Automation of expert decisions in delayed line haul deliveries: an application of the Behavioural Artificial Intelligence Technology

J. A. Smeets

Master Thesis

TU Delft

Automation of expert decisions in delayed line haul deliveries: an application of the Behavioural Artificial Intelligence Technology

by

J. A. Smeets

Master Thesis

at the Delft University of Technology.

Degree:	MSc Transport, Infrastructure & Logistics	
Track:	Policy	
Faculty:	Civil Engineering & Geosciences	
Student number:	4440714	
Project duration:	May, 2022 – November, 2022	
Supervisors:	Prof. dr. ir. L.A. Tavasszy,	TU Delft, Chair
	Dr. Ir. A. van Binsbergen,	TU Delft, First supervisor
	PhD. Ir. A. Nadi Najafabadi,	TU Delft, Second supervisor
	Prof. Dr. Ir. C.G. Chorus,	Councyl, External supervisor

November 13, 2022

Preface

Throughout this Master's Thesis period, I had the privilege to work with a great graduation committee. First of all, I want to thank Ali Nadi Nafabajadi. You have been by my side for the entire thesis period during our weekly meetings at the Civil Engineering faculty. Especially in the beginning, when I struggled to find a participating company for the case study, you were a helping hand in keeping that fire burning. Secondly, I was to say a special thanks to Caspar Chorus. Although we have not had many long meetings on my thesis progress, the moments we had were very insightful and provided me with the confidence to make some crucial decisions. I also want to thank Arjan van Binsbergen who was not on my graduation committee until quite late in the process. Nevertheless, your theoretical perspective always gave me new insights and accelerated me to think deeper about what actually entails Artificial Intelligence and the purpose of cognitive biases in decision-making. Lastly, I want to thank my chair Lori Tavasszy for the sharp & crisp advice during our kick-off, mid-term and final defence meeting. Your concluding remarks at the end of each of these meetings really summarized the information and paved the foundation for me to continue my work. Then, I also want to say a huge (!) thanks to the Councyl team. Nicolaas, thank you for the warm reception and immediate inclusion into the Councyl team. Especially the lunches at the Adam Tower, stand-up meetings at the Bouwcampus, and, of course, the 2nd anniversary karaoke night in Amsterdam are things that I will remember for the long term. Also a huge thanks to the Councyl ladies Annebel, Stella & Monica, for always being available on short notice to hear me out about my struggles and provide constructive advice. Also, Mark & Guus, as the Councyl Graduation Intern team, it was always fun to share my thoughts with you and above all, drink a beer and have a good laugh together!

Now, being graduated, it is good to shortly look back and say thanks to the people who helped me achieve this. Taking it back six years, I started my Bachelor of Systems Engineering as a recently graduated high school student. And I must say that it's been a good ride which I could not have finished successfully by myself. During the program, there was always one particular group of fellow students that helped me through the hard times, and with whom I experienced the most fun, being: the Technische Matrici. Without this group, I am sure it would have been a lot harder. Also, during my master's program in Transport, Infrastructure & Logistics, Daan, Thom & Joost, without our endless conversations, group projects and of course the famous coffees in the train seats at the Civil Engineering faculty this period would have been a lot more difficult for me.

Then lastly, I want to say special thanks to my housemates from the Statensingel in Rotterdam. You, at some times, trying to understand my topic just to hear me out and blow off some steam has really helped me to advance. Finally, a huge thanks to my parents and brother. Yvonne, Harm & Mark, thank you for your involvement, genuine interest in my topic and curiosity to learn more. Coming from a purely medical family background, I know that it is not your particular field of expertise, however, I sincerely hope that I have been able to learn you all something about Choice Modelling, the MultiNomial Logit model and Artificial Intelligence.

Now it is time for you to start reading. Don't be scared by the high number of pages, some of the stuff looks more difficult than it actually is ;). And if you ever want to have a nice discussion about my results, please feel free to reach out to me. Enjoy!

J.A. Smeets
Rotterdam, November 2022

Summary

E-commerce distribution service decisions are becoming increasingly complex due to increased demand for personalised services, incorporating green delivery assets and outsourcing distribution services to Logistic Service Providers (LSPs). As a result, decision-makers more often encounter multiple objectives translated into contradicting goals. Literature review and expert interviews with two online retailers (E-tailers) and three LSPs, showed that E-commerce distribution service decisions are made in *strategic*, *tactical* and *operational* decision spheres. Within the strategic sphere, *fulfilment network* and *partner contracting* decisions are made. These decisions consider internal strategic goals and external influences, which are often difficult to quantify. Considering these goals and external influences requires human expertise and non-repetitive decision-making. In the tactical sphere, E-commerce actors encounter *channel assortment* and *inventory planning* decisions. These tactical decisions are often made by humans considering historical customer purchasing, shopping behaviour or click-stream & browsing data. Tactical decisions are made more often than strategic decisions due to changing customer needs. In the operational sphere, *channel allocation*, *routing & scheduling* and *last-mile allocation* decisions are made. These decisions constitute day-to-day operations and are made by supervising personnel in distribution centres. Generally, these decisions rely partly on the supervisor's expertise and partly on available information. Operational decisions are repetitive and benchmarked to desired Key Performance Indicators (KPIs) metrics to execute day-to-day operations efficiently.

Over the years, several decision-making tools have been posed to either support decision-makers with complementary information or to completely replace the decision-maker and constitute decision automation. Newer decision-making tools incorporate Artificial Intelligence (AI) technology. With this AI technology, decision tools can correctly interpret and learn from external data and use these learnings to provide complementary advice to decision-makers. One company active in the decision support industry is a spin-off from the TU Delft called Councyl. Councyl provides decision support to hospitals, immigration services and job interview procedures through their Behaviour AI Technology (BAIT) based decision-making tool. BAIT derives expert preferences through statistically designed choice experiments according to Discrete Choice Analysis (DCA) theory. It incorporates these preferences into a decision model that can accurately replicate choices. Especially in decisions that are repetitive and formed through human expertise, BAIT is found to add value to companies. Up to this point, BAIT has yet to be tested in E-commerce distribution services.

To address the applicability of BAIT to e-commerce distribution decisions, we note from earlier applications that BAIT can replicate high-complexity decisions by creating and analysing its own data. Also, decision explainability can be preserved. Referring to the E-commerce distribution services identified in the first section, we find that BAIT applies most to *partner contracting* and *routing & scheduling decisions*. Due to their complex nature, these decisions are not optimised by structured data and, thus, rely (partly) on human expertise. Also, they must be explainable and aim for high accuracy, which can be accommodated by BAIT. Out of these two decision types, we find that routing & scheduling decisions are the most repetitive and reliant on human expertise. Therefore, we decided to model decisions concerning delayed linehaul deliveries, which are a sub-decision of routing & scheduling, according to BAIT methodology. A case study was performed in close collaboration with a group of Operational Supervisors at DHL Express. We hypothesised that automating the delayed linehaul decisions with BAIT's decision-making tool would reduce the discussion and assessment time in delayed linehaul decisions to close to zero. As a result, Supervisors free-up valuable time to supervise other, more demanding sorting & distributing processes. Consequently, a more efficient in and outflow of parcels can be established, which translates to more on-time deliveries and a strengthening of the competitive position of DHL Express in the E-commerce industry.

