, and arguably easiest CEs for retailers are conducted online, where it is straightforward to show users different versions of the same UI. Kohavi & Longbotham (2017) describe this A/B testing as an efficient and easy, but critical tool to learn whether the software or the business it offers need to be changed. Retailers are already doing this in vast amounts (Thomke & Manzi, 2014). A recent example of Amazon shows a

[3] Recently a supermarket in Finland experimented (albeit not controlledly) with dessert placement within the store. Putting desserts on the right-hand side of the aisle increased their sales notably. News in HS.

similar, but differently built approach: it launched a pricing experimentation platform that helps determine the success of certain pricing policies by applying different policies to treatment and control groups of products instead of users (Cooprider & Nassiri, 2023). The fact that Amazon has decided not to price discriminate, but instead conduct experiments between product groups, speaks for the flexibility of A/B testing as a concept that can help online businesses understand their domain better.

Technically anything that can be altered, and from which data can be extracted, can be experimented with in online settings. But as this thesis focuses more on CEs in the traditional essence of retailing, more emphasis is put on the extant literature and case examples within offline applications. Plenty of experimenting probably happens without the retailer disclosing it or without an academic study contributing to it. Nonetheless, several records of CEs still exist to demonstrate the concept.

Academia has seen first retail CE papers published already decades ago, with e.g., Applebaum & Spears (1950) describing how to plan and implement a CE in grocery stores to test packaging, displays, pricing, promotion plans, new products, store equipment and work methods. Brunk (1953) discusses CEs in retail merchadising, particularly in product assortment with Latin square designs, which means that n treatment differences (assortments) are rotated between n test units (stores) for fixed periods and thus each store-assortment-period combination can be visualized in a $n \times n$ matrix.

Several studies have been published that incorporate CEs as the research methodology: Cunningham & O'Connor (1968) trialed with insecticides, rotating four different price and display combinations in eight stores, effectively a "two 4×4 Latin square" experiment. Anderson & Simester (2003) present results from field experiments on prices ending with $9, concluding interestingly that $9 endings are more effective with new products than when customers have vast information of a product.

More recently, e.g., Bradlow et al. (2017) tested a new predictive analytics-driven pricing algorithm with treatment and control groups of stores for 12 weeks. The regression response variable being gross margin dollars, there was a 40c net improvement in the treatment stores ($p < .01$) compared to control stores with 100,000+ SKU observations in each group. It is a great example of a fairly simple analysis of treatment and control store cluster data where the test groups were not large in size (21 stores each) but as the scope encompassed 14 product categories and 12 weeks, massive amounts of data

could be obtained, and thus statistically significant evidence of the causal linkage between the new pricing policy and gross margins was collected, helping the retailer make confident decisions on pricing strategies.

Fisher et al. (2019) conducted a quasi-experiment in which the effects of quicker (online order) delivery due to a new DC were examined against a control group of customers that did not geographically benefit from the new opening. Using the slightly more complex Poisson regression with propensity scores as regression weights allowed the researchers to obtain more comparable results, as the quasi-experimental design (due to the geographical nature) made the groups somewhat heterogenous by several important sample characteristics.

The Fisher et al. (2019) example underlines the importance of carefully designing the experiment, particularly items 3 and 6 of the guidelines presented in Table 1. In addition to these guidelines, experimenting in a retail context has several requisites that distinguish a succesful experiment from a failed one. Apart from a justifiable business problem and well formulated hypothesis, experimenters need to understand the cost benefit analysis of covering all variable and fixed costs with the profits that the experiment is expected to generate. Another thing to consider is how large the sample sizes need to be to obtain a desired statistical power (e.g., 95%). These stages of the experiment planning are crucial to determine whether the risks are too high, and if the experiment ought to be scrapped.

After this, the retailer needs to architect the sampling so that parallel experiments do not contaminate each other, and that the groups represent the population as homogenously as possible. This can be achieved with stratified sampling where the population is first divided into homogenous subgroups, after which the actual sampling to treatment and control groups is done to ensure that different characteristics (e.g., store sizes, customer segments) are represented in both groups. (Kerkhove, 2022, p.255).

While these literature examples show that CEs can be used to improve various parts of retail businesses, the fact that there are not many anecdotes of companies adopting a continuous cycle and culture of experimenting is rather interesting. Instead, the examples are more or less run by academics. Perhaps the understanding of the rigor and requirements of CEs is not widespread in the business realm. Another reason could be that not many case studies are published outside the academic world. For instance, Thomke

& Manzi (2014) and Thomke (2020) only seem to touch upon the business examples where companies administer CEs by themselves.

2.6 Summary and theoretical synthesis

This section summarizes the findings of the literature review and synthesizes a theoretical framework of the characteristics and positioning of CEs within the overall context of retail analytics and optimization.

The framework, presented in Figure 2, builds upon the analytics continuum by Rooderkerk et al. (2022), which is developed from the technology corporation Intel's guide on advanced analytics (Intel, 2017). Rooderkerk et al. (2022) define descriptive (what happened?) and diagnostic (why did it happen?) analytics as traditional analytics, providing hindsight and insight, respectively. Of more advanced analytics, predictive (what will happen?) and prescriptive (how can we make it happen?) analytics provide foresight to the future. Further down the continuum, there is a new frontier with autonomous analytics that is defined as continuous learning and optimization with minimal human intervention. Figure 2 shows the analytics continuum at the base of retail analytics and optimization, with CEs overarching across the continuum as a concept that can enable and benefit from all analytics types.

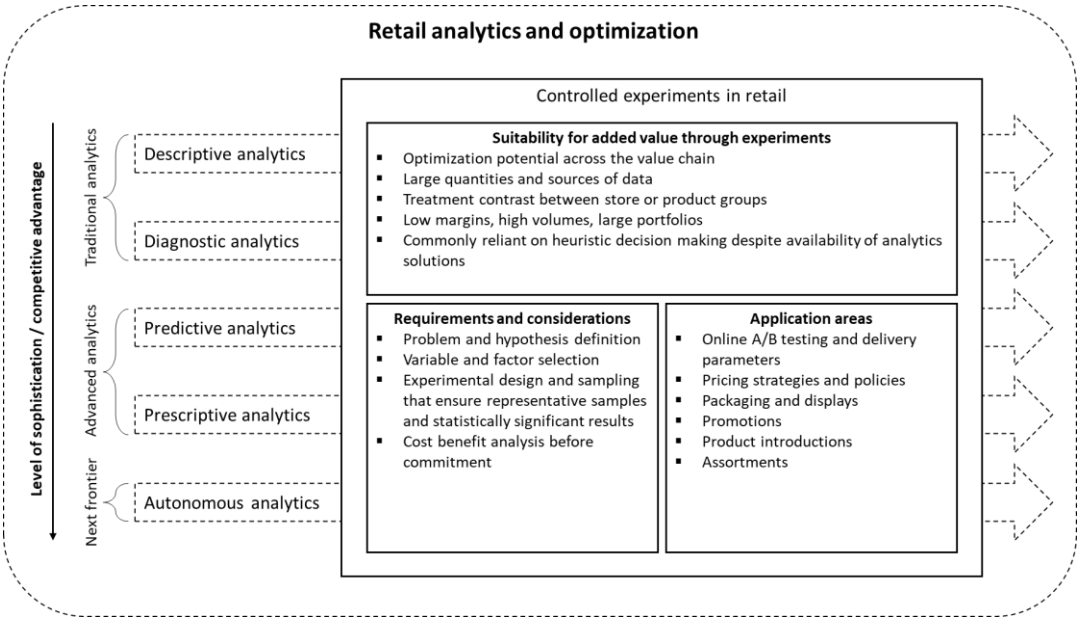

Figure 2: Controlled experiments in the context of retail analytics and the analytics continuum by Rooderkerk et al. (2022)

As the analytics continuum encompasses essentially all kinds of methods and levels of sophistication from centuries old statistical concepts to modern autonomous tools (e.g., AI and ML models), it would intuitively feel bold to claim that CEs have relevance from all analytics types. But this can be illustrated with an example of promotion optimization. A controlled experiment on a new promotion will itself bring data that can be *descriptively* presented. Having a control group allows the retailer to isolate the effect of the promotion and thus *diagnostically* understand the causal effect of the promotion. Several weeks of experiment data can be extrapolated to *predict* the value of new promotions in the long run and used in demand forecasting. Finally, within a promotion management tool, a *prescriptive* algorithm could be built that suggests the optimal promotion strategy based on the trial, and if combined with intricate ML tools, made *autonomous* and self-supervised.

Ranging across the continuum, CEs should be viewed as a tool that at best aids managerial decision making and other methods of creating value with data and analytics. By no means seems CEs as a panacea to retail optimization, but rather a method that needs careful thought behind the process that in certain cases might prove beneficial. The practicalities that

experimenting necessitate seem rather challenging, which might explain the modest recognition of the methodology in business settings.

In Figure 2, the suitability characteristics, requirements and application areas of CEs in retail context are synthesized based on theoretical findings in sections 2.1 through 2.5. A retailer needs to have a holistic comprehension of all three in order to make decisions on whether an experiment should be undertaken or not. Understanding the enablers of the business is obviously important in all decision making, but in order to make successful experiments, retailers need to recognize which characteristics allow them to thrive. The requirements and considerations need to be addressed to ensure robustness and risk mitigation. Finally, rarely does a decision maker manage all parts of the operations, and hence the sphere of influence plays a major role in which application areas experiments should be considered. The characteristics presented in each part of Figure 2 are not exhaustive, but rather the commonly understood and discussed ones based on the reviewed literature.

3 Methodology

In this chapter, the research methodology is discussed. First, research design and process is introduced. Then, the means of collecting data are presented, continued with discussion about the analysis, and finally validation of the selected methodology and results.

The thesis follows design science principles where the proposed design framework is synthesized and validated through phases of data collection and subsequent validations. This process is depicted in Figure 3.

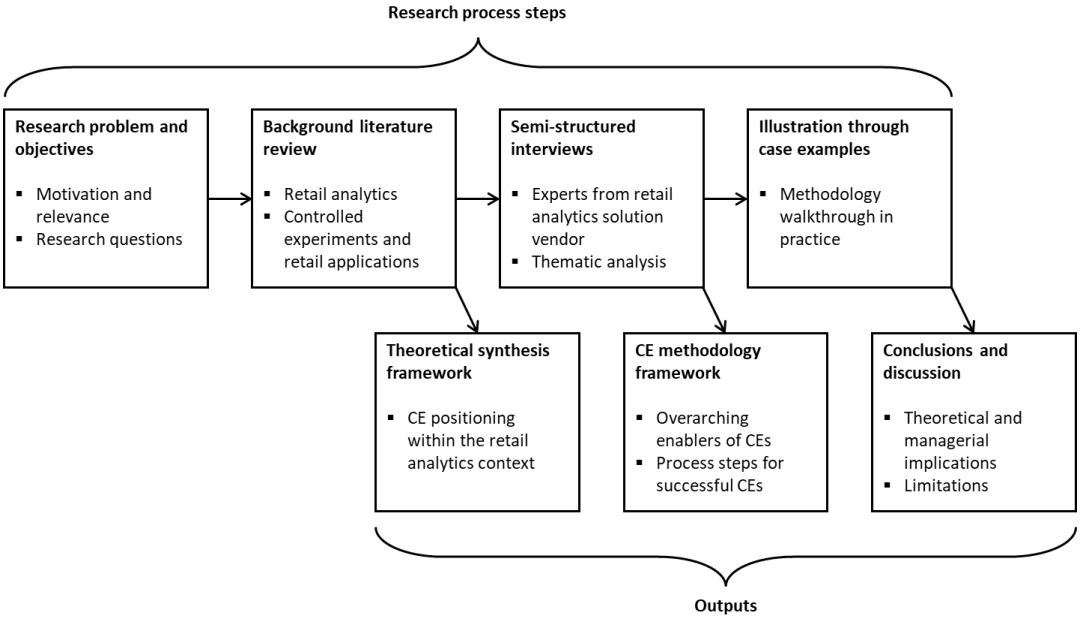

Figure 3: Research methodology summary

3.1 Research design

Whereas natural science research approaches that either build or test theories are concerned with explaining the current state of things, there is room for science that rather tries to address how things should be in artificial settings (Simon, [1969] 1996). In problems that are vaguely structured and associated with technological problem solving in practice, Holmström et al. (2009) call for a design science approach that instead of explanation touches the research problems from an exploratory point of view. Saunders et al.

(2012) divide all reseach designs into exploratory, descriptive and explanatory studies, pointing out that exploratory studies ask open questions and adapt to whatever data is obtained. Exploratory studies are hence typically qualitative.

As the reseach problem of this thesis has more relevance from a practical perspective and the research questions direct the approach to a more conceptual output, a qualitative design science methodology serves the purpose well. What makes design science particularly relevant in a solution development research is the possibility to bridge the gap between practical matters and theoretical understanding (Holmström et al., 2009) and its innate characteristic of creating new and innovative artifacts (Hevner & Chatterjee, 2010, p.5). Oftentimes the term *design science* is used concurrently and overlapping with many other rubrics, such as *action research*. Holmström et al. (2009) point out that any nomenclature can be used, but for instance *action research* must have a clear focus on the design and implementation of an artifact in order to be considered design science. In this thesis, the term design science is used throughout as the output is predominantly a concept design.

As Holmström et al. (2009) mention, design science – particularly at its solution incubation stage – incorporates scientific reasoning outside the generally more traditional deductive and inductive approaches. In particular, abductive reasoning, first introduced by Peirce (1878) as an alternative to deduction and induction, emphasizes a more creative and intuitive angle on making scientific conclusions (Kovács & Spens, 2005). But as methods of reasoning discussed by Peirce (1878) are more or less philosophical in nature and presented through syllogisms, a more practical view on abduction from a design process standpoint is needed. Dorst (2011) proposes a formula for how the different reasoning methods solve the following pattern:

$$What_{(Thing)} + How_{(Working\ principle)} \rightarrow Result_{(Observed)}$$

In deduction, the *result* is unknown but it can be obtained as the *what* and *how* are known. In induction, the *how* is unknown and the researcher has to propose a mechanism or law in order to explain the observed *results*. Together, deduction and induction form a loop of discovery and justification, respectively. But in design problems where the desire is to create value, the equation can be reworked, as follows:

$$What_{(E.g.,object,service,system)} + How_{(Working\ principle)} \rightarrow Value_{(Aspired)}$$

Typically in this formulation, both the *how* and the *value* are known and understood, but the artifact (*what*) has to be created in order to solve the problem. This is what Dorst (2011) calls *Abduction-1*. As to the research problem introduced and examined in chapters 1 & 2, *Abduction-1* functions as the fundamental logic of this thesis;

$$What_{(CE\ methodology)} + How_{(Causal\ inference)} \rightarrow Value_{(Optimized\ processes)}$$

where the exact essence of the *what* has not been yet discovered based on the literature review in chapter 2. A controlled experiment methodology is formulated as a practical framework in section 4.2 to represent the *what* and define the solution proposition as the main output of this thesis. Sections 3.2 and 3.3 touch on the rigor of the data collection and analysis, and research methodology, respectively.

3.2 Data collection and analysis

In addition to the examination of literature on retail decision making and optimization procedures including CEs, semi-structured interviews were conducted to obtain practical understanding of the themes. Interviews were held with experts from a retail optimization software company that specializes in analytics and provides solutions to major retailers globally. Interviewing experts from a specialist solution provider can be justified with the opportunity to combine extensive industry expertise with understanding of analytics and data as tools for decision making and optimization. The company has been contemplating developing an experimentation capability on top of its current offering that comprises supply chain, merchandising and staff optimization solutions for retailers. Still, controlled experiments as an offering is merely a concept, and not widely discussed within the organization. In total, 8 interviews were conducted, summarized in Table 2.

Semi-structured interviews were chosen as the primary data collection method as they give flexibility in approaching informants by allowing the restructuring of questions as well as omission of some whenever suitable (Saunders et al., 2012, p.374). As the point of these interviews is to gain understanding of the processes between decision makers and analytics systems, the interview questions are more open-ended and thus require more room for open discussion and context understanding. Harvey-Jordan & Long (2001) emphasize thorough literature review and planning of interview topics and subtopics to ensure that semi-structured interviews add knowledge. Chapter 2 provides adequate prior research to function as the

theoretical background that helps conduct the interviews with the solution vendor experts.

Table 2: Summary of the solution vendor expert interviews

Informant	Area of influence	Region	Timing
A	R&D, data science	Europe	4/2023
B	Business strategy	Europe	4/2023
C	Product management	Europe	4/2023
D	Product management	Europe	4/2023
E	Customer project management	Europe	4/2023
F	Business consulting	North America	4/2023
G	Business consulting	Europe	4/2023
H	Sales	North America	5/2023

Saunders et al. (2012, p.377) suggest that in exploratory research, either unstructured or semi-structured interviews should be used, with the former being more common. However, as the nature of this thesis requires rigorous understanding of solution requirements and human decision making from an extremely multifaceted optimization perspective, the interviews must have some structure to avoid derailing the dialogue to topics that are of little or no relevance to the process.

The interview sample was selected with versatility and experience in mind. All of the interviewed participants had 10+ years of experience in the retailing industry, from practitioner, consulting or software solution provider side, oftentimes a combination of at least two of these. A notable emphasis was placed on previous positions as retail decision makers to avoid merely examining the solution provider's point of view that might cause bias. Within the company, several responsibility areas and departments were represented, as can be seen in Table 2. As the company has not provided straightforward solutions to business experimenting, the informants could be approached with neutral questions, and subsequently their thoughts and

ideas have less bias towards any direction. Interviews were held either physically on-premises or remotely, depending on the location and other practicalities. All interviews were recorded, and confidentiality and anonymity was assured to make the dialogue more open and relaxed (Saunders et al., 2012, p.389). Interviews lasted for about an hour, with varied sequence and duration on the questions. The interview template can be viewed in Appendix 1.

Analysis on the interviews started already beforehand with understanding the solution vendor's business and offering, and the interviewees' responsibilities, areas of influence and experience. These were then also discussed in the beginning of the interview to adjust how the individual should optimally be approached. In addition to recording, the interviews were transcribed so common patterns were more easily found. As the research methodology is a combination of design science, abductive reasoning and qualitative data from semi-structured interviews, no specific data analysis method fits the purpose perfectly. But from different methodologies of qualitative data analysis, thematic analysis fits the best as it suits semi-structured interview data well, as Braun & Clarke (2012) describe with a case example. Braun & Clarke (2012) remark that thematic analysis is flexible and offers avenues for providing qualitative results that are easily comprehensible for wide audiences. For a design science study with an intentionally broad scope, this goes hand in hand with the goal of providing an exploratory solution concept that solves a practical research problem.

Thematic analysis has been widely used for a long time, but its prominence can be associated with the highly popular Braun & Clarke (2006) paper on using the methodology in psychology studies. In the paper, the authors present a six phase guide for qualitative data analysis with thematic analysis. The phases are described in Table 3. This process was used in analyzing the interview data. Section 4.1 presents the process and results of the analysis.

Table 3: Phases of thematic analysis, adapted from Braun & Clarke (2006)

Phase	Description
1: Familiarizing with the data	Transcription, re-reading, initial ideas
2: Generating initial codes	Coding relevant features across the dataset, matching data and codes
3: Searching for themes	Compiling codes into wider themes
4: Reviewing themes	Validation of themes w.r.t. the codes and the entire dataset
5: Defining and naming themes	Building the narrative and descriptive definitions of themes
6: Producing the report	Providing examples, discussing the analysis w.r.t. the literature and research questions

3.3 Methodology validation

Hevner et al. (2004) emphasize the iterative and incremental nature of design science. While several papers (e.g., van Aken, 2004; Peffers et al., 2008; Holmström et al., 2009; Saunders et al., 2012) merely touch on this cyclic and additive characteristic of design science from a retrospective field-testing point of view, an argument can be made that even in the early stages of solution development, an iterative process is of valid use. Kovács & Spens (2005) point out that abductive reasoning through simultaneous data collection and theory development is very common in action research. As the abductive research process differs from purely deductive and inductive counterparts by incorporating this kind of subsequent data collection, analysis and design, it can be concluded that for a practice-oriented design problem, reflective iterations through phases of data accumulation is worthwhile to ensure valid results.

The iterative essence of this thesis is applied in the data collection, design and validation that happen in phases. First, a literature review is carried out in sections 2.1 – 2.5, after which a general understanding of the topic of discussion, the problem field and relevance is formed in section 2.6. This notion is then assessed by the interviews with retail solution vendor experts. After these two rounds of data collection, the framework for CEs in retail operations is constructed in section 4.2. After this pivotal step, the

framework is even further validated with case examples in section 4.3. Consequently, several layers of validation is present in the research process, increasing the cogency of the results.

Hevner & Chatterjee (2010) discuss the qualities of design science in information systems research (p.9) and collect several frameworks (p.23) for the research process, for instance the popular design science research methodology (DSRM) devised by Peffers et al. (2008). The DSRM consists of six activities that in sequence serve as the backbone of design science process: *problem identification and motivation*; *define the objectives for a solution*; *design and development*; *demonstration*; *evaluation*; and *communication*. Common for all the research frameworks is that they strive to ensure quality and pertinence in design science research that – without such guidelines – could be at risk of producing unrealistic and impractical outputs lacking academic rigor. Most noteworthily, Hevner & Chatterjee (2010) return to the design science research guidelines (p.12) first introduced by Hevner et al. (2004). Table 4 presents these guidelines and discusses them from this thesis' point of view.

As can be seen from Table 4, this thesis can fulfill all guidelines and the methodology serves the purpose of producing a valid solution to a practical problem. The guideline framework by Hevner et al. (2004) and its derivatives are well established in design science literature (e.g., Gregor & Jones, 2007; Peffers et al., 2008; Gregor & Hevner, 2013). Hence its use as the main validation tool for this thesis is justified. Also, as the design problem of this thesis only covers the solution incubation (Holmström et al., 2009), without the practical solution refinement through real life implementations and pragmatic analysis, the guidelines fit better than e.g., the DSRM framework that more strictly calls for a real demonstration before evaluation and communication. This kind of demonstration is outside the scope of the thesis, and a valid potential avenue for further research, as discussed in section 5.3.

Table 4: Design science reseach guidelines, adapted from Hevner et al. (2004)

Guideline	Description	Thesis contribution
1: Design as an Artifact	Production of a viable artifact e.g., construct, model, method, instantiation	A framework for practical application of CEs in retail
2: Problem Relevance	Production of a technology-based solution to an important and relevant business problem	A rarely seen way of making profitable business decisions in a low-margin, high-volume industry
3: Design Evaluation	Rigorous evaluation of utility, quality and efficacy of the design	An iterative data collection and evaluation process for the design artifact
4: Research Contributions	Clear and verifiable contributions in the design artifact, foundations and/or methods	Bridging the gap between a methodology and a practical application area
5: Research Rigor	Reliance on rigorous methods in the construction and evaluation of the artifact	Scrutiny through DS principles and frameworks, constant validation
6: Design as a Search Process	Utilizing available means to reach desired ends; satisfying laws in the problem environment	Literature review and expert interviews shaping the artifact design and governing the requirements
7: Communication of Research	Presentation to both technology-oriented and management-oriented audiences	Discussion (chapter 5) including both, publication of thesis

4 Findings

This chapter presents the results and findings of the empirical part of the study. Common themes found in the interviews are presented in section 4.1, followed by a practical framework on retail CEs in section 4.2. Section 4.3 further examines the conceived framework through illustrative example cases. Finally, when moving to chapter 5, the findings are synthesized as the implications of this study.

4.1 Interview analysis

The interview transcripts were analyzed in a spreadsheet containing manually picked quotations that were of meaningful interest. Most of the quotations were then transformed into a code, characterizing phase 2 of the thematic analysis methodology presented in section 3.2. In total, the 8 interviews produced 275 codes. Some of the codes, while particularly interesting, were not directly related to a research question or question from the interview template (Appendix 1), which speaks for the flexibility and benefit of semi-structured interviews.

The codes were first compiled into five common themes, and after a review as per phase 4, two of the themes were united due to heavy overlap, ending up with four wider themes. Afterwards, in another review round, some codes were further moved to another to distinguish themes more clearly. These review rounds refined the overall intention of each theme. Table 5 portrays the process in which the four themes build from the codes. Only the most relevant codes are displayed for comprehensibility. The four found themes, in no particular order, are:

- Modern analytics are growingly helping retailers across their business
- Retail operations have plenty of opportunities for CEs
- To make the most out of CEs, various questions need to be addressed
- A standardized retail CE solution has potential, but proving value is difficult

This section discusses the themes within the overarching context presented in section 2.6 together with excerpts from the interviews.

Table 5: Codes and themes from interviews

Codes	Themes
Analytics types do not work in silos Descriptive analytics is the golden level Practitioners don't want their "art" to be disturbed Focus shift from revenues to cost efficiency Solution adoption is easier when organizational structures and governance do not change Modern tools break data silos and integrate business areas Lock-in to a single solution vendor Systems and processes are typically old and inflexible Predictive and prescriptive models are already developed and proven Many retailers are fascinated about modern analytics Technology maturity is generally low Improved future prediction instead of repeating old habits More automation across the value chain	Modern analytics are growingly helping retailers across their business
CEs have clearest potential with areas where humans make intuitive decisions Successful anecdotes exist of CEs in marketing, merchandising and supply chain Planogram changes Store layout changes Assortment management Promotion optimization Pricing strategies CE on software implementation as a PoC Program roll-out to understand demand and customer demographics Logistics, delivery schedules	Retail operations have plenty of opportunities for CEs
If a retailer has centralized processes, reluctancy to experimenting is lower Technology maturity key in how largely experiments can be benefitted from Retailers might not have competent personnel to supervise CEs CEs done today are typically narrow and siloed, value comes from integration CEs need to comply with the business model and organizational structures Possible conflicts of interests between stakeholders A platform would ease experimenting process ERP systems need to be configured, sometimes impossible Who within the organization gets informed about ongoing CEs needs consideration Experimenting between stores is more complex than between consumers	To make the most out of CEs, various questions need to be addressed
Hard to demonstrate value because the problems are arbitrary Great opportunity in being a thought leader against competitors CE solution cannot be sold as generic, it must be specified and definite Anecdotes and case studies are persuasive in proving value A solution would require approaching right decision makers Selling the idea would require training consultants New solutions must be co-developed with a pilot customer Retailers are different, thus a solution needs much configurability CEs could function as the extension of scenario simulations The more disturbance experimenting causes to current systems, the less tempting it is Testing on small groups to understand the impact could add plenty of value	A standardized retail CE solution has potential, but proving value is difficult

Modern analytics are growingly helping retailers across their business

On a general level, retailing has seen significant progress in how data and analytics are used to gain insight into the business. As maturity levels vary greatly, the level and speed of adoption of new solutions is also mixed.

Informant A pointed out that analytics types do not work in silos, but rather they overlap and at best organizations can benefit from all types at the same time. Less mature players can first gain large benefits from the less advanced analytics, whereas more mature ones can adopt more innovative tools. All in all, new tools, most prominently predictive and prescriptive tools, allow circumventing traditional decision-making processes and react quicker to new trends and external disruptions. For instance, informant F reminded that the pandemic brought the importance of lead times and safety stocks into the light and raised the awareness of long-term planning and preparedness.