In the situation of a delayed line haul, parcels which are included in the morning shift routes are delayed. As a result, the cargo spaces of the morning shift delivery vans are incomplete, and delivery drivers have to wait for the delayed parcels, leading to extended delivery times for the other morning shift parcels. Therefore, a pressurized situation arises wherein Supervisors have to choose between departing a share of the morning routes, moving the delayed parcels to the afternoon or moving the delayed parcels to the next day. In total, Supervisors have four choice alternatives. Of these four choices, we regard the definition of departing a share of the morning shift as too comprehensive because it is executable in various ways. As a result, the alternative is difficult to express in a static choice experiment. Moreover, the Supervisors explained that moving parcels to the next day is the worst possible choice alternative. Hence, it is only chosen when the delayed line haul does not include any urgent parcels, which is seldom the case for DHL's Express parcels. Therefore, we decided to only include the *delay morning shift* and *move to afternoon shift* as choice alternatives in the choice experiment.

In total, we had three group discussions with three Supervisors to determine a list of decision criteria (i.e. decision attributes). In total, eight decision criteria were identified, being the *Notification time*, the *Arrival time*, the *Waiting time*, the *Number of stops in the delayed line haul*, the *Number of twelve-hour pieces in the delayed line haul*, the *Number of non-timebound pieces in the delayed line haul*, the *Capacity of the afternoon shift* and the *Closeness of delayed stops to the midday routes.* Next to criteria identification, attribute levels were determined. Attribute levels are a prerequisite for the choice experiment for each choice scenario to indicate a unique set of criteria values. Also, the Supervisors clarified criteria importance scores and criteria-specific characteristics. With this information, we established a survey design consisting of two calibration scenarios and thirty actual choice scenarios. The thirty choice scenarios were generated with Ngene software to ensure most criteria trade-off valuations with the minimum possible choice scenarios.

The choice experiment was executed by nine DHL Express Operational Supervisors deployed in four service centres. We obtained a data set consisting of 270 trade-offs revealing choices. Using this data, criteria weights were estimated using the Apollo library in the programming language R, corresponding to the preferences of the Supervisors. Due to the small sample size, not all criteria weights are scalable to the population according to statistical theory. Therefore, we estimated three different decision-making tools for DHL and tested their validity on the choices of the Supervisors in the choice experiment. First, a *sample model* was estimated incorporating all choice experiment-derived criteria weights. This decision tool can predict 97% (29 out of 30) of the choice scenarios correctly. The choice scenario that was wrongfully predicted consisted of a borderline case where Supervisors were highly divided. Second, a *population model* was estimated incorporating only statistically sound, thus scalable to a population, criteria. This decision-making tool predicted 87% (26 out of 30) of the choices of the choice experiment correctly. Again, wrongfully predicted choice scenarios consisted of borderline choice scenarios. Lastly, the *prior model* consisted of criteria weights determined by the Supervisors prior to the experiment. This decision-making tool predicts 73% (22 out of 30) of the choices correctly and thus is the worst-performing model. Although this analysis indicates that the best-performing model is the sample model, we must place a nuance. The model's performance depends on the intended purpose and goals of the model, which we will discuss in the next section.

With the use of BAIT to accurately replicate expert choices and provide detailed insight into criteria trade-offs, informational black boxes can be opened. The replication of expert choices allows DHL to create valuable Supervisor time by automating delayed line haul decisions or providing a safety net of accurate collegial advice. This backup check would urge the Supervisor to reassess his judgement in case of deviation from collegial advice. In either of the two cases, discussion and assessment time is reduced, which allows Supervisors to allocate valuable time to sorting & distributing tasks. From an academic perspective, the BAIT method is complementary to standard DCA theory since it offers an easy-to-use decision-making tool applicable to real-life choice situations, which we see as the first step towards decision automation. For this particular purpose of replicating choices, we find BAIT complementary to prescriptive Multi-Criteria Decision-Analysis (MCDA) methods. However, the decision tool should be validated in future research with real-life decision scenarios. According to us, this can be done by performing a simulation or a before-and-after study in a controlled setting. Also, we assumed rational and consistent decision-makers in our model. Future research might break these assumptions by including non-rational or inconsistent decisions and assessing model performance accordingly. To DHL Express, we recommend discussing the desirability of the resulting criteria trade-offs. Also, we recommend starting to use the decision-making tool as a safety net for decision-makers. Simultaneously, the tool allows the registration of new choices, which can subsequently be used to update criteria weights accordingly. By doing so, the scalability of the model can be improved. Lastly, we recommend that Councyl continues its partnership with DHL Express. Deriving from the enthusiastic reception of the model by DHL's representatives and the size of its operations, we see the applicability to automate decisions along the full supply chain and within other partly human decision areas.

Acronyms

AI Artificial Intelligence. 2

ASC Alternative Specific Constant. 39, 45

AT Arrival Time. 43, 50

BAIT Behavioural Artificial Intelligence Technology. 1, 3, 9, 11, 25, 33, 57, 61

CA Capacitiy Afternoon Shift. 43, 53

CDC Centralized Distribution Centres. 13

CL Closeness to Midday Routes. 43, 54

DC Distribution Centres. 12

DCA Discrete Choice Analysis. 5, 6, 9, 25, 62

DCM Discrete Choice Models. 25, 26, 62

DSS Decision Support Systems. 4, 7

IT Information Technologies. 1

KPI Key Performence Indicators. 34, 62

LL Log-Likelihood. 27

LSP Logistic Service Providers. 1, 12, 64

MAD Mean Absolute Deviation. 27, 48

MCDA Multi-Criteria Decision-Analysis. 4, 25, 62

MNL Multinomial Logit Model. 27, 33

NO Notification Time. 43, 50

NS Number of Stops. 43, 52

NT Number of Non-timebound Pieces. 43, 53

OR Operational Research. 7

RP Revealed Preferences. 26

RQ Research Question. 8

RUM Random Utility Maximization. 26

SE Standard Error. 27, 39

SP Stated Preferences. 6, 25, 26

TH Number of Twelve-hour Pieces. 43, 52

WT Waiting Time. 43, 51

Contents

Summary v

Acronyms vii

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Problem context 1

1.1.1 Decision-making in E-commerce distribution services 2

1.1.2 Prescriptive decision-making tools 4

1.1.3 Descriptive decision-making tool 5

1.1.4 Behavioural AI Technology 6

1.2 Knowledge gap 7

1.3 Research Objective 8

1.4 Research Questions 8

1.5 Research Outline 8

2 Conceptualization 11

2.1 E-commerce decision hierarchy 11

2.2 E-commerce market decisions 12

2.2.1 Logistic Service Providers 12

2.2.2 Retailers 16

2.3 Conceptual framework E-commerce distribution services 19

2.3.1 Causality of decisions 20

2.4 Behavioural AI Technology to improve decision making 21

2.4.1 Classification of E-commerce distribution services decisions 21

2.4.2 Conclusion on applicability BAIT 23

3 Methodology 25

3.1 Behavioural AI Technology 25

3.2 Choice model estimation 26

3.2.1 Model fit 27

3.3 Choice experiment design 28

3.3.1 Orthogonal design 28

3.3.2 Efficient design 29

3.4 Choice set composition 30

3.4.1 Choice sets 30

3.4.2 Constructing choice sets 31

4 Modelling & Survey Design 33

4.1 Scoping of the decision 33

4.1.1 Decision-makers 33

4.1.2 Decision outcomes 34

4.1.3 Decision time spent 34

4.1.4 Information availability 34

4.1.5 Decision context 34

4.2 Decision criteria inventarisation . . . 35
4.2.1 Criteria identification . . . 35
4.2.2 Criteria range determination . . . 35
4.2.3 Criteria importance determination (priors) . . . 36
4.2.4 Criteria knock-out levels . . . 37
4.2.5 Criteria interaction effects . . . 37
4.2.6 Criteria constraints . . . 38
4.3 Final survey design . . . 38
4.3.1 Choice experiment theoretical framework . . . 38
4.3.2 Survey context statements . . . 38
4.3.3 Survey demographic questions . . . 39
4.3.4 Number of required choice sets . . . 39
4.3.5 Efficient design construction . . . 39
4.3.6 Survey choice sets . . . 42
4.4 Data collection . . . 43
4.5 MNL model estimation . . . 43
4.5.1 Model estimation . . . 43
5 Data Analysis & Implications 45
5.1 Sample description . . . 45
5.2 Calibration scenarios . . . 46
5.3 General MNL Parameter estimates . . . 48
5.3.1 Alternative Specific Constant . . . 49
5.3.2 Notification time (NO) . . . 50
5.3.3 Arrival time (AT) . . . 50
5.3.4 Waiting time (WT) . . . 51
5.3.5 Number of stops (NS) . . . 52
5.3.6 Number of twelve-hour pieces (TH) . . . 52
5.3.7 Number of non-timebound pieces (NT) . . . 53
5.3.8 Capacity afternoon (CA) . . . 53
5.3.9 Closeness to afternoon routes (CL) . . . 54
5.3.10 Interaction Number of Stops (NS) & Closeness to afternoon routes (CL) . . . 55
5.4 Relative importance comparison . . . 55
5.5 Decision-making tools . . . 57
6 Discussion, Conclusion & Recommendations 61
6.1 Discussion & Conclusion . . . 61
6.2 Recommendations . . . 62
A Appendices 65
A.1 Scientific paper . . . 65
A.2 Logistic Biases . . . 73
A.2.1 Judgement biases in logistics . . . 73
A.3 Expert interview questions . . . 74
A.4 Ngene code for choice experiment generation . . . 75
A.5 Choice experiment snapshots . . . 75
A.6 Apollo code . . . 78

List of Figures

1.1 E-commerce distribution services decision classification . . . 3
1.2 Research outline with research questions . . . 9

2.1 Fulfilment network with operational decision flow LSP 1 . . . 13
2.2 Fulfilment network with operational decision flow LSP 2 . . . 14
2.3 Fulfilment network with operational decision flow LSP 3 . . . 16
2.4 Fulfilment network with operational decision flow Retailer 1 . . . 17
2.5 Fulfilment network flowchart with decision flow Retailer 2 . . . 19
2.6 E-commerce decision hierarchy comprising decisions from market practices . . . 20
2.7 Applicability BAIT to decisions . . . 22

3.1 Example attribute level balance . . . 28
3.2 Basic plan example . . . 29
3.3 Choice sets . . . 30

4.1 Conceptual explanation interaction effect . . . 38
4.2 Theoretical framework survey design . . . 38

5.1 Scenario -2 utility maximization scenario . . . 47
5.2 Scenario -1 utility minimization scenario . . . 47
5.3 Utility contribution Notification Time (NO) . . . 50
5.4 Utility contribution Arrival Time (AT) . . . 51
5.5 Utility contribution Waiting Time (WT) . . . 51
5.6 Utility contribution Number of stops (NS) . . . 52
5.7 Utility contribution number of Twelve Hour pieces (TH) . . . 53
5.8 Utility contribution number of Non-Timebound pieces (NT) . . . 53
5.9 Utility contribution Capacity of the Afternoon shift (CA) . . . 54
5.10 Utility contribution Closeness to afternoon routes (CL) . . . 54
5.11 Utility contribution interaction Number of Stops (NS) & Closeness to afternoon routes (CL) . . . 55
5.12 Relative importance parameter estimates and priors . . . 56
5.13 Sample model prediction example . . . 58
5.14 Choice experiment - Sample - Population - Prior model comparison . . . 58

A.1 Ngene code for generating choice experiment . . . 75
A.2 Informed consent statement choice experiment . . . 76
A.3 Example choice scenario experiment . . . 77
A.4 Service centre location question choice experiment . . . 77
A.5 Deployment years question choice experiment . . . 78
A.6 Apollo code part 1. MNL model estimation . . . 79
A.7 Apollo code part 2. MNL model estimation . . . 80

List of Tables

1.1 An overview of direct and choice-based preference elicitation MCDA methods 7

4.1 Overview attributes, type, levels and range 36
4.2 Overview attribute priors 36
4.3 Notification time dummy coding scheme 37
4.4 Arrival time dummy coding scheme 37
4.5 Demographic factor questions 39
4.6 b0 constant calculation 40
4.7 Choice set probabilities 41
4.8 Ngene choice experiment output 42

5.1 Demographic characteristics & completion time respondents 46
5.2 Different prediction model comparisons 48
5.3 Parameter estimates Model 1 49
5.4 Different model utility contribution based on Scenario -1 with and without ASC 50
5.5 Comparison relative importance parameter estimates and priors 57

1 Introduction

Increasing customer demand for personalised and green delivery services increases the importance of thorough consideration for E-commerce actors when making choices regarding their distribution services. In e-commerce distribution services, complex in- or outsourcing, product-channel allocation, fulfilment network set-up, and sustainable delivery asset decisions must be made. Human decision-makers often encounter multiple objectives, sometimes contradicting goals and incomplete information within these complex decisions. As a result, data analysis & AI-driven tools are utilised to support decision-makers. Although these methods improve efficiency on many occasions, particular decisions where experts' attitudes are decisive are difficult to capture by conventional AI. The Behavioural Artificial Intelligence Technology (BAIT), developed by Councyl (2020), identifies expert knowledge and replicates decisions with a decision-making tool. This tool might be used to replace decision-makers for the purpose of automation. As a result, valuable time is created which can be dedicated to other processes. BAIT's applicability and added value have been tested in health, insurance and the public sector with positive results. However, other industries where humans are involved in complex decision-making, might also benefit from BAIT's decision-making tool. E-commerce is one of these industries. In this research, we investigate the decision-making processes of E-commerce actors in distribution services and assess the applicability of BAIT to enhance the efficiency of decisions. In this introductory section, First, we set a problem context by discussing literature concerning decision-making in E-commerce distribution services. Second, a distinction is made between prescriptive and descriptive decision tools, and the BAIT method is discussed. Third, we address a resulting knowledge gap, propose a research objective and translate that objective into several research questions. At last, we elaborate on the research outline.