> *I feel like almost everybody is getting okay at [AI & ML] in a kind of short term span, but to be better at more long-term buying and to be leveraging more future leading data on things like promotions, on what promotions to schedule, how to promote them… …leveraging more future looking tools to try to determine what's the right thing that's gonna drive consumer behavior as opposed to just repeating "what we've done in the past". – Informant H*

Regardless of the pandemic, the overall attitude of retailers has shifted to a more proactive and intricate direction. Traditionally, retailers have had issues finding competent personnel to oversee the operations and utilize software solutions. Even today, as raised by informant C, new solutions applying predictive and prescriptive analytics require help from external consultants who tailor the solutions to the retailers' processes. Overall, the focus today is not just to drive revenue, but also to reduce costs, and informant E pointed out that it is not either or, but actually some cost optimization actions might also increase sales such as improved replenishment causing better shelf availability.

> *When I started [20+ years ago], the mindset was that there is no more room for improvement. You need a different approach to realize that there is still potential to cut millions in costs, and recently there has been a change of mind in optimizing*

everything, because in retail margins are so small. – Informant E

There are still some barriers to adopting novel tools and analytics. Often, the people managing retailers' merchandising or supply chain have decades of tenure and thus the processes are deeply ingrained to the organization. Decision making and organizational structures are typically very hierarchical and anything disrupting the status quo is contested. The element of trust becomes challenging when a data-driven solution recommends actions in areas where the manager would normally make intuitive decisions based on their experience.

Another significant challenge for making the most out of new innovations is the lock-in to systems and certain software vendors. Building a partnership with a vendor for over a decade hinders the opportunities to look outside for competing solutions. Retailers might realize that a change would ultimately be for the better, but do not still see it worthwhile to move away from the existing solutions due to seemingly substantial costs and risks.

You can think of [decision making] as art vs. science. In [supply chain optimization] you can get close to 100% science, but in other parts [e.g., merchandising] it is more art. But even in these areas it would be strange if things could not be automated further. – Informant B

Retailers struggle with multiple overlapping and imperfectly integrated systems that aid different areas of the business. For instance, supply chain management can itself have several platforms and tools that together make the architecture complex and burdensome to manage. Informant G discussed that a retailer can have over a hundred modules containing different processes and data. The interplay of these is a main challenge for both the retailer and the vendor that tries to optimize the processes. Also, organizations are still siloed and information does often not fluently flow across different functions. Modern systems are integrating decision areas and shifting decision making to a more holistic direction, which together with the increased speed of analysis makes organizations more agile.

How we breakdown the silos of retailers, I think will be something that is facilitated by the [advanced analytics] tools. – Informant D

> *I think there is a shift in collaborative data and collaborative teams and really kind of it's more of a gray area of when you do business analytics. That's so much broader now than just being someone who is looking at, say, sales and inventory. It really takes into account what does that basket look like? What does the marketing and promotion results look like? Where is halo and cannibalization[4] happening? – Informant F*

Informant A described prescriptive and autonomous analytics as the *golden level* that all organizations should strive to reach. The world is seeing AI and ML develop with a speed that some find frightening, but retailing has its own characteristics that slow the advancements down. As discussed in section 2.1, the fundamentals of retailing will still remain the same and rather than revolutionizing retail completely, new innovations help execute the same processes with more ease, insight and speed. Retailers are often extremely large and slowly moving organizations, and protecting the critical daily processes often surpasses making disruptive changes in priority. But the field is also extremely competitive and consumer loyalty is important to attract repetitive buying. This creates pressure to adopt new technologies and stay ahead of the curve.

Retail operations have plenty of opportunities for CEs

Continuing the findings of sections 2.5 and 2.6, retailing seems like an area where CEs could be applied in many use cases. Decision-making and optimization has been proven to be effective through experimenting. Informant F had previously worked at a retailer in a managing position and conducted CEs that helped the enterprise reduce its inventory levels and increase revenue by several percentages. The experiments were managed and analyzed in a spreadsheet and reported to the executive level. The following anecdote speaks for the fact that even non-scientifically conducted experimenting without complex tools can be effective as long as the

[4] Halo effect is defined as the increase in product Y's demand due to an increase the demand of product X. For example, a promotion on laptops typically increases the sales of complementary accessories such as laptop bags. Cannibalization means the opposite; an increase in product X's demand decreases the demand of product Y. For example, a promotion on a certain minced meat product decreases the sales of substitute varieties of meat products.

experimental design and overarching business objectives are defined adequately:

> *I've done it predominantly based around inventory levels. I typically took a subsection of stores and then a subsection of products. So, I would take a large format store, a medium format store, a small format store. I would take a high volume store, medium volume store and a low volume store and take some sort of blend of those six options. That would become essentially my test cases and I would go through those and I would make changes to their assortment, delivery schedules, safety stock values. Make changes to just anything and run through that and see ultimately with the goal being "Was I able to reduce the inventory but increase the sales?" – Informant F*

Aligning with the application areas recognized in scientific literature and textbooks, the informants mostly agreed with the potential areas where experimenting could provide value and alleviate risks associated with making chain-wide decisions hoping to bring a positive impact. Opportunities are available throughout the value chain, ranging from supply chain to merchandising to staffing.

Most noticeable potential was by consensus seen in assortment management, promotion optimization, pricing and space planning. Informant F thought the easiest potential is in supply chain, as the excerpt above illustrates, because supply chain personnel are typically intermediaries in the organization whereas e.g., category managers or marketing department are typically detached from the store operations where the experiments would eventually happen. This, most interviewees see, is however not that large of an issue in the future as the solutions that are used in management of different areas become more all-encompassing and integrated.

> *I could see [controlled experiments] on a new floor plan for a store of the future kind of layout that goes hand in hand with the different planograms that you'd be setting with that… … I could see it in the [supply chain & merchandising] side, both from supporting our new features, but then also well, what happens if we introduce a new product line to a subset of stores? What's the impact of our forecast moving forward?… …I could see it in the promotions. What if I introduce in a brand-new promotion type or update my loyalty program, how is that going to impact my*

> *overall demand… …and then laboring into workforce, what happens if I change some major strategic aspect in my workforce. How is that going to impact? It's just different use cases feeding into each scenario. – Informant H*

> *So if you are able to create some scenarios: What if you push this [order] today or not? So this can be impacting also the buying and the external suppliers. Or to rearrange external logistics for example. And this can also be taken into account in the warehouse themselves. So for pick and pack. Is this really necessary to ship this today? Can I do it tomorrow? So there are a lot of cases where this can bring value when it works properly. – Informant G*

Interestingly, as the interviewees discussed the topics also from the vendor company's perspective, another level of opportunities with CEs in retail was seen in how the company can prove value of its products to retailers that either are interested in purchasing new capabilities or already using them.

By conducting proof-of-concepts (PoC) of the capabilities that a software solution has, the vendor can effectively convince the retailer about the tangible benefits that it can get by purchasing the solution. This can be beneficial in all stages of the partnership; demonstrating value to enable sales, confirming solution fit and understanding pain points during the implementation stage, and testing new features to persuade upsell opportunities after initial setup.

> *When we implement our solutions, we could even more accurately follow the project. Today, we follow the directions to which KPIs are moving… …but for long rollouts we could use better control groups, as typically we just compare to historical numbers… …This could have even larger benefit for us than [developing an experimenting product per se]. – Informant B*

From the retailer's perspective, this kind of thinking should not only be limited to new software initiatives, but also changes to any part of the business model. Informant H had history at a retailer, being in charge of a new program in which a new category of products was rolled out to generate vertical growth in an area that was previously unheard of in the field. The program was first implemented in a small heterogenous group of stores, and when a comparison was made, the retailer was able to conclude the demographics in which the new category affected the most, and eventually

make the decision to expand the scope to a national level in the particular demographic stores. This kind of culture of experimentation, as also discussed in e.g., Thomke (2020), makes the organization much more innovative and helps it find out the consequences of decisions without full commitment of resources.

In retailing, this opportunity to gain understanding of the cause and effect of actions with smaller risk is the essence of experimenting. Some application areas are cheaper, easier to play with and quicker to adjust, while some, for instance store layout changes, are expensive and slow to implement. This does not necessarily imply anything to whether a CE makes sense or not as a validation method. An expensive change such as a layout or a planogram change will anyway yield costs and have uncertainty, so a CE will at least alleviate the risk and provide hints to how to proceed on a larger scale.

In retail, CEs are not as much a tool to play around with, as in online examples where players such as Amazon and Google make thousands of UI experiments daily (Kohavi & Longbotham, 2017). Informant A discussed that in retail, CEs are rather more unique schemes that need careful considerations. A retailer cannot operate many simultaneous experiments because of the risk that an experiment has implications on another experiment. Moreover, changes in retail are more expensive to implement compared to UI variations and thus a retailer has to be more careful in which things to experiment with and which not. This characteristic is what calls for a wider understanding of the requirements for using the CE methodology in a retail context.

To make the most out of CEs, various questions need to be addressed

Continuing from retailers needing to understand that CEs are more suited as thoroughly planned projects rather than random acts of curiosity, retailers need to understand what is being tested and which results indicate a positive impact.

In the example of informant H rolling out a new category in only the subsection of stores in which it made sense, understanding the strategic big picture was key to reap the ultimate benefits of the program. The retailer did not have to make an all-or-nothing decision, and the new category could be seen as a natural vertical expansion for the retailer.

Another example of this can be seen in a sales PoC case in which the vendor and the retailer validated the benefits of taking a planogram solution into use for a selected category:

> *We had maybe about 15 stores that we were looking at from 300 stores. So they [the retailer] were looking to make these changes in a small proportion, then extrapolate out and both validate the tool and use these results to say well, have they got an opportunity with their drinks categories for further assessment, further changes being made in both the overall space allocated to those categories and then subcategories within them… … And in this case [the retailer] then took [the tool] into use and we worked it through for a couple of years with their macro space team. – Informant D*

Experimenting means in any case deviating from the existing ways of operating. Therefore, the way an experiment reflects to the customer perception needs to be considered. A retailer needs to understand its positioning in the market and the value proposition it has. Informant C discussed this in the following excerpt. In the same citation, informant C illustrates why retail stores pose a more challenging arena for experimenting than trialing between individual consumers.

> *Probably easiest place to see [experimenting] is in customer loyalty. We gave these customers a coupon, we didn't give those customers a coupon. Comparing these there's it's easier to do because there's less noise around the customer potentially and there's bigger samples. So you can do a 10,000 customer sample whereas you be doing a 10 store sample. So, they would know which customer responded better to a monetary discount… …What is generally hard to do is if your proposition is universal. It's very hard to trial pricing if all your stores hold the same price. It's a really interesting topic because should retailers be getting into not charging the same price everywhere, or you charge the same price in 90% of places? There's always a bit of difference or it's broadly the same price, but you create instability. – Informant C*

Existing processes and relations are often the results of years of development, and there are multiple stakeholders involved indirectly in decision making. Informant E pointed out that oftentimes assortments are contractually

agreed on with the supplier, which means that the retailer cannot make arbitrary changes to it. Typically in these cases, new product introductions are just plainly pushed in predefined clusters without even a possibility of trialing with them.

More generally, maturity of processes and system architecture strongly dictates the range in which an organization can conveniently conduct experiments. Informant A lamented that even some large retail chains manage critical decisions with spreadsheets, and informant E gave a practical example of many retailers still today practicing supply chain replenishment with haphazard methods instead of basing it to recent demand forecasts. The IT landscape in this kind of players can be scattered and stiff, inevitably narrowing the practical potential to adopt a culture of experimentation.

> *I know the five biggest retailers in Germany and I know they their IT systems and setups in total. The IT architecture and systems they're using are not capable to do [experiments]... ...[Retailer X] has had a custom developed solution for over 30 years and the solutions are not capable to do such kind of analysis in an easy way. – Informant G*

Retailers have various operating models and organizational structures. As discussed earlier, retail enterprises often fail in facilitating information flows and collaboration across different parts of the business, for instance marketing decisions are hardly ever discussed with supply chain directors and vice versa. This naturally affects the prospects of experimenting as a tool that drives the company forward, since experiments in one area have an impact on another ones. A trial on a new promotion will affect demand and thus reflect in the forecast and replenishment of said products and ultimately supply chain KPIs. Communication needs to be transparent and different internal stakeholders must be incorporated in collective decision making to avoid conflicts of interest. When informant F conducted CEs at the retailer, store managers did not always even know about an experiment, but sometimes there needed to be a heads-up for the stores that e.g., inventory levels would seem notably lower for a period.

> *If the retailer is centrally managed, it is much easier to conduct experiments but if the chain is decentralized, it is difficult to govern which stores participate; it wouldn't happen naturally between the stores. – Informant A*

> *But I think a lot of [controlled experiments] that are done today are very narrow and only based on the objectives of the team. Implementing that decision, not necessarily what it means to store operations or to, you know, let's say it's category manager, you know to supply chain or to the commercial team or whoever's not making the decision. And I think [integration] could be a real benefit that goes out there. And because it's very poor at the moment and it goes back to retailers being quite siloed, being quite traditional in their decision making and not necessarily thinking about the big picture as such when they go to make a change. – Informant D*

A standardized retail CE solution has potential, but proving value is difficult

The consensus between the interviewees was that a software solution for CEs could be relatively easily developed and made configurable to various customer needs and use cases. Even the prescriptive part of recommending actions based on experiment results seemed rather trivial as the company has already quite advanced capabilities in making automated suggestions based on historical data and forecasts. But how to position a CE capability with respect to the existing features was not clear, which makes sense as this concept has not been widely discussed within the company.

All in all, the informants were in unison about the need to have easy configurability and flexibility that builds on top of current solutions. Informant H thought of it as a natural continuum to scenario simulations, and several others pointed out that each experiment should first be simulated in the software before commencement. However a solution is potentially built, it needs to have a clear scope of capabilities and a clear message. Informant F noted that there is a risk of overwhelming retailers if a software solution has no focus and no clear value message.

> *Why are we not offering [experimentation as a solution] yet? I think the biggest reason is that we have not yet identified where the biggest value is. Saying generically that we can help experiment with anything is too vague, it has to be a specific solution, like "we can help you see how it affects sales if we move bananas to another shelf in a randomized set of stores". – Informant A*

> *In most cases, I think that retailers assume that return on investment will be bad due to high rampup costs [of experimenting]. And bypassing existing structures, for example you would have to configure a permanent experimenting structure in your category management tool or space planning tool in order to apply experiments to them. I just think it's too much fuss over marginal resources. – Informant E*

Because the statement of the problem can be defined in practically limitless ways, a tangible value for a solution is hard to quantify. Retailers are eager to make decisions based on assumed return on investment (ROI), and thus the financial aspects of the entire course of an experiment needs to be inbuilt in a solution.

Approaching the right people in the retailing organization is also key to assert the nature of experimenting and the tangible benefits of the methodology. Building and presenting case studies is a cogent way of demonstrating the value of a product, but the process is slow as the case study would have to be agreed on with a pilot customer that agrees to help build the practices and logics to the software. But the strength of CEs as a tool is that the strongest arguments against it tend to be that human intuition is enough. Informant E remarked that in retail, decision making is often intuitive because the people in charge are consumers themselves, too. Still, retailers are keen on understanding the tangible value of actions, and as science shows, CEs are a powerful method to convey it. Creating reference cases is vital to productize and sell CEs as a solution.

> *Retail loves anecdotes. They love case studies, they love making things tangible and real and simple, you know, [for example] 0.2% uplift on this and everything's gotta go through an internal capital process to get signed off and needs to have a business case associated. – Informant C*

> *It's almost like our [classified product]. We had a [sales] case recently where we had a really hard time demonstrating measurable value with the exposure, with basically licencing the module, because we don't know exactly what the problem you're trying to solve is. So I feel like it's a very bespoke challenge of: "Are you introducing a new set of of consumer products? Are you introducing a new store of the future?" The different challenges are going to result in different tangible financial values. But I*

> *think from a soft benefit standpoint, this kind of [experimenting solution] would provide flexibility to implement new programs that are gonna drive value down the road for the retailer. So it's almost an enabler for whichever programs that they are ultimately going to implement. – Informant H*

From the solution vendor company's perspective, differentiating from competitors is critical to win new customers and keep existing ones happy with the solutions. Sales cycles in which a retailer tenders solutions from several vendors always have a strategic fit aspect, and for a retailer that cultivates curiosity and understands the fundamentals of experimenting as a information collection method, an offering that enables CEs to be conducted on top of all daily processes could have potential. This also requires identifying the right people at the retailer who to sell the idea to, and also internal development of best practices around the product. Informant H mentioned that currently their consultants working with customers would not have the sufficient knowledge to assist retail customers with CEs.

> *Being able to be a thought leader and go beyond scenarios testing, but true A/B in market testing I think would give us a discrete advantage and show us that we're kind of moving past where a traditional players sit in the space. So yeah, I think there's a great opportunity. – Informant H*

Interview summary and ramifications

Retailing seems like a dispersed industry in terms of adoption of analytics to support decision-making. Maturity levels play a heavy role in how quickly and effortlessly a retailer is able to reap the benefits and accumulate ROI with new tools and corresponding processes. It still seems that CEs could have potential regardless of the technological maturity, partly due to the vast opportunities to utilize them for, but also due to the challenges of CEs being more related to organizational characteristics rather than technological issues. Hence, and because the purpose of this study is not to concentrate on technological maturity but instead on CEs as a concept, the following sections focus on the essential questions that need to be addressed in order to benefit from them. As the interviews brought light to also a solution provider's perspective, the ensuing parts will consider the concept on a more general but practical level in the form of a methodology framework.

4.2 Controlled experiments' methodology in retail

This section continues upon the findings from chapter 2 and the results and inferences from section 4.1. As mentioned before, a practical methodology for CEs in a retail context would be of relevance for retailers, and as section 4.1 shows, a data-driven software vendor could also benefit from it. Section 4.1 provided several interesting and crucial aspects of what to take into account when considering CEs as an optimization methodology in a retail context.

As discussed already in chapter 1, the scope of the thesis is intentionally wide due to retailing sector being extremely heterogenous, and thus the outputs of both section 4.1 and this section are more overarching rather than specific for a certain type of retailer. This does not necessarily deteriorate the applicability of the framework since there does not seem to be a widely understood collection of the most fundamental aspects of CEs in a retail context. Every retailer would in any case have to adjust the process to its peculiarities that either enable or limit CEs from happening successfully.

The following methodology framework is designed with a practical approach, without a commitment to the viewpoints of either retailers or software product vendors. It is loosely based on the guidelines for designing an experiment by Montgomery (2013, p.14) in Table 1. The framework postulates that the retailer is to a large extent centrally operated, meaning that the planning is done at the chain level and store managers have limited governance in customizing their stores. For a retail chain that is decentralized through e.g., a franchising business model, CEs become more difficult as the processes cannot be overseen by a centralized function that can dictate what happens across the chain and who gets to be a part of any new initiatives. Technically, franchisees could organize CEs together without centralized functions, but getting enough stores involved and administrating the process would most likely be difficult and slow.

Figure 4 presents the overview of the methodology, and this section discusses the parts of it in detail. A detailed summary of it is further depicted in Figure 5. Section 4.3 focuses on the practicalities of the methdology in more detail through fabricated case examples that follow the process from beginning to end.

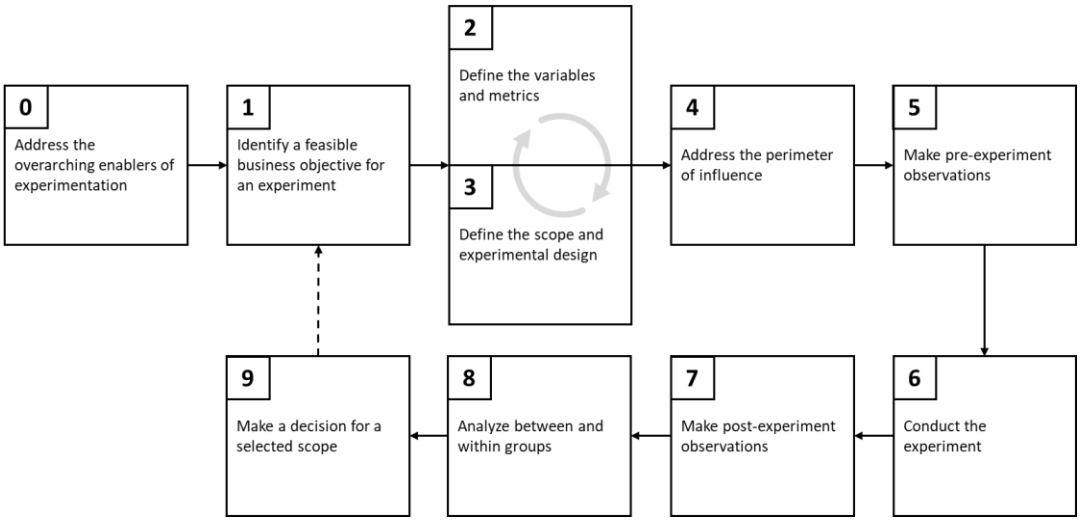

Figure 4: General overview of the retail CE process framework

In the framework, process steps are numbered from 0-9. Step 0 encompasses the overarching enablers in the retailing organization for conducting experiments. Steps 1-9 are the practical steps to be taken in CEs, in the numeric order with the exception that steps 2 (*Define the variables and metrics*) and 3 (*Define the scope and experimental design*) are in conjunction, meaning that they form an iterative process where the results of one affects the other. Proceeding to step 4 (*Address the perimeter of influence*) can thus only happen after the iteration and validation of steps 2 and 3 together. As step 9 (*Make a decision for a selected scope*) is the last one, indicating the ultimate decision based on the experiment, there is an option for follow-up run of the experiment or an iteration of the entire experiment. Thus, there is a dashed arrow from step 9 to step 1 (*Identify a feasible business objective for an experiment*) making the process iterative in situations that call for it.

0: Address the overarching enablers of experimentation

As discussed in the previous section, there is a plethora of issues to be considered if a retailer would like to utilize the science-esque CE as a way to gain knowledge about the business that typically is far from scientific and more leaning towards art. Each organization is unique and CEs can be conducted in certainly many ways and contexts, but the fundamental qualities that make the organization more capable of mastering the concept and gain benefits from them can be considered rather universally. These

enablers are not necessarily mandatory, but rather something a retailer should keep in mind to make the processes more effective and efficient.

First, retail organizations are typically large, and an experiment has effects on several managers and employees' field of responsibility. Therefore, the organization needs to promote and spread knowledge and logical understanding of what CEs are and why anyone, let alone the company, could use them. Otherwise, there is a risk of reluctance and dismotivation for the changes that experiments trial with. Second, and related to the conceptual understanding of CEs, the company culture should encourage curiosity and breaking the status quo if there are ideas on how improvements could be made. This is by no means an easy shift in any organization, as Thomke (2020) also describes. Entire new models of leadership need to be embraced, but once a culture of experimentation exists, the hurdles to internally regard CEs become lower.

Third, the organization needs to make sure that its internal communication is active and that the departments whose operations interplay with each other are not siloed in decision making. The same applies to data flows between systems that together build the IT architecture used in managing the operations. Modern ERPs are increasingly consolidating data sources and realms of analysis to break the walls between organizational sections. Fourth, the planning levels within the retail organization need to be aligned and the information not only needs to flow horizontally as described above, but also vertically from strategic to tactical to operational planning levels. Experiments typically happen in the operational or at most tactical level, from supply chain optimization to assortments to promotion management to name a few areas. But an experiment and its implications need to comply with how the retailer conducts business and what its strategic aspirations are in a much larger picture. This stipulates the executive and director levels to actively communicate the company vision and direction to lower levels, and to stay aware of the elementary processes happening at the core of the business.

1: Identify a feasible business objective for an experiment

To start the process of a CE, a distinct business opportunity should be discovered and formulated. However relevant the scientific method typically is as a means to gain understanding in a retail context, a research question and a hypothesis need to be well articulated before an experiment can happen. An argument can be made that an explicit hypothesis that captures

the identified opportunity is the foundation of all experimenting. The power of CEs lies in the precision of creating an arrangement where the isolation of a variable produces results that are not approximate but in fact precise. Kerkhove (2022, p.247) emphasizes that the hypothesis must be *specific* and *measurable*, going as far as calling it *impossible to design a successful experiment based on a vague hypothesis*. For example, a bad hypothesis could be "Coupons drive revenue", whereas a better example would be "Offering n coupons with promotions $[p1, p2, p3, \dots]$ generate XM€ sales uplift and YM€ profit uplift during the period".

When the research question and hypothesis are formed, the decision maker needs to consider whether experimenting is the right way to answer them. Sometimes there is beauty in simplicity and a rigorous experiment would only slow down an opportunity to react quickly to unplanned phenomena[5]. Also, in many cases making minor tweaks to drive sales and cut costs are low in risk and costs, and thus an experiment might not be worthwhile.

At this stage of planning, the financials of the experiment need to be addressed. Linking to the business objective, the expected business value of the treatment must be considered and mapped out. Both how the treatment is expected to drive revenue as well as how it affects costs and margins have to be taken into account. A CE might have notable fixed costs that are always part of a larger budget, and the person in charge of the experiment has to be able to provide a tangible benefit estimation that is plausible and fits the strategy that the business function executes. When an estimate is prepared and the experiment is in line with what the company strives to achieve in a bigger picture, an initial decision must be made whether the experiment will be conducted, scrapped or postponed. If the decision is to proceed with the experiment, another issue to consider at this stage is how the results of it will be utilized in decision making. For instance, a decision tree on how to continue with different results should be drawn already at this stage, since

---

[5] In May 2023, the Eurovision song contest roused massive fanaticism in Finland. The Finnish contestant wore a bright green outfit which inspired thousands of fans to adopt the color green in all kinds of objects. Many stores initiated successful campaigns to promote green products that actually had nothing to do with the event. Testing with a CE whether a color associated to an event leads to increased demand of items in that color would not have made sense in this example as the boom was very short in duration. However, for recurring color-specific events such as international sports competitions or Christmas, a CE could provide informative results on how to plan e.g., merchandising and promotions.

without such a process the intuition of the manager in charge might trump the data, eventually dismissing the entire idea of a CE as a methodology.

2: Define the variables and metrics

After considerations on the overall objective and feasibility of a CE, and a decision to execute the plan, the experiment needs to be carefully designed and planned to fit the purpose and be as effective as possible in addressing the hypothesis. This step of the framework forms an iterative process with step 3, where the outcomes of each influence the other, and thus the sequence is not linear, but rather iterative before moving on to step 4. First, the treatment variable has to be specified. Depending on the use case, a treatment can be a major change, for instance a vertical expansion to a new business area or a store layout change. But the treatment can also be a minor adjustment that retains the general status quo, such as a new product within a category, a new promotion or a more/less frequent ordering cycle. Regardless of the magnitude of the treatment, it has to be planned carefully to enable isolating its effect in the treatment group, and to ensure results that can be discerned from natural noise in the data.