1.1. Problem context

In the early days of E-commerce, retailers used phone calls and paper letters to communicate with customers. These customers did not count delivery times in days but weeks. As opposed to current days, customers did not derive convenience from quick deliveries but from the ability to order goods while continuing life as usual. However, this changed with the rise of the internet from the 2000s onwards. As a result, E-commerce saw an enormous increase in demand, which incentivised online retailers (E-tailers) to coordinate multiple distribution centres in strategic locations. Consequently, these E-tailers were left with complex fulfilment network set-up and coordination decisions, which proved too challenging to combine with other company operations (Robinson, 2014). As a result, many E-tailers decided to focus on the product and partnered with Logistic Service Providers (LSP) to outsource (part of) their logistic services. As a result, production firms focused solely on core competencies like research & development and large-scale production while simultaneously, unique logistic experience accumulated over the years within LSPs (Li et al., 2012). The latter created an entirely new business case in which personalised services like flexible delivery place and time, next-day delivery and flexible methods for collecting & returning parcels were added to the logistic mix and offered to customers. According to Vakulenko et al. (2019), offering these personalised services has become essential to improve the online shopping experience and thus ensure a competitive advantage over other firms. Luk et al. (2018) confirm this view and conclude that consumer satisfaction not only depends on product quality but also on offered product distribution services. The offering of these additional services, by either LSP or E-tailer, requires the integration of new Information Technologies (IT) with existing business practices. IT systems have become essential for firms to facilitate these personalised distribution services and identify and adapt to ever-changing customer needs. Although these systems have proven to strengthen a firm's competitive position, IT systems are rapidly evolving and thus require a firm to have a flexible business model. Therefore, E-tailers and LSPs must be aware of new emerging IT systems, which might give them a competitive advantage.

One of the newest developments in e-commerce IT systems is Artificial Intelligence (AI). Kaplan and Haenlein (2019) describe AI as: "a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation". Using these learnings and flexible adaptation, e-commerce firms may offer more personalised services to enhance customer loyalty and win an additional customer base to strengthen their competitive position. According to Kaplan and Haenlein (2019), AI knows three different types; according to Akter et al. (2021), depending on the context, AI might be expressed via a tool, technique or algorithm. First, purely analytical AI inhabit characteristics which are consistent with cognitive intelligence. This cognitive intelligence is learned through big data consumption to reflect and add to past experiences. Increased computing power and internal memory enable the AI system to detect (to the human) non-observable patterns in the data, providing more accurate advice. Analytical AI is the AI method that is most commonly used (i.e. fraud detection, image recognition, self-driving cars) and thus best known. Second, next to cognitive intelligence, human-inspired AI also inhabits emotional intelligence. By doing so, human emotions tend to be identified and considered in the AI system's output. Third, are humanised AI systems. These systems inhabit next to cognitive and emotional competencies also socially intelligent characteristics and thus reach a high level of self-consciousness. These systems are non-existent but might be in the (distant) future.

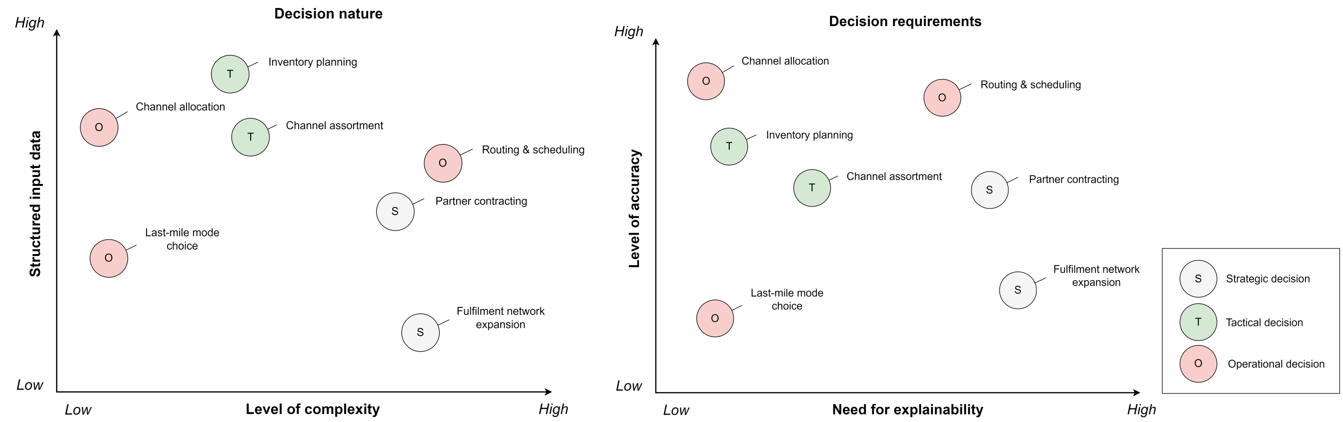
BAIT's methodology identifies decision makers' preferences, mimics their choices with an interactive prediction model, and uses this model for introspection to simulate and/or automate decisions. Corresponding to the AI terminology of Kaplan and Haenlein (2019), BAIT interprets and learns from external data and captures decision-makers' intelligence into user-friendly software to automate the decision-making process. As a result, valuable time is created for decision-makers to assign to more demanding task. This decision-making tool accounts for cognitive and emotional intelligence and provides detailed insight into existing decision criteria trade-offs. Comparing the BAIT method to the three AI methods described above corresponds most to the human-inspired AI classification. Therefore, the intelligence captured by such software can automate decision making processes by reproducing expert choices. To understand the decision-specific applicability to e-commerce distribution services of the method, first, we need to classify decisions within this industry, which we conduct in the next section.

1.1.1. Decision-making in E-commerce distribution services

In chapter 2, we take a structured approach to identify a decision hierarchy and inherited decisions that shape E-commerce distribution services. First, we review the literature to identify a three-step decision hierarchy. Second, we conduct expert interviews to determine underlying decisions within each decision sphere. Third, we propose a conceptual framework incorporating decision spheres and underlying choices. Fourth, we assess the applicability of BAIT to these decisions based on each decision's nature and requirements. In this section, we shortly summarise these analyses.

Altogether, we identified three decision spheres. Strategic decisions are based on customer service objectives, which are translated into company goals. These decisions are made on a multiple years basis and therefore have a long-term nature. Second, tactical decisions entail certain decisions that organise the logistic set-up according to customer expectations. These decisions are usually supported by customer data and the availability of technical resources to analyse this data. Third, operational decisions are made on a day-to-day basis to facilitate operations. As thoroughly described by Riopel et al. (2005), these decision spheres and the underlying decisions are interconnected. Although this is the case, we note that inter-connectivity is purposely left out of our analyses since we deem it outside our research's scope. Within the decision spheres, seven decisions were identified by interviewing E-commerce actors. These decisions are: *Fulfilment network expansion*, *Partner contracting*, *Channel assortment*, *Inventory planning*, *Channel allocation*, *Routing & scheduling* and *Last-mile allocation*.

These decisions tend to be optimised through various supporting mechanisms. However, there is not one supporting IT mechanism that fits every decision. Some decisions are best supported by non-human knowledge-based decision-support tools, like machine learning, while others are best supported by knowledge-based decision-support means, like decision rules. For example, data-driven algorithms consume big sets of historical data to support decisions and, by doing so, can reach high levels of accuracy. However, they are often tricky to explain, require specific expertise and, as mentioned, need a lot of data. Other decisions are made under incomplete information, are the product of a consensus between stakeholders or are constrained by limited resources and are thus sub-optimal. These decisions require human expertise and might be supported by decision rules or standards. As a third option, decisions can also be partly rule-based and partly data-driven. Literature offers a multitude of Multi-Criteria Decision-Making methods that support human decision-makers in selecting the best alternative. To link each of the decisions mentioned above to a particular support mechanism, two decision classification categories and four decision classification criteria are distinguished. A decision's nature is measurable by the degree of *structured input data* and the level of *complexity*. Moreover, we rank the decisions based on their decision requirements. Decision requirements consist of a need for *explainability* component that indicates the importance of explaining how decision-makers made a decision. And a decision *accuracy* component that indicates the desired accuracy level. By classification of the decisions based on these criteria, we intend to visualise which decisions are subject to which supporting methods.



(a) Decision classification based on decision nature

(b) Decision classification based on decision requirements

Figure 1.1: E-commerce distribution services decision classification

In Figure 1.1, all seven decisions are classified according to the four decision classification criteria. As can be seen, some decisions, like channel allocation and inventory planning, might benefit from a high degree of structured data input. With the availability of structured input data (e.g. customer locations, customer demand and warehouse capacity), these decisions are subject to computer-driven optimisation. They can therefore reach high levels of accuracy. Moreover, these decisions do not require high levels of explainability. However, other choices, like routing & scheduling and contracting decisions, are often made with incomplete information, which requires more cognitive effort from human decision-makers. As a result, these decisions are increasingly complex and consist partly of the decision-makers behaviour that is difficult, if not impossible, to capture by solely computer-driven algorithms. Placing BAIT somewhere on this spectrum to measure its applicability requires an understanding of its purpose.

We classify Behavioural Artificial Intelligence Technology (BAIT) as a decision-making tool since its purpose is to describe and replicate expert decisions accurately. To this end, the replication of decisions serves to exclude decision-makers from the process by automating the decision. This is an important area of research because full automation of decisions would reduce information gathering and discussion time to zero and allow decision-makers to allocate valuable time to other company processes. Especially in operational processes, decisions are often repetitive and based on best practices. Although this is the case, human decision-makers are continuously involved in repetitive and best-practice decisions, which indicates the potential to improve operational performance through automation. In addition, time margins are generally low, and decision-makers have multiple processes running simultaneously, contributing to a pressurized environment. As a result, we see decision automation as a means to allocate more valuable decision-maker time to the sorting & distributing processes. Consequently, increased time allocation to these processes increases efficiency, which translates into more on-time deliveries and, thus, strengthens the competitive position. Next to BAIT, we find several decision tools which use sophisticated Multi-Criteria Decision-Analysis (MCDA) methods to establish decisions. These methods are classified as Decision Support Systems (DSS) and serve the purpose of assisting decision-makers by providing complementary information, which makes them more prescriptive in nature. In regard of purpose, we distinct decision automation or assistance, and descriptive or prescriptive tools to achieve one out of these two. In the next sections, we elaborate on these two approaches.

1.1.2. Prescriptive decision-making tools

The literature proposes various Decision Support Systems (DSS) that intend to support decision-makers by indicating how decision criteria should be valued. Decision criteria valuations indicate the weight to which the criterion is decisive for the outcome of the decision. These DSS use sophisticated MCDA methods to adjust directly chosen criteria weights for emotional obscurity, biases and incomplete information into *prescriptive* weights. These prescriptive weights are inhabited in the DSS and used by decision-makers as a matter of support. In this section, we identify 6 MCDA methods: Analytical Hierarchy Process (AHP), Fuzzy Analytical Hierarchy Process (FAHP), Interpretive Structural Modelling (ISM), Case- & Rule-based Reasoning (CBR), Best-Worst Method (BWM) and Decision Making Trial and Evaluation Laboratory (DEMATEL). We will discuss these different MCDAs next.

Analytical Network Process (ANP) & Analytical Hierarchy Process (AHP)

The ANP and AHP criteria weight estimation methods are fundamentally very similar. Therefore they are combined into one section. In ANP & AHP, after identification of decision criteria, the decision maker makes pairwise criteria comparisons to address their importance (Meade and Sarkis (2002)). An example of such a question is: 'How much more important is the delivery time to the delivery amount, on a 1-9 scale?'. By doing so, the importance of each criterion is defined and accordingly valued as such. Although this allows for varying the decisiveness of each measure, the weighting is very dependent on the judgement abilities of the decision-maker. To account for this, several mathematical techniques are utilized to account for biases, emotional obscurity and incomplete information.

Fuzzy Analytical Hierarchy Process (FAHP)

FAHP was used by Singh et al. (2018) to support the selection of LSPs in a cold chain of perishable goods in Inda. The fuzzy method proposes linguistic variables (i.e. equal - perfect) instead of numbers, like ANP & AHP, to decision-makers to make pairwise comparisons. These linguistic variables are connected to fuzzy-scaled numbers that address the decision maker's fuzzy environment of incomplete information. By doing so, FAHP is able to support decision-makers in situations of incomplete information.

Interpretive Structural Modelling (ISM)

Thakkar et al. (2005) combined ANP with ISM to address a real-life case of India's organic food sector. In addition to ANP, the ISM method first conducts a literature review to identify decision criteria. Second, brainstorming sessions are organised with a group of experts. These sessions aim to verify the decision criteria and establish contextual relationships among them. Through discussion, the available information among decision-makers tends to be increased, and thus more reliable criteria weights are derived as input.

Case-Based Reasoning (CBR) & Rule-Based Reasoning (RBR)

Isiklar et al. (2007) propose a framework that includes CBR and RBR to support the selection of LSPs. CBR retrieves information by analysing previous situations and reusing that information in new cases. It is applicable when knowledge acquisition is time extensive, the risk for repeated mistakes is high, and reasoning happens under incomplete information. On the contrary, training a CBR algorithm for selecting historic cases requires significant learning time, and the algorithm's quality depends on the quality of historical instances. RBR establishes knowledge through 'if-then' rules. Therefore, decision outcomes of RBR are mostly uniform since they follow a predefined set of rules. As a result, it tends to be a poor method for supporting decisions that deal with dynamic contexts, which is often the case in logistics. In their overarching framework (Isiklar et al., 2007), the decision maker is asked to rate the importance of each decision criterion by linguistic variables to which fuzzy numbers are connected. This process provides criteria weights that are compared with historical cases to select appropriate outcomes.

Best-Worst Method (BWM)

A method proposed by Rezaei (2015) that tends to deal with the uncertainty of human judgement is the BWM. At first, the best and worst criteria are selected from a predefined set. Second, for both the best and worst criteria, an importance sequence is defined regarding all other non-best and non-worst criteria. By doing so, decision-makers don't have to assign a strength directly between criteria (i.e. 2x,3x,4x better) but only define a sequence. Third, a maximin function is executed to derive criteria weights. Compared to AHP, BMW needs fewer pairwise comparisons, contributing to the quickness of the method. Also, it does not ask about the strength of one criterion over the other, but only the direction of importance. BMW eases the decision-maker job since a 'worse or better' assessment is made more quickly than a 'how much worse or better assessment. As a result, less reliance is put on the judgement abilities of the decision-maker to derive more reliable criteria weights.