The experimenter also needs to be aware of everything else that varies in the experiment setup as well as during time. Variations can be internal (individual test units are heterogenous such as different size or customer base) or external (test units are exposed to different occurrences such as events and weather). Large enough sample sizes and time periods, and randomized samples usually account to both internal and external noise, but the experimenter needs to be aware of them since modelling and controlling for all variation within the world we live in is unrealistic.

At the same time, the level on which the experiment is conducted will be clarified. As traditional retailing is a difficult context to experiment between individual customers, the most trivial level is having the test groups be clusters of stores, as this allows homogenous groups. Having homogenous groups is the *control* of *controlled* experiments, and the only way of achieving this is retail is either test between large enough sample of customers, or homogenous clusters of stores with time. Geographical sampling, testing features between product groups, and within-store testing of time-dependent actions such as shelf filling schedules are all examples of test setups that might provide interesting data, but lack the homogenous control that truly isolates the effect.

For causal models such as the CE, the independent variables are connected to a dependent variable that is the eventual point of interest within the experiment, effectively the measure of success. In a retail context, this variable is typically one or many KPIs that are of interest to the manager overseeing the experiment. For a case of a new pricing strategy tested within certain product groups, the experimenter can, for instance, track the total revenue, profit margin and profit margin % within the product group between treatment and control groups. For a case of trialing with different safety stock levels, the experimenter should look for changes in inventory value, inventory turnover, availability and stockout rate. The selected KPIs to measure should be in line with the overall business objective of the experiment, as well as the performance measures used outside the experiment. Also, the extent to which KPIs correlate with each other needs to be understood, because otherwise false conclusions could be made about what the mechanism is through which the KPI has changed due to the experiment. For instance, increased availability drives revenue growth but at the same time might increase inventory levels.

3: Define the scope and experimental design

If the treatment is done on product or category level, the scope has to be considered. For example, a new pricing strategy has less risk if it is being implemented in only a few categories, but the scope needs to be wide enough and representative of as large population of products as possible to make results more applicable. The attributes of products dictate the scope in which the CE should be applied, but also the general applicability of the treatment regardless of the CE.

In the example of a new pricing strategy, there has to be a rudimentary understanding of the price elasticity of demand of products to make sensible decisions. For instance, products with high brand loyalty typically have a low price elasticity and thus trying new pricing policies has less risk. By contrast, a pricing strategy that seeks to reap larger margins from widely available and substitutable commodities such as dairy products might not be successful due to consumers' tendency to select the cheapest alternative in any case.

Next, the experimental design is selected. The most common approach would be the *pretest-posttest control group design* (Campbell & Stanley, 1963) but some situations call for a more complex setup. For instance, some experiments can test multiple adjacent treatments. Another possibility is to incorporate time as a factor in not just making pre- and post-experiment

observations but to see effects at multiple points or rotational periods. Generally, the simpler the experimental design is, the less risk it has. The experimental design should then be converted into a statistical hypothesis based on the KPIs that were selected in step 2. The corresponding analysis method needs to be planned and formulated to fit the selected metrics, scope and experimental design.

At this point, the basis of what the experiment is about are formulated, and next, the test groups need to be formed. This process preferably utilizes randomization, as with large enough samples it ensures homogeneity and thus allows much easier statistical methods to analyze the data. For groups formed of stores within a retail chain, a combination of randomization and selection by a variable can be the best approach. This is called stratifying (Kerkhove, 2022, p.255), and in a retail context it could mean for instance dividing all stores first into large, medium and small size stores, and randomizing within these categories so that treatment and control groups get stores from all three. Full randomization assumes that the sample sizes are adequately large, which might not be possible. The number of units in a group has to be in balance between practicality and statistical power, as too large groups only increase costs and the workload. In the first place, the advantage of CEs is that only a sample of the population can yield data that can be extrapolated to represent the entire population.

During this step, the timelines of the experiment are defined. An experiment needs setup time before it can successfully start. These might be even months depending on the experiment; a store layout change program requires considerable resources in not just money but also time. Also, a normalization period, in which the treatment is already applied but the experiment has not yet commenced, makes sense in many cases. In major changes such as the aforementioned store layout change process, comparing just recently revamped stores to untouched ones will provide biased data since consumers are typically eager to try novel experiences and thus the footfall in the changed stores is already larger, leading to an unfair comparison. On the other hand, with minor changes such as introducing a new product, it might take time for the customers to react to its availability and thus a short normalization period could be justified.

Also, the experiment test periods need to be decided based on how long time is seen as sufficient to acquire enough data to make conclusions. In the literature examples discussed in section 2.5, the periods were in the range of a week to a year. Generally, the longer the experiment is, the more

statistically significant results it can be expected to yield. But in many cases, it is not reasonable to stretch the test period as this slows down the decision-making process unnecessarily. The whole idea of a CE in a retail context is to balance between making timely decisions that react to the external environment and use statistical rigor by applying an isolating methodology.

As to the treatment and the groups formed based on it, having exactly two groups is not necessary, but there can actually be several simultaneous treatments, as long as the sample sizes are large enough to allow this. A controlled experiment like this can be built with the same idea than a two-group experiment, so that treatment groups can also be compared against each other on top of comparing against the control group. Another way is to rotate the treatments within the groups in a predefined schedule, i.e., a Latin square design, as discussed in section 2.5 with the Brunk (1953) and Cunningham & O'Connor (1968) examples.

The iterative essence of steps 2 and 3 is in how the outputs of step 2 affect the decisions made in step 3 and vice versa. The treatment, variables and KPIs all affect how the scope should be selected, since the scope of the CE influences heavily the results that the experiment produces. The KPIs that are tracked as per usual practices might restrict the opportunities to design experiments within a certain business area. There might not simply be tools and processes to support making a treatment and following it through.

To the opposite direction, the experimental design affects how the test will be monitored and which variables can and must be accounted for. Also, a decision on the scope can change the perspective of which KPIs should be tracked and how weighty the treatment should be compared to the control. Furthermore, the selected time frames for the experiment might impact which KPIs make sense to monitor. Overall, going from step 2 to step 3 has to be followed by reflecting the outputs of step 3 in how the treatment, variables and metrics are planned. If there are changes in either step, the other one needs to be inspected before moving on.

4: Address the perimeter of influence

Steps 0 through 3 are more or less general planning that happens on the experimenter's desktop. After these crucial steps the action starts by factoring in the stakeholders that the experiment influences. Responsibility areas need to be clearly mapped and understood within the personnel running the experiment as well as the people touched by it. As retailing and

its underlying processes are extremely multifaceted and different parts of the business have direct impacts on the others, the experimenter has to consider all other areas of the business that the experiment will reflect in.

For instance, a category manager wanting to trial with assortments has to be aware that supply chain management will see the changes and that there might be conflicts of interest in the KPIs that each department follow. An expanded assortment will at the same time expand the inventory value, and while the category manager might see this as perfectly fine as long as the revenues are increasing, an inventory manager might not be happy with having to cope with larger balances.

The communication between these kinds of stakeholders needs to be open and the message clear about what is the ultimate objective of the experiment and why all stakeholders should cherish it rather than looking at their own siloed areas of interest. When the communication is open and stakeholders are kept informed, the ultimate decision after the experiment has still to be made, and this governance is a question that might require interference from a director that oversees the different business areas. Otherwise, a conflict of interest might cause friction that only hinders the experiment in the first place.

Not everyone in the organization needs to be aware of an experiment if the experiment itself allows this. Not only does making blind experiments reduce the needed communication and change management, but it also ensures that stakeholders touched by the experiment do not influence the results by knowing that there is an ongoing experiment with a certain goal.

Informant F had been conducting CEs at a retailer, and oftentimes the stores acting as the treatment group units were not told about an experiment. Only when the CE clearly had a visible impact on e.g., the inventory levels, did F inform the store. Transparency is rarely a bad thing, but if the retailer is centrally managed and the experiment is approved by a person that has overarching power, it is for the good of the entire chain to let experimenting happen even if some stakeholders such as store managers might be left outside the communication loop. If a CE is successful, it might be beneficial to announce its existence to spread awareness and show an example. After all, the whole idea of why chain wide decisions are made and operations are optimized is to drive benefits across the business and stay ahead of competition.

Another facet of communication and awareness is the one between the retailer and the customer. Since customers are ultimately the ones generating the data that becomes the experiment results, informing them about an ongoing experiment could distort their behavior and deteriorate the public image of the retailer since it could no longer be seen as equal and fair. It is essentially the same in online A/B testing; users are presented with a version of the UI without explicitly stating so, and whatever happens afterwards is just a piece of data for the company to analyze. Thus, retail customers should be kept oblivious of any experimenting that – while affecting them – does it implicitly.

5: Make pre-experiment observations

To be able to witness any transformations that a treatment might have, a baseline needs to be set for the figures to allow comparison with post-experiment data. Thus, before a CE is commenced, a thorough scan and documentation of all relevant KPIs and financials is needed. At this point the treatment and control groups are already determined, so the review gives an option to re-evaluate whether the groups are homogenous enough, or if adjustments are needed. Also, the aggregate figures give an opportunity to make more tangible preliminary estimates than the ones made in step 1. Overall, the pretest observations should be considered both on an aggregate basis and test unit basis. When the experiment is conducted at a store level, it is fairly straightforward to obtain individual store data and KPIs, and analyze if there are notable deviations lately that might need to be flagged beforehand.

The units to be used in hypothesis testing have to be coherent and fit for the purpose. For instance, monitoring the daily average of a KPI needs consideration on the length of the period of which the average is calculated.

6: Conduct the experiment

Once the previous steps are taken, the experiment is prepared and a possible normalization period has elapsed, the experiment can begin. Throughout the course of the experiment, which can be weeks or months, it is important to ensure that the operations are disturbed as minimally as possible other than the experiment treatment. It is impossible to avoid changes completely, but having large sample sizes alleviates the noise.

Notably important is that there are no overlapping experiments that could contaminate each other. The experiment supervisors must also watch the

fidelity to the experiment and take actions to correct anything that diverges from the agreed procedure.

During the test period, continuous data should be collected, and any noteworthy deviations must be recorded for further examination. For instance, a single store might have a notably larger footfall due to an unforeseen local event. With large group sizes and long time periods, the effect of this kind of deviations will eventually decrease but for further analysis conducted in the later steps, it is important to track all abnormalities.

7: Make post-experiment observations

After the experiment, similar data and KPI review has to be taken as in step 5, with all post-experiment counterparts to the data points captured pre-experiment. At this point, the aggregate results can be easily compared with the pre-experiment results, and the found deviations should be accounted for, whether the experiment was affected by internal or external variation that skews the results.

8: Analyze between and within groups

At this part the data-heavy analysis will be conducted. First, the hypothesis testing defined in step 3 is conducted to obtain the main results between the groups. The results should rather straightforward indicate whether the treatment was successful or not, and by how large a magnitude. But even if the results seem good when reflecting on the hypothesis, the experimenter needs to consider the extending implications of the experiment, going back to step 4 in which all impacted stakeholders with their own KPIs were addressed. If there are no glaring conflicts of interest in KPIs across the perimeter of influence, the experimenter should go on with extrapolations of the results to obtain an estimate of the ROI of the experiment.

After the between-groups analysis, the experimenter should dive deeper into the data and inspect any within-group points of interest. While the treatment and control groups are at best very homogenous on aggregate levels, the groups themselves contain units that might differ by various dimensions and characteristics.

A within-group analysis seeks to understand which demographics responded particularly well or badly to the treatment and why. For instance, a new promotion for take-away coffee bundled with a breakfast snack could end up

being successful in stores that are located near city centers, but ineffective elsewhere due to different commuting patterns. Therefore the, analysis at this step should not only cover the data but also take a step deeper into the features that explain the data. This helps planning the ultimate intervention based on the experiment.

9: Make a decision for a selected scope

After the statistical analysis between and within the groups is complete, the experimenter needs to address the research question set in step 1. If the results are in line with what was hoped for, the experimenter has to consider whether the treatment would at this point comply with not just the business objective of the experiment, but also the strategic objectives of the business as a whole. Good numbers are valuable only when they can be seen as organic results of what the company is doing right. Through the course of an experiment, there are plenty of opportunities to gain understanding from not just data but also the people that supervise certain areas of the business. Any decision, be it backed up by an experiment or not, should take a holistic approach, and in retail this is even emphasized since it is characterized by operations that require meticulous interaction between its parts.

The main reason to perform a within-group analysis is to understand the scope in which a treatment is eventually valid to implement. There are essentially three types of decisions to make: [1] For the treatment group, which units should retain the treatment, [2] For the control group, to which units should the treatment be added, and [3] For units outside the experiment, to which ones should the treatment be added. Sometimes these decisions might not seem unambiguous even if a decision tree with thresholds was drawn already in step 1. This might call the experimenter to only partially accept the results, and for the rest conduct a follow-up run. In this case, the entire process pivots back to step 1, as the steps might need readjustments upon the initial experiment run.