Decision Making Trial and Evaluation Laboratory (DEMATEL)

Govindan and Chaudhuri (2016) used the DEMATEL method to select LSPs. This method takes a risky approach. First, a list of logistic risk categories is composed that might influence the selection. A literature review is conducted, and expert opinions are gathered. Next, interrelationships are decided between these risk categories. An example of a posed question to a decision-maker is: "what is the extent of influence of planning risk and process design (R1) on quality risk (R2) on a scale of 0–4"?. By repeating this process for all risk categories, a causality string is made, indicating the importance of each risk category. Eventually, this string is used for decision support. By taking a risk categorisation approach, the DEMATEL approach is different to other criteria weight estimation methods.

Most existing MCDA methods provide prescriptive criteria weights through sophisticated methods which are complementary to direct judgement criteria weights. By doing so; falsified or incomplete information (Jharkharia and Shankar, 2007), the obscurity of human emotion (Karthikeyan et al., 2019), human perception bias, and group perception bias (Qureshi et al., 2007) tend to be addressed for, and thus not incorporated into the DSS. Therefore, by using these MCDA methods, complementary information is embedded in the DSS to assist decision-makers. For automation purposes, descriptive decision-making tools might be better fitting as a means to replicate decisions. We elaborate on one of these methods next.

1.1.3. Descriptive decision-making tool

The behavioural method BAIT uses Discrete Choice Analysis (DCA) to derive existing criteria valuations implicitly. According to this theory, expert preferences are made elicit through statistically designed choice experiments that gather maximum criteria trade-off valuation with as few questions as possible (Akiva and Lerman, 1985). The method BAIT captures these preferences as descriptive criteria weights into an interactive prediction model that replicates the expert group's choices at a high accuracy rate. Utilizing the *descriptive* DCA method, decision-maker preferences including bias, emotional obscurity and incomplete information, are embedded into the decision-making tool. We see this ability to reproduce choices including the exact decision-maker preferences as the first step towards decision automation. Because the DCA method has this sole purpose of reproducing choices, potentially inherited decision biases should not excluded from the decision-making tool. To inform the reader on the meaning of bias, and their existence in logistics, we deem it essential to address and investigate their presence. We do so, first, by posing the definition of a cognitive bias by Tversky and Kahneman (1974). Second, we refer to Knapp et al. (2021), who identified three cognitive decision-maker biases within each of the three before-mentioned decision spheres: strategic, tactical and operational. For this elaboration, we refer the reader to section A.2.

The conclusion that a large part of MCDA methods is rather prescriptive, we see a purpose for BAIT and its descriptive DCA in decision automation. Chorus (2015) argued five reasons why descriptive criteria weights should be based on choices instead of judgements if we want to reproduce human decision-making. First, people often do not know which particular trade-offs they make in day-to-day choice situations. Second, people hesitate to be open about their trade-offs. Third, judgements are more prone to bias than choices due to their direct nature. Fourth, criteria trade-offs are the product of historical decisions, not explicit judgement. Moreover, fifth, economic theories are commonly based on choices, not judgements. According to these arguments, criteria weight determination, to replicate decision-maker choices would be more accurate using DCA theory than the other MCDA methods.

Next, we will discuss the Behavioural AI Technology, BAIT.

1.1.4. Behavioural AI Technology

This section describes Behavioural AI Technology (BAIT) in general terms. We refer the reader to section 3.1 for a more detailed description. BAIT can be dissected in two parts. The first, descriptive, part is fundamentally based on the Discrete Choice Analysis (DCA) theory. In the second, rather underdeveloped part of BAIT, descriptive weights are translated to prescriptive weights based on the introspection given by the prediction model. This process of adjusting weights is usually based on commonly agreed expert consensus or available ground truth cases. Although this second part is proposed as part of the method, it is unclear how translation of criteria weights would take place. Therefore, in our research, we focus on the descriptive DCA part of BAIT. Regarding academic mentions, the BAIT method is in an infancy stage and was only used by ten Broeke et al. (2021) in a medical context.

In total, the descriptive DCA part of BAIT consists of six steps: *Decision scoping*, *Decision characteristics inventarisation*, *Choice experiment design*, *Choice experiment execution*, *Model estimation* and *Model validation*.

In steps 1 to 4, what we see as the **behavioural** part of the method, expert decisions are collected using the choice experiment. First, the decision is scoped to identify the decision outcomes, decision-makers, information availability, decision context and decision time. Next, the decision characteristics inventarisation step identifies a list of decision criteria and corresponding criteria values. Also, criteria-specific characteristics, like knockout values and criteria constraints, are determined. For a better understanding of these concepts, the reader is referred to section 3.1. These first two steps are executed in close collaboration with the group of experts. In the third step, the choice experiment is designed. Specific decision characteristics ask for specific choice experiment designs which should be taken into account. Ngene software is used to generate the most efficient choice experiment design. In the final behavioural step of the method, the choice experiment is executed by a group of experts. Each choice scenario comprises two or more alternatives with decision criteria and varying criteria values. By asking the respondents to choose one option over the other(s), no explicit attribute weights are requested of the respondent. This way of gathering preferences through the responses of a choice experiment is called Stated Preferences (SP) collection.

In steps 5 & 6, an **optimization** problem is solved. Step 5 estimates criteria weights in the programming language R using the DCA library Apollo based on the SP data. Criteria weight estimation is an iterative process in which each new criteria weight is tested on the choices made by the respondents. This process is repeated until no improvements can be made. As a result, the most reliable criteria weights are obtained. In the second optimization step, the criteria weights are loaded into a decision-making tool to predict the choice scenarios of the choice experiment. This validation step is performed to assess the accuracy of the criteria weights and the corresponding advice. In the end, the decision-making tool can be used to assist a group of experts, or automate a decision according to the underlying weights.

Table 1.1: An overview of direct and choice-based preference elicitation MCDA methods

Author	**Methodology**	**Criteria weight determination**	**Decision tool nature**
Direct elicitation methods			
Meade and Sarkis (2002), Jharkharia and Shankar (2007), Gol and Catay (2007)	Analytical Network Process (ANP) & Analytical Hierarchy Process (AHP)	Pairwise criteria comparisons, ratio scale 1-9	Prescriptive
Singh et al. (2018), Zhü (2014), Saaty and Tran (2007)	Fuzzy Analytical Network Process (FANP)	Pairwise criteria comparisons, ordinal scale, linguistic terms, equal (1) - perfect (9)	Prescriptive
Thakkar et al. (2005), Qureshi et al. (2007)	Interpretive Structural Model (ISM) & ANP	Pairwise criteria comparisons, binary choice 0 (criteria i influences j) - 1 (criteria j influences i)	Prescriptive
Isiklar et al. (2007)	Case Based Reasoning (CBR) & Rule Based Reasoning (RBR)	Criteria rating, ordinal scale, absolutely important - important	Prescriptive
Govindan et al. (2016)	Decision Making Trial and Evaluation Laboratory (DEMATEL)	Risk category (criteria) pairwise influence comparisons, ratio scale 1-4	Prescriptive
Rezaei (2015)	Best-Worst Method (BWM)	Direction categorization best-to-others & worst-to-others, binary scale	Prescriptive
Choice-based elicitation methods			
ten Broeke et al. (2021)	Discrete Choice Analysis (DCA)	Application of choice experiments to collect decision criteria trade-offs, followed by weight estimation by statistical analysis	Descriptive

Table 1.1 indicates an overview of direct and choice-based elicitation methods. For each of the methods, the criteria weight determination method is summarised in one sentence. As can be seen, all direct elicitation methods, except for DEMATEL, make use of a sophisticated method involving human or group judgement. In contrast, DCA, uses choice-based elicitation through choice experiments and statistical analysis. The existence of cognitive biases by Tversky and Kahneman (1974), the identification of cognitive biases in logistics (Knapp et al., 2021), and the shortcomings of human judgements (Chorus, 2015) indicate that choice-based elicitation tops direct elicitation when the goal is to replicate human decision making. Therefore we deem BAIT and its underlying DCA methodology an interesting method for further research when replicating expert choices for automation purposes.