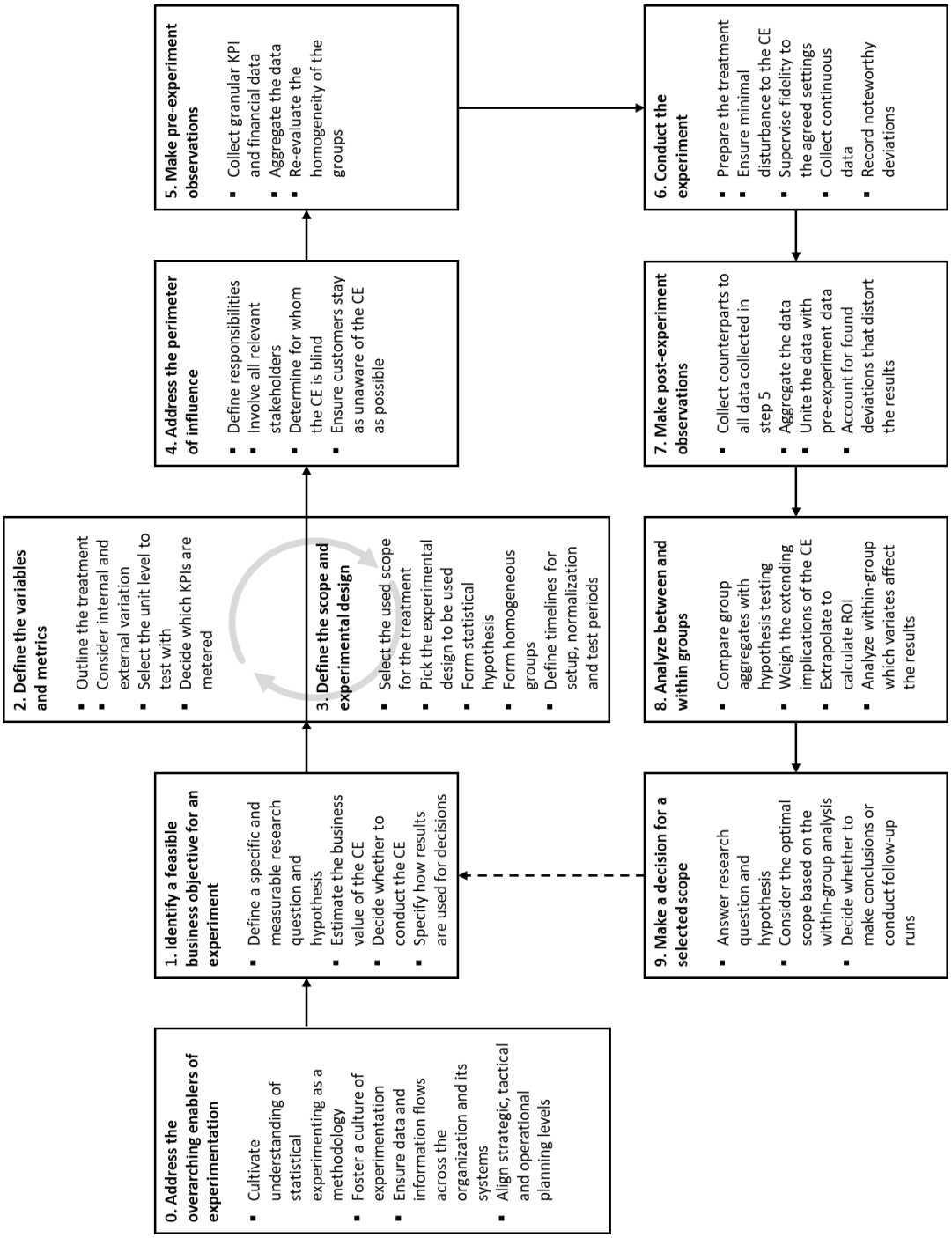

Figure 5: Detailed summary of the retail CE process framework with overarching enablers (0) and process steps (1-9)

4.3 Illustrative review

In this section, two fictional case examples of the methodology presented in section 4.2 are described and discussed. These examples depict how retailers could use the framework to plan and execute a CE that helps them to make better decisions on discovered business opportunities. As with the framework in general, these cases assume centrally managed organizations in which a centralized function manager can make widespread decisions on selected scopes of stores and items. Furthermore, step 0 of the framework (*Overarching enablers of experimentation*) is not covered, as the cases are fabricated, and it would not bring additional value to the methodology validation to address general characteristics of the imaginary companies. Focusing on steps 1-9 brings concreteness to the case and how a CE can be planned and executed in practice.

Case: ValueGrocers

This case example walks through a grocery supermarket chain experimenting with a new promotion type. ValueGrocers operates 400 supermarket stores in North America, having a focus on a wide selection of products for different customer segments and ranging from fresh produce to meat, bakery, canned food, beverages and snacks, home and personal care products and pet supplies. Its business model focuses on competing by providing low prices and frequent discounts on its products both to all consumers and especially to loyalty program customers. Nearly half of its revenue is promotion-based, which means that margins are thin and often promotions are sponsored by the suppliers. This kind of funding drives promotional activity and provides mutually beneficial opportunities for the parties involved.

So far, ValueGrocers has had promotions that are based on a discount percentage, giving percentage discounts where either the percentage itself (e.g., 30% or 50% off) or the outcome price (e.g., $1 or $2) seem attractive. Now, the company has been considering adding promotions that more effectively drive the quantities sold. One way to push larger sales quantities is seen as to offer multi-buy promotions where a discount would only be valid when several products are bought together.

Being unsure how this type of promotion performs for ValueGrocers, the vice president of merchandising has decided to pursue the concept with a CE that tests multi-buy offers against the uplift that conventional percentage discounts have had. For the character of a multi-buy offer, they have

concluded that the offers would be most logical as [buy X for $Y] instead of [buy X get X+1] or [buy X get Y% off] since the majority of promotions and the pricing strategy overall favor even numbers. Therefore, this multi-buy type is seen as the best brand image fit that customers would likely feel the most comfortable with.

For step 1 of the section 4.2 framework, the research question is stated as: "Do multi-buy promotions generate more uplift on revenue than percentage discounts?" A research hypothesis can thus be stated as "Multi-buy promotions generate more revenue uplift than percentage discounts". Statistical hypotheses are defined in step 3 because the outputs of steps 2 and 3 heavily influence the statistical analysis that is conducted to obtain valid results.

For step 2, the vice president of merchandising considers this promotion type treatment as a direct substitute to the traditional percentage promotion that can be applied in certain stores to calculate its revenue uplift. The KPIs that are followed are thus the revenue uplifts on product, product category and complementary category levels, with the product uplift being the most relevant. Making this KPI choice goes with the assumption that the baseline revenue is standard and that everything else affecting the final profit margins stays the same. Considering product, category and complementary category levels helps understanding the halo and cannibalization effect that the promotion might have.

For step 3, the merchandising team selects a line of canned foods and a line of personal care products as the scope, as these are ones that typically benefit from multi-buy because they are practically non-perishable and generally cheap commodities that consumers tend to stockpile in case of a tempting offer. As the uplift effect of the percentage promotion type on the selected scope is known beforehand, the experimental design should be selected so that a comparable uplift figure for the new promotion can be obtained. Hence, a two-group controlled experiment is selected where the new promotion type is applied to a group of stores and omitted in a control group of same size. Campbell & Stanley (1963) call this design the *pretest-posttest control group design*. A statistical hypothesis that seeks to give answers to the research question with this experimental design can be formulated as:

$$H_0: \mu_X - P\mu_Y = 0$$
$$H_1: \mu_X - P\mu_Y > 0$$

where μ_X and μ_Y are the mean revenues for the multi-buy promotion and no promotion, respectively, during a period similar to and as long as the experiment period. P is a constant denoting the uplift $[(100 + a)\ \%]$ of the conventional percentage discount known prior to the experiment. The alternative hypothesis H_1 is a one-sided hypothesis, which means that H_0 will not be rejected if hypothesis tests indicate a smaller uplift for the new promotion type (Montgomery, 2013, p.38). Hypothesis testing for this kind of settings can be done with a two-sample t-test (Montgomery, 2013, p.38), and for this example the test statistic can be formulated as follows:

$$t_0 = \frac{\bar{X} - P\bar{Y}}{S_p\sqrt{\frac{2}{n}}}; S_p = \sqrt{\frac{(n-1)s_X^2 + (n-1)P^2S_Y^2}{2n-2}}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means for the treatment and control group revenues, respectively. S_X^2 and S_Y^2 are the corresponding sample variances and n is the sample size of the groups. Here, due to the hypotheses having a constant multiplier P for μ_Y, the sample mean $\bar{Y}$ must also be scaled with P, and the sample variance S_Y^2 with P^2. This hypothesis testing scheme assumes randomly selected groups of units and identical variances for the samples X and Y. For the former, randomizing the samples is possible since the store base is wide. The latter can also be assumed for simplification, even if in reality stores within a retail chain are always somewhat different.

As to selecting the most reasonable sample size n, there is no general formula, since it is a fine balance between statistical power and practicality. Montgomery (2013, p. 45-47) suggests that one way to determine proper sample sizes is to compute power curves which plot the statistical power $(1 - \beta)$ against sample sizes n for certain differences in means $\frac{|\mu_X - \mu_Y|}{\sigma}$, where β is the probability for type II error $P(fail\ to\ reject\ H_0 | H_0\ is\ false)$ and σ is the population standard deviation. The larger the desired difference in means i.e., the effect size is, the smaller the sample sizes need to be in order to achieve a certain power. There are several statistical packages that can be used in this kind of plotting for balancing the tradeoffs between the significance level α, effect size, statistical power, and sample size.

Another aspect to consider at this stage is the timelines for the experiment. A promotion typically only lasts for some weeks, and thus to best find out the effect of the promotion type, the merchandisers see a realistic duration as the

best fit. Typically, its promotions last for two weeks, but for an experiment it wants to ensure enough data so a three-week experiment period is justified.

As to stakeholder management in step 4, the experiment is supervised by the vice president of merchandising, who together with the merchandising team manage the initiative and make the ultimate decisions. The most relevant stakeholders outside merchandising are the suppliers for the products in question, and demand and operations planners. Financial planners are also incorporated as the promotions are funded by the suppliers. The stores involved in the experiment are also informed about a test of a new type of promotions for certain products. An experiment like this has very low risk of becoming distorted due to store personnel being aware of it. Thus, ValueGrocers has opted to be internally open about the experiment. Naturally, customers are being kept unaware of the experiment itself, while advertising the promotions in the affected stores as per ordinary practices. This experiment is conducted outside the loyalty program to ensure larger amounts of POS data.

Steps 5 and 7 (pre- and post-experiment observations) are taken with the company's typical practices from the tools that contain all relevant records, mainly from ERP and financial planning tools. Conducting the experiment in step 6 starts with preparing the promotions in the stores, marketing them in the relevant channels and configuring the POS system for the new type of promotion. During the experiment, the focus is to monitor the sales figures and ensure that the supply chain complies with the demand. Because the sample sizes are relatively large and each group consists of similarly sized stores with geographical distribution, the team can be fairly confident that both internal and external noise is minimized, and the experiment can bring valid results to be interpreted.

Moving on to the aftermath of the experiment in steps 8 and 9, the team starts with making group comparisons on the chosen statistical hypothesis. The effect size that it looks to see between the groups has been determined in step 3, and this has affected the sample sizes and statistical power, too. At this point, the between-group analysis is highlighted by the product-level hypothesis testing, but as the vice president also wants to understand the effect on category and complementary category levels, hypothesis testing is done on these levels as well to complement the main results.

Besides hypothesis testing, at step 8 the extending implications of the experiment need to be considered. Going back to the stakeholder

management in step 4, the merchandising team needs to address how the KPIs of other areas were affected and if the treatment was successful as a strategic fit. Also, the analysis should at this step drill down to within-group level, to determine which stores responded particularly well to the new promotion type. Parameters relevant in this analysis could be for instance the areas in which the store is located, store size, customer demographics and proximity of competitors. As promotions play a fundamental role in ValueGrocers business model and multi-buy offers are a rather distinctive way to operate, the vice president sees it as a strategic move to either implement the new promotion type in all stores or none, while the percentage discounts can still remain as an alternative.

Case: PrimePharma

This case example examines a European pharmacy chain experimenting with a vertical expansion with a new product category. PrimePharma operates 100 pharmacies that are small in size. It offers prescription and over-the-counter medicine, alongside various health and personal care products. Its strategic focus has been on specializing in high-quality products and expert customer service. Thus, its price point is relatively high, and it only has occasional discounts and campaigns.

PrimePharma wants to seek vertical growth by expanding its offerings to areas that it sees as natural extensions to its current portfolio. During the past decades, the demand for fitness-related products has been rising, and fitness supplements are nowadays very popular from professional athletes to occasional exercisers. The vice president of strategy sees a potential growth area in fitness supplements that it could offer for its customer base that values curated brands and high quality. Unsure whether this product category is profitable, they look to explore it with a CE to test the success of introducing the new category. The research question is formulated as "Should we expand into offering fitness supplements?" From this, a research hypothesis is derived as: "Fitness supplements generate XM€ total revenue annually while being a profitable category that increases total footfall."

The independent variable being a binary decision on whether to include fitness supplements in the assortment for a certain store in the experiment, the company has several metrics that it hopes to impact with it. For KPIs in step 2, the strategy team decides to track the new category revenue and profit margins (€ and %) to determine answers for the research question and hypothesis. It also wants to understand how the new introduction affects its

total capability to attract customers, and thus it decides to monitor total footfall, which it also sees as an adequate proxy for total sales growth in areas outside the new category.

For step 3, a decision is made to experiment with a rather narrow scope of fitness supplements from one supplier, as it eases the supply chain and contractual rigidity. While the supplier could offer a vast assortment of supplements, PrimePharma starts with only the most common products, such as protein, pre-workout supplements, creatine and vitamin-mineral supplements. Similarly to the ValueGrocers case, the classic *pretest-posttest control group design* (Campbell & Stanley, 1963) is selected as the experimental design as it is easy to implement in a context that is conceptually simple.