1.2. Knowledge gap

Besides an abundance of Operational Research (OR) papers to optimise delivery routing, scheduling and warehousing practices, some critical decisions in designing and managing distribution services remain human-driven. As advocated in the literature review above, existing Decision Support Systems (DSS) use human judgements to establish decision criteria weights which are subsequently used to provide complementary information to decision-makers. However, as argued by Chorus (2015), when the goal is to replicate choices, which we see as first step to decision automation, criteria weights should be based on choices instead of direct judgements. If not, criteria weights potentially deviate from their intended criteria valuation and as a result, decision outcomes deviate from what is desired, potentially affecting decision performance. We find that incorporating a behavioural component to establish decision tools for automation is a promising and not sufficiently covered research area in e-commerce. Including human behaviour in logistic decision-making was advocated by both Hofstra and Spiliotopoulou (2022) and Cui and Zhang (2018), who obtained positive results.

As discussed in the sections above, BAIT incorporates a behavioural component in criteria weight estimation by modelling expert choices instead of direct judgements. Therefore, we will investigate the application of BAIT in e-commerce distribution services. In the remainder of this research, first, we construct a framework which can be used as a classification tool to identify decisions in E-commerce distribution services. Second, we assess the applicability of BAIT to one particular decision. Third, we perform a case study at an E-tailer or LSP wherein we model a decision according to BAIT methodology.

1.3. Research Objective

Following the knowledge gap that is described above, the objective of this research is stated as:

To assess the added value of Behavioural AI Technology to the decision-making process of Dutch E-commerce actors by conceptualising and modelling one of their distribution service choices

1.4. Research Questions

We raise several research questions to address the knowledge gap and research objective. By doing so, we structure the study systematically. The research questions answered during this research are the following:

1. What decisions are included in setting up, managing and executing E-commerce distribution services from an E-tailer and Logistic Service Provider perspective? (section 2.3)
2. Where in this decision sphere might Behavioural AI Technology be of added value? (subsection 2.4.1)
3. What is the most suitable decision to test the application of Behavioural AI Technology in the E-commerce sector? (subsection 2.4.2)
4. How do directly judged criteria trade-offs differ from those obtained by modelling choices with Behavioural AI Technology? (section 5.4)
5. How do different criteria trade-off valuations affect choice outcomes in a sample, population and prior model? (section 5.5)
6. In what ways can Behavioural AI Technology contribute to improve decision-making in E-commerce distribution services? (section 6.1)

By subsequently answering these questions, we will accomplish the intended research objective. Question 1 provides insight into the decision-making processes inherited in e-commerce distribution services. Question 2 aims to identify if and where Behavioural AI Technology might add value to these distribution services. Question 3 addresses which particular decision we choose to model in the remainder of the research. Then, after modelling and analysis for which we choose DCA over MCDA because of our intended automation purpose, question 4 addresses the potential difference between initially judged criteria valuations and those obtained with BAIT. Obtaining this insight is relevant because it indicates what criteria are expected to be important based on direct judgement and what criteria are important based on implicit choices. This difference highlights unknown decision-making, which signifies the presence of bias or hidden expertise. Subsequently, question 5, discusses the subsequent differences in choice outcomes when applying these different criteria valuations. At last, question 6 will address BAIT's potential contribution to improving efficiency in the e-commerce sector.

1.5. Research Outline

In this last section of the introductory chapter, we provide an overview of the research outline. A visualization of the sequence in which the chapters were written is is posed in Figure 1.2. In this overview, six research chapters are posed, and an indication is given as to where each Research Question (RQ) is answered. By doing so, we guide the reader through the research in a structured way.

We perform the research according to three overarching discovery, implementation and evaluation stages. In the first discovery phase, an extensive literature review is performed on E-commerce decision hierarchy, decision support tools and Behavioural Artificial Intelligence Technology (BAIT) to derive a problem definition. This problem definition is subsequently used as a base to perform the conceptualization in the second chapter. In the conceptualization step, the literature review results and expert interviews are conceptualized into a framework that E-commerce actors can use to classify decisions. After framework finalization, we assess the applicability and added value of Behavioural AI Technology by decision classification. This marks the end of the discovery phase, and the first three RQs are answered. In the second implementation phase, an individual E-commerce distribution service choice is modelled according to BAIT methodology in the modelling & analysis chapter. To do so, the methodology BAIT and its underlying Discrete Choice Analysis (DCA) theory is discussed in chapter 3. Subsequently, decision modelling is performed & the final survey design is determined in the modelling & survey design chapter. Next, the evaluation phase is originated in which the obtained data from the survey is analyzed and implications discussed. By doing so, we answer RQ4 and RQ5. Finally, in the discussion, conclusion & recommendation chapter, the answers to the first five RQs are synthesized and we answer the final question. Also, limitations of the study are discussed, and recommendations are made for further research, DHL Express and Councyl. Figure 1.2 shows a comprehensive overview of the research outline.