Providing variety with the ValueGrocers case, this example will utilize a difference-in-differences (DiD) regression as the statistical hypothesis testing methodology. A two-sample t-test would eventually yield the same results, but for the sake of demonstration, a regression is an interesting alternative. The DiD concept builds around the idea of measuring the formula:

$$\left(Treatment_{after} - Treatment_{before}\right) - \left(Control_{after} - Control_{before}\right)$$

(Huntington-Klein, 2021), which can more mathematically be written as:

$$[E(Y|T = 1, P = 1) - E(Y|T = 1, P = 0)] - [E(Y|T = 0, P = 1) - E(Y|T = 0, P = 0)]$$

Where Y is the selected response variable, T is a binary variable; 1 for the treatment group and 0 for the control group, and P is a binary variable; 1 for post-experiment and 0 for pre-experiment data. As a regression equation, this logic can be written as:

$$Y_{it} = \beta_0 + \beta_1 T_i + \beta_2 P_t + \beta_3 T_i P_t + \varepsilon_{it}$$

Where β_0 is the intercept term, β_1, β_2 and β_3 are the regression coefficients for the variables and ε_{it} is the error term for the observation at the particular group i and time t and. β_3 corresponds to how much larger Y is for the treatment group after implementing the treatment (Huntington-Klein, 2021). In other words, it is the causal effect that can be attributed to the treatment.

In this case, the merchandising team is interested in understanding the effect that the new category has on revenues, profit margins and footfall. Therefore,

the response variables should be chosen accordingly. For instance, with revenue and profit margin, Y could be formulated as €/day per store, and with footfall, Y could be formulated as the number of individual customers/day per store. Now, the statistical hypotheses can be formulated as:

$$H_0: \beta_3 = z$$

$$H_1: \beta_3 > z$$

where z is the chosen threshold value that is seen as significant enough to satisfy the research objective. Technically, z could be set to 0, meaning that within the selected significance level, any improvement would lead to rejecting H_0. But in practical terms, this is usually not enough and there should be a target value for each KPI. For instance, with the financial KPIs, all costs associated with taking the category into the assortments should be considered. The revenues and profits need to be large enough to not just breakeven but to actually have a notable profit & loss impact.

The DiD approach assumes that the control and treatment groups are subject to the same internal and external variation, and that they would have followed similar trends had the treatment not been applied. Again, the effect of noise can be minimized with proper sample sizes that can be determined with power curves, as explained in the ValueGrocers case example above.

Step 4 consists of responsibility and stakeholder management, and the vice president of strategy sees the category manager of health products as the best fit to supervise the experiment. The experiment requires close collaboration with demand and operations planners and other category managers. The category manager must also be in close contact with the vice president to align business objectives and make sure the new initiative makes strategic and financial sense and moves the company to the right direction. Another main stakeholder is the marketing department which has to make sure that customers are aware of the new category.

Stores themselves are not necessarily mentioned about a CE, but rather the experimenting team communicates about a new category being rolled out cautiously prior to larger decisions. As with the ValueGrocers case, customers are kept fully blind from the experiment. The company does not see it as a risk to discriminate some cities or regions by not initially launching the new category. It conducts the experiment to reduce the risk of making rash decisions and to understand the performance of a new category. Contrary to

the ValueGrocers case example where it most likely either fully adopts the new promotion type or discards it totally, PrimePharma has to be more careful about which stores and demographics would best benefit from the expanded assortment.

Pre-experiment observations are taken for the selected KPIs and standardized to a unit that can also be monitored after the experiment, such as the average revenue in €/day per store. The experiment includes monitoring the footfall impact that the new category is expected to cause. For the footfall measures, the company has been using infrared sensors to get accurate numbers. Post-experiment observations in step 7 contain the corresponding revenue and profit margin figures on both category and product levels.

The experiment has a fairly long setup time, with the supplier negotiations, supply chain preparations and marketing efforts taking weeks ahead of the launch. Similar to the ValueGrocers case, during the experiment the category manager monitors the sales figures and collaborates with supply chain management to ensure supply and demand are in line. During the experiment in step 6 the category manager might also visit the stores to observe the personnel point of view and oversee that the experiment moves ahead accordingly.

After the experiment, in steps 8 and 9 the analysis focuses on the profitability of introducing fitness supplements as a new product category. The analysis starts with running the DiD regression with a capable software, producing the coefficient values with associated p-values. From these the KPI-specific hypotheses can be tested, leading to the general findings of the experiment. These results are then extrapolated to estimate the ROI for the initiative, by also taking into account all fixed costs.

On top of the general regression analysis, the team should have substantial emphasis on the within-group analysis to understand why in some stores it might not be profitable to start selling fitness supplements. Compared to the ValueGrocers promotion type example that covers a more universal part of how to operate a retailing business, this case example is more granular; different stores can and should have different assortments based on how well they perform.

General remarks on the controlled experiments' methodology

Altogether, the process framework is designed as a practical methodology that fits the *what* in the *Abduction-1* pattern described in section 3.1:

$$What_{(CE\ methodology)} + How_{(Causal\ inference)} \rightarrow Value_{(Optimized\ processes)}$$

The framework is a product of iterative data collection and validation; first a literature review and synthesis, after which round of interviews and qualitative data analysis. From a design science point of view, this methodology bridges the gap between theoretical understanding and a vaguely structured technological problem that by this point has not seen a pragmatic approach. Generally, design science with qualitative interviews as the main data collection method yields unique results as the data is ambiguous and the research process requires creativity. This might intuitively degrade the academic reliability and generalizability, but in its defense, the overall purpose of design science is different than in traditional sciences that rely on deductive and inductive reasoning and lack the need for creative problem solving.

Reflecting on the design science research guidelines depicted in Table 4, the process so far only lacks some parts of guideline 7 (C*ommunication of research*). This section has contributed to guideline 5 (*Research rigor*) by validating the methodology steps with practical examples. The case examples are not exhaustive illustrations nor do they fully embody the issues to address with controlled experimenting in a retail context. Nonetheless, the methodology does not fail in illustrating the sequence and most important parts of each step. The case examples should not be taken literally, but rather as rough validation of the methodology's applicability in two distinct use cases.

Other case examples could look far different and shift the methodology's focus to other directions. Namely, the cases had an impact on the final design, as they shed more light on the importance of a specific and measurable hypothesis that is formulated upon steps 2 and 3, and statistically tested in step 8. But the cases were made generic to illustrate that the framework delivers a purpose while being flexible.

5 Discussion

This chapter concludes the study and synthesizes the main findings. The research questions posed in chapter 1 are addressed along the theoretical implications in section 5.1. After this, in section 5.2 managerial implications are discussed, from both retailer and solution provider perspectives, as this twofold aspect was also apparent in the empirical part of the study. Finally, section 5.3 addresses the limitations of the thesis and motivates further prospects for academic exploration within the area of CEs in a retailing context.

5.1 Theoretical contributions

It is rather interesting that controlled experimentation has not received widespread interest in the retail research world. As an application area, it is almost as good a match for the methodology as possible due to abundance and variety of data, customer-centricity and a constant need to improve and optimize. The objective of this thesis was to bridge this gap and provide practical understanding on how CEs can and should be used in aiding decisions that typically are made intuitively and backward-looking. This chapter answers the research questions and discusses their implications.

> *1: How do modern analytics help retailers make better decisions and optimize their processes?*

While retailing as a field of industry is in constant flux and new innovations change all parts of the value chain, the fundamentals of what retailing is, remain the same. Modern analytics, ranging from descriptive analytics all the way to autonomous analytics, help retailers optimize their processes and improve parts of the business where maturities vary broadly. Many retailers manage their day-to-day processes with spreadsheets, while others have adopted autonomous ML models to rationalize the operations ranging from forecasting to replenishment, assortment planning, promotion optimization and pricing. Generally, it seems that the more eager a retailing company is to purchase solutions from specialized vendors instead of developing tools internally, the better results it can expect to achieve in the long run.

At the baseline, demand forecasting drives all analytics, and the adoption of tools that utilize sophisticated forecasting algorithms is accelerating, leading to better compliance of supply and demand. Everything else must essentially be based on the demand forecasts, including the planning of assortments and

promotions. Even marketing needs understanding of the demand, and having integrated systems that consider all parts of the operations helps close the gaps between siloed parts of the business, of which marketing is typically a great example. Importantly, the coordination of strategical, tactical and operational planning levels is critical in the realm of retail analytics, as the boundaries shift and the cooperation between them strongly dictates how retailing will evolve in the future.

> *2: How do controlled experiments fit the landscape of data-driven retail analytics?*

Controlled experiments can be applied in nearly any use case to give causal inference on a treatment, but in a retail context they have a strong linkage to the other analysis and decision-making. Extant literature shows that the practical use cases for which CEs can be used extend across the value chain, from pricing, promotions, assortments, supply chain configurations and store displays. Regardless of the maturity that the retailer has, CEs can be an effective way to isolate an effect that wants to be examined.

Fitting CEs in a collection of analytics systems that a retailer uses is not as burdensome as it might sound. The retailer – like with any analytics – has to have a clear business objective to strive for, but with the difference that a CE needs more thorough planning and better collaboration. Understanding the potential that CEs can bring to a retail decision-maker has been low in the academia, as the methodology has merely been seen as a way of doing research rather than business.

CEs have relevance on all types of analytics in the analytics continuum. With the recent trend of widespread generative AI, statistical data-analysis will perhaps get less overall interest in the academic world in the near future. But the demand for theories and frameworks that tackle the complexity of different types of statistical analytics in a particular field will only increase.

The theoretical framework presented in section 2.6 synthesizes the discovery that CEs can be a powerful glue in navigating the analytics types in a retailing context. It is still only a method to help make better decisions, among other analytical solutions commonly used. Building from this conception, a more practical grasp of the usability of CEs for retail optimization is required, and thus the third research question is a logical next step.

3: How should a systematic CE solution be designed in order to help retail decision makers in improving their business?

In section 4.2, a methodology framework was designed. From a design science perspective, it does what design science is all about: introducing a solution to a practical problem. While the framework is more targeted to managerial audiences, it bridges the gap between a scientific methodology and a use case domain that is widely studied in many disciplines such as business, engineering, design and sociology. The framework can further be adjusted to function as a baseline concept from any research point of view.

As theoretical remarks, the framework highlights the need to form strong, measurable hypotheses for business decisions. Managers tend to have a strong intuition that drives decision-making. Formulating this into a hypothesis that can be statistically tested, has relevance from not only a CE process but in any business decision context. Understanding what the desired effect is and what needs to be scrutinized to see it, is crucial in doing the right thing in a field where mistakes are costly.

Another major aspect in the framework is the process step of addressing the perimeter of influence. In typical use cases of CEs, this is not relevant, but in a retailing context the need to define responsibilities and communicate what is happening is critical. Furthermore, traditional use cases of CEs typically end up in a binary decision on whether the treatment is universally valid or not. As depicted in the framework steps 8-9, in a retailing context a within-group analysis can bring deeper insights that help extend the treatment to only a valid subset of units.

Design science has its challenges in producing theoretical contributions that can further be addressed in science. But science is all about understanding our world better, and artificial domains are equally important as natural ones in making sense of the world and our societies that are changing at an accelerating pace. The three research questions were posed to successively build the concept presented as the methodology framework. This practical artefact should generate interest in anyone researching retailing and how decisions are made in order to remain competitive in such a dynamic industry.

5.2 Managerial implications

While the theoretical part of this thesis is strongly weighed from a retailer point of view, the empirical part brought comprehension on also a software solution provider's side. Creating systematic CE tools is not advisable for retailers themselves, but instead they should look to adopt specialized tools that integrate with existing systems and provide ways to not only monitor but efficiently adjust the operations. This is because software providers have superior capabilities in research and development and typically an ideal combination of industry expertise and solution incubation skills.

For a retailer, the main benefit of a CE methodology is to delineate the steps it needs to take to successfully utilize experimenting. Managers considering running an experiment might be confused about where to start and which things to take into account. For a solution provider, designing and developing a product that systematically allows the planning and execution of experiments requires a conceptual background. The framework can help it comprehend the critical parts that a software needs to address in order to be of practical value to a customer company.

Retailers should encourage a culture of curiosity to even begin with the idea of experimenting. A common theme brought about by the interviews was the reluctance to adopt new processes and challenge the status quo. Whether it is making an assortment decision or purchasing a new planning software, being open minded and staying up to date with the external world might be a question of survival in the industry that is characterized by heavy competition. Modern analytics are not making managers useless, but rather giving them more potential to prosper when better, data-driven decisions can be made. The CE methodology augments the possibilities in how an optimal outcome can be reached while alleviating the risks associated with a decision under uncertainty.

To reap the largest benefits from running controlled experiments, retailers must follow the framework meticulously, including the overarching enablers mentioned in step 0. The framework is designed as a sequential process that steers how the CE will provide results that are useful. It is by no means a silver bullet to how decisions should be made, but there are not many alternative ways to make similar causal inference. It is for a good reason that the CE is so widely used in fields that require extreme precision such as medicine. In retailing, the requisites are not this rigorous but there are no

reasons not to use the precision of a CE if it is convenient and fits the business objective.

Retailers might not understand this without someone offering CEs as a commercial solution. Creating a business around the idea of CEs as a standalone software might not be comprehensive enough to convince retailers. But for software vendors that already offer solutions for the different parts of retail planning and optimization, offering an extension that enables the end-to-end CE process for different decision areas might have tangible value as a product. Retailers that already have decided to adopt analytics are probably also more likely to understand the concept of an experiment. Thus, selling the idea of a tool that helps plan and supervise an experiment with a prescriptive suggestion mechanism might be worthwhile after all.

A software company specializing in retail solutions should consider the entire process instead of only parts of it. The consensus between the interviewees was that a solution needs to be intuitive to use and provide a clear workflow to the user with little manual configuration. It is crucial that a CE solution has a built-in logic of defining the hypothesis, experimental design, sampling, KPIs, data flows and statistical testing automatically. Moreover, as an extension to a planning platform, the solution has to consider whether there are risks of the experiment being contaminated by an external factor or another ongoing experiment. Overall, the way in which a solution is developed to a certain use case area might differ from other areas. A supply chain CE tool is certainly different from a pricing CE tool. But the general methodology of a CE in retail should still follow the one depicted in the framework.

Applying the framework as it is, should be done with caution. Like any managerially targeted framework, making a practical case of it is unique in every instance and this is just one way of approaching the linkage of CEs as a methodology and retailing as an application area. The interviews indicated that the understanding of the applicability of CEs is low both at retailers and even within the solution vendor company. Thus, a sequential concept framework can at the very least start the considerations of CEs as a potential way to optimize retail processes with data-driven causal inferences.

5.3 Limitations and further research areas

As mentioned in the previous section, this study is merely an approach to the topic. A coincidental study with same reference material could end up with partly different types and qualities of results depending on the focus. This study offers a practical perception into the possibilities that CEs can have for retailing.

As retailing was defined intentionally vaguely in chapter 1, the constructed methodology framework is also purposely generic. This is both a strength and a weakness, and thus a more refined version for a selected scope could bring more pragmatism. The methodology could more accurately take into account the maturity levels of how analytics are adopted within the focal retail organization. The methodology is also generic on the actual application areas of CEs, which might not direct the implementation of CEs strongly enough. These limitations related to the universality of the framework are not necessarily jeopardizing what the intent was in the first place. There seem to be no prior studies that create a practical methodology framework on the topic, so a generic artefact can be justified. The exact adoption of it might still need further refinement, but section 4.3 provided at least some practical examples to using it.

As with nearly any study, data collection is a balance between validity and realism. The interviewees were manually picked from a single solution vendor company which might narrow the range of insights. On the other hand, the company operates globally and several nationalities and regions were represented. The sample was also fairly limited in size at $n = 8$. The study could also have interviewed retail decision makers from a point of view that lacks the conception of analytics being a self-evident answer to problems.

Another potential issue with the interview process is that the answers might have been biased due to the introduction of the topic prior to the interview instead of asking the same questions unanticipatedly. There is also a risk of interview questions directing the answers to a certain direction. Future research potential can be seen in studying how retailers see CEs as a decision-making tool and how a practical process would look upon those results.

As to the output of the thesis, the framework was not piloted in a real life application. A future study could implement the concept step-by-step and discuss its applicability. This limitation was recognized in the early stages of

planning of the study, and is not an inevitable issue when design science is chosen as the research design.

A future study building on top of the outputs of this thesis could also refine the framework even deeper into the process step descriptions, for instance how the balancing between statistical power, significance level, sample sizes and effect sizes is optimized to a particular case. Another area to deepen into is the between-group analysis and which tools and statistical models are required to produce insights about compliant demographics within the test units.

Finally, a further study could investigate how CEs compare with other causal inference methodologies. Experimentation is an ancient way to test an intervention and one of its strenghts is that it does not require computing power. An effective CE that provides statistically significant results could be run with pen and paper. But modern tools are improvingly good in capturing nonlinearities and optimizing processes in dimensions that human judgment fails to capture. Mazzanti (2022) discusses the differences between A/B testing and causal ML in a marketing example. A study comparing CEs with causal ML in a retail context in the covered application areas, would be interesting since the latter does not require the rigorous planning and execution of an experiment. But before ML penetrates retailing at a global magnitude, traditional methods such as the controlled experiment combined with contemporary analytics at least do not let retailers down.

References

Anderson, E.T., & Simester, D. (2011). A Step-by-Step Guide to Smart Business Experiments. Harvard Business Review, 89(3), 98-105.

Applebaum, W., & Spears, R. F. (1950). Controlled Experimentation in Marketing Research. Journal of Marketing, 14(4), 505-517.

Bradlow, E.T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. Journal of Retailing, 93(1), 79-95.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

Braun, V., & Clarke, V. (2012). Thematic Analysis. APA handbook of research methods in psychology, (ed. H. Cooper), vol. 2: Research Designs, 57–71.

Brunk, M. E. (1953). Controlled Experiments in Retail Merchandising. Journal of Farm Economics, 35(5), 916-923.

Campbell, D.T., Stanley, J.C. (1963). Experimental and Quasi-experimental Designs for Research. Houhgton Mifflin. ISBN 0-395307872.

Cooprider, J., & Nassiri, S. (2023). Science of price experimentation at Amazon. Business Economics, 58, 34-41.

Cunningham, A.C., & O'Connor, N.J. (1968). Consumer reaction to retail price and display changes. European Journal of Marketing, 2(2), 147-149.

Dekimpe, M.G. (2020). Retailing and retailing research in the age of big data analytics. International Journal of Research in Marketing, 37(1), 3-14.

Dorst, K. (2011). The core of 'design thinking'and its application. Design studies, 32(6), 521-532.

Elgendy, N., & Elragal, A. (2016). Big Data Analytics in Support of the Decision Making Process. Procedia Computer Science, 100, 1071-1084.

Fernie, J., & Sparks, L. (Eds.). (2009). Logistics and retail management: emerging issues and new challenges in the retail supply chain (3rd ed.). Kogan page publishers. ISBN 978-0749454074.

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. International Journal of Forecasting, 38(4), 1283-1318.

Fisher, M.L, & Raman, A. (2018). Using Data and Big Data in Retailing. Production and Operations Management, 27(9), 1665-1669.

Fisher, M.L., Gallino, S., & Xu, J.J. (2019). The Value of Rapid Delivery in Omnichannel Retailing. Journal of Marketing Research, 56(5), 732-748.

Fisher, R.A. ([1935] 1971). The Design of Experiments. British Medical Journal, 1(3923).

Gregor, S., & Hevner, A.R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. MIS quarterly, 37(2), 337-355.

Gregor, S., & Jones, D. (2007). The Anatomy of a Design Theory. Association for Information Systems, 8, 312-335.

Guha, A., Grewal, D., Kopalle, P.K., Haenlein, M., Schneider, M.J., Jung, H., Moustafa, R., Hedge, D.R. & Hawkins, G. (2021). How artificial intelligence will affect the future of retailing. Journal of Retailing, 97(1), 28-41.

Harvey-Jordan, S., & Long, S. (2001). The process and the pitfalls of semi-structured interviews. Community Practitioner, 74(6), 219-221.

Helm, S., Kim, S.H., & Van Riper, S. (2020). Navigating the 'retail apocalypse': A framework of consumer evaluations of the new retail landscape. Journal of Retailing and Consumer Services, 54, 101683.

Hevner, A.R., Ram, S., March, S.T., Park, J. (2004). Design Science in Information Systems Research. MIS Quarterly 28(1), 75–105.

Hevner, A.R., & Chatterjee, S. (2010). Design research in information systems. Theory and practice. Integrated Series in Information systems, Volume 22. Springer. ISBN 978-1441956521.

Holmström, J., Ketokivi, M., & Hameri, A.P. (2009). Bridging Practice and Theory: A Design Science Approach. Decision sciences, 40(1), 65-87.

Huntington-Klein, N. (2021). The Effect: An Introduction to Research Design and Causality, (1st ed.) CRC Press. ISBN 978-1003226055. [Accessed 12.6.2023]. Available at https://theeffectbook.net/.

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: Principles and Practice, (3rd ed.) OTexts. ISBN 978-0987507136. [Accessed 19.4.2023]. Available at otexts.com/fpp3/.

Hänninen, M., Kwan, S.K., & Mitronen, L. (2021). From the store to omnichannel retail: looking back over three decades of research. The International Review of Retail, Distribution and Consumer Research, 31(1), 1-35.

Intel. (2017). Guide to Getting Started with Advanced Analytics. [Accessed 25.4.2023]. Available at https://www.intel.com/content/www/us/en/artificial-intelligence/getting-started-advanced-analytics-planning-guide.html

Kerkhove, L.P. (2022). Data-driven Retailing: A Non-technical Practitioners' Guide. Springer. ISBN 978-3031129612.

Kiil, K., Dreyer, H.C., Hvolby, H.H., & Chabada, L. (2018). Sustainable food supply chains: the impact of automatic replenishment in grocery stores. Production Planning & Control, 29(2), 106-116.

Kohavi, R., & Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. Encyclopedia of machine learning and data mining, 7(8), 922-929.

Kovács, G., & Spens, K.M. (2005). Abductive reasoning in logistics research. International Journal of Physical Distribution & Logistics Management, 35(2), 132-144.

Mazzanti, S. (2022). Using Causal ML Instead of A/B Testing. Towards Data Science. [Accessed 15.6.2023]. Available at https://towardsdatascience.com/using-causal-ml-instead-of-a-b-testing-eeb1067d7fc0

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review, 90(10), 60-68.

McArthur, E., Weaven, S., & Dant, R. (2016). The Evolution of Retailing: A Meta Review of the Literature. Journal of Macromarketing, 36(3), 272-286.

McKinsey (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. [Accessed 16.3.2023]. Available at https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

McKinsey (2020). Automation in retail. An executive overview for getting ready. McKinsey & Company Retail Insights. [Accessed 16.3.2023]. Available at https://www.mckinsey.com/~/media/McKinsey/Industries/Retail/

Our%20Insights/Automation%20in%20retail%20An%20executive%20overview%20for%20getting%20ready/Automation-in-retail-An-executive-overview-for-getting-ready-FINAL.pdf

Montgomery, D.C. (2013). Design and Analysis of Experiments (8th ed.). John Wiley & Sons. ISBN 978-1118146927.

Mou, S., Robb, D.J., & DeHoratius, N. (2018). Retail store operations: Literature review and research directions. European Journal of Operational Research, 265(2), 399-422.

Ngoc, N.M., Viet, D.T., Tien, N.H., Hiep, P.M., Anh, N.T., Anh, L.D.H., Truong, N.T., Anh, N.S.T., Trung, L.Q., Dung, V.T.P. & Thao, L.T.H. (2022). Russia-Ukraine war and risks to global supply chains. International Journal of Mechanical Engineering, 7(6), 633-640.

Pandey, P., & Pandey, M.M. (2015). Research Methodology: Tools and Techniques. Bridge Center. ISBN 978-6069350270

Peffers, K., Tuunanen, T., Rothenbergerm, M.A., Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems 24(3), 45–77.

Peirce, C.S. (1878). Deduction, Induction, and Hypothesis. Popular Science Monthly, 13, 470–482.

Roggeveen, A.L., & Sethuraman, R. (2020). How the COVID-19 Pandemic May Change the World of Retailing. Journal of Retailing, 96(2), 169-171.

Rooderkerk, R.P., DeHoratius, N., & Musalem, A. (2022). The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners. Production and Operations Management, 31(10), 3727-3748.

Saunders, M., Lewis, P., & Thornhill, A. (2012). Research Methods for Business Students (6th ed.). Pearson education. ISBN: 978-0273750758.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company. ISBN 978-0395615560.

Sillanpää, V., & Liesiö, J. (2018). Forecasting replenishment orders in retail: value of modelling low and intermittent consumer demand with

distributions. International Journal of Production Research, 56(12), 4168-4185.

Simon, H.A. ([1969] 1996). The Sciences of the Artificial (3rd ed.). Cambridge, MA: MIT Press. ISBN 978-0585360102.

Sorescu, A., Frambach, R.T., Singh, J., Rangaswamy, A., & Bridges, C. (2011). Innovations in Retail Business Models. Journal of Retailing, 87, S3-S16.

Syntetos, A.A., Babai, Z., Boylan, J.E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. European Journal of Operational Research, 252(1), 1-26.

Thomke, S., & Manzi, J. (2014). The Discipline of Business Experimentation. Harvard Business Review, 92(12), 70-79.

Thomke, S. (2020). Building a Culture of Experimentation. Harvard Business Review, 98(2), 40-47.

van Aken, J.E. (2004). Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. Journal of management studies, 41(2), 219-246.

Verhoef, P.C., Kannan, P.K., & Inman, J.J. (2015). From Multi-Channel Retailing to Omni-Channel Retailing: Introduction to the Special Issue on Multi-Channel Retailing. Journal of Retailing, 91(2), 174-181.

Zentes, J., Morschett, D., & Schramm-Klein, H. (2017). Strategic Retail Management (3rd ed.). Springer. ISBN 978-3658101824.

Appendices

Appendix 1 – Interview template

- How do you see the retail industry shape in the future with diagnostic, predictive and prescriptive analytics instead of the typical descriptive analytics?
    - What kind of analytics solutions do you think will be offered in addition to what vendors currently provide?
- Describe the typical attitude of retailers in how they see data-driven optimization and decision making? How do they approach the capabilities of your solutions? Is there often reluctancy to replace old ways of working?
- What are the main drivers of retailers wanting to implement your products?
- Which pitfalls does your offering have that might make the customer purchase a competitor solution or decide to continue with proprietary tools?
- Do you know of any cases where your customers have conducted experimenting to test a business decision? Please describe in detail.
    - If yes, did your platform support in the experiment? In what ways?
- What do you think are the main reasons from your side not to offer an experimenting solution?
- What do you think are the main reasons from a typical customer's side not to conduct business experimenting?
- How would you see experimenting as a part of your offering in the big picture? How would such a concept fit your product strategy?
- In which product areas of your offering do you see potential for experimenting? In which ones is it the highest? In which cases would experimenting not be a good fit, e.g., current solutions provide adequate results that do not need the rigor of an experiment?
    - What distinguishes an area being more valid for experimenting than another one?
    - Would the value creation of an experimenting solution be of similar type than with your current solutions?
- If an experimenting capability would be built on top of your current solutions, what kind of data and implementation requirements would that make compared to current implementations?