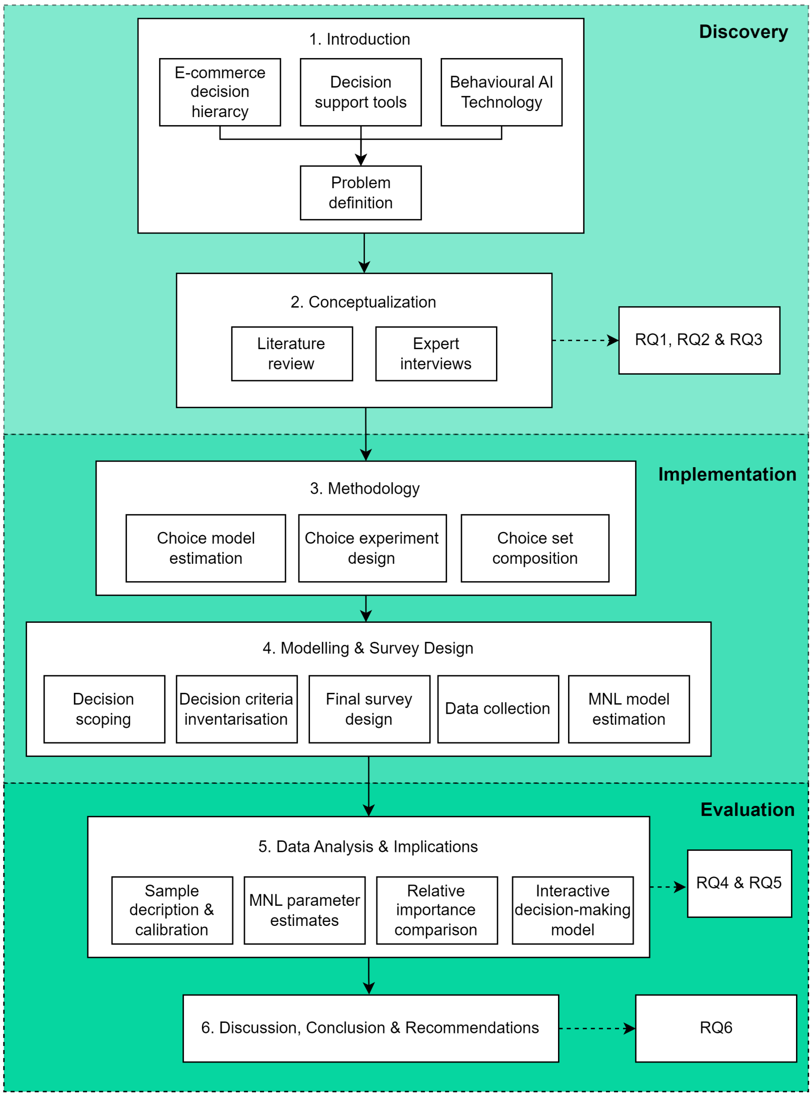


Figure 1.2: Research outline with research questions

In total, we write six chapters to provide answers to four research questions. We combine the problem definition and literature review steps into one Introduction chapter. Next, in chapter 2, Conceptualisation, we discuss the literature review, which entails a hierarchical classification method to identify decision processes and elaborate on the expert interviews. Next, in chapter 3, Methodology, we discuss the BAIT method and DCA. After that, in Chapter 4, Modelling & Survey Design, we elaborate on the modelling steps and pose the final survey design. In Chapter 5, Data Analysis & Implications, we discuss the experiment's results and validate its performance. In Chapter 6, Discussion, Conclusion & Recommendations, we conclude on the applicability of BAIT to increase efficiency in the decision-making processes of Dutch E-commerce actors. Additionally, the study's limitations and new directions for research are discussed.

2 Conceptualization

In this chapter, we conceptualize the decision hierarchies that comprise E-commerce distribution services and identify their underlying decisions. Both a theoretical and a practical (market) perspective is acquainted. In terms of theoretical perspective, we first assess the different decision hierarchies to which Behavioural Artificial Intelligence Technology (BAIT) might contribute using a literature review. Secondly, we conduct expert interviews with industry players' representatives to identify the presence of underlying decisions and their subjectivity to decision support. After identification, thirdly, a comprehensive conceptual framework is posed consisting of both decision hierarchies and incorporated decisions. This framework serves as a theoretical contribution that researchers might use in the future for decision identification purposes while dissecting e-commerce fulfilment networks. At last, we assess the applicability of BAIT to the decisions of the conceptual framework.

2.1. E-commerce decision hierarchy

We used various search engines to find relevant research papers. Search engines include, but are not limited to, Scopus, Google Scholar, ScienceDirect, TU Delft Repository & ResearchGate. In terms of search strategy, the initial search used to find different decision support methods in chapter 1 was extended. This initial search was performed in Scopus and consisted of keywords: E-commerce and decision-support with multiple synonyms to increase reach for articles. ("e-commerce" OR "ecommerce" OR "E-fulfilment") AND (decision-support OR "decision support") resulted in 573 research papers. The list was considered top-down based on titles by sorting the papers on the number of citations. Papers that propose a decision support method or conceptualize decision-making of distribution services were selected into a separate list. After this assessment, we read the abstracts and conclusions of all separated papers. The Snowballing strategy was executed whenever the abstract and conclusion were deemed relevant. We consulted both reference lists (backwards snowballing) and citation lists (forward snowballing) to find new papers. These new papers, if deemed relevant, were added to a separate list. The papers that resulted in the separate list were used in this literature review.

Within setting up, facilitating, maintaining and operating distribution services, many decisions arise. In their comprehensive book on logistic operations, Riopel et al. (2005) identify a myriad of forty-one decisions in logistic processes consisting of a.o. Inventory management, product packaging and warehousing operations. According to the authors, logistic decisions are categorised into three categories: Strategic planning, network, and operations. Rooderkerk and Kök (2019) attained an omnichannel perspective regarding channel assortment planning of distribution services. In omnichannel distribution, channel boundaries are broken down, and performance is not regarded per channel but as a holistic brand experience. The authors address planning assortment concerning strategic, tactical and operational decision challenges. Moreover, Melacini et al. (2018) gather a holistic view that again indicates three decision spheres within the set-up of an e-commerce fulfilment network. First, the distribution network should be considered. This design has underlying decisions concerning channel expansion: how many fulfilment centres do I need? Where do I locate them? Are logistic services outsourced or facilitated by owned delivery assets? Second, inventory management and assortment planning were mentioned by the authors. These particular decisions correspond to the tactical decision sphere of Rooderkerk and Kök (2019): what range of products do I want to offer through my online and physical channels? And do I want to personalise assortments for specific customers based on, e.g. cookies and GPS locations? But also inventory management decisions: do I want to hold inventory in my stores to serve demand? or use stores for showrooming only and nudge customers to buy online? Do I pool all my stocks at one location or decentralise inventories across Distribution Centres (DC)? Third, delivery planning and execution decisions should be considered. How do I decide the transport mode for particular packages? How do I plan and schedule my delivery tours? And which package goes to which DC for fulfilment?

These three themes of Melacini et al. (2018), similar to Rooderkerk and Kök (2019), and to a large extent Riopel et al. (2005), can be brought down to a three-step decision hierarchy, consisting of *Strategic, Tactical* and *Operational* decision spheres. Although the papers are aligned on this, a contradiction arises between the operational decision spheres of Rooderkerk and Kök (2019) and Melacini et al. (2018). In contrast, the latter describes day-to-day organisational decisions, while Rooderkerk and Kök (2019) attain a more strategic (longer-term) angle, considering assortment choices and pooling strategies. It is important to note that we see assortment and pooling strategy decisions as present in the tactical sphere and, therefore, in the remainder of this research, regarding the operational sphere as consisting of day-to-day organisational decisions.

To further discover the decisions within each sphere, expert interviews are summarized in the next section with both Logistic Service Providers (LSP) and Retailers to identify E-commerce market practices. Next, we will combine the resulting decisions with the literature above to make a conceptual framework for E-commerce distribution services. Doing so provides a clear overview of the current choices in E-commerce decision-making.

2.2. E-commerce market decisions

Expert interviews were conducted with several key players in the E-commerce industry to assess the presence and relevance of the decisions identified above. LSPs and E-retailers were contacted through Linkedin. We sent messages to employees in management, network efficiency, or other logistical job positions. In total, we spoke to three LSPs and two Retailers.

The interviews were semi-structured, corresponding to the qualitative research lecture of Moerman (2022). More specifically, the first part of the interviews was structured, while the second part was unstructured. At first, the topic of E-commerce distribution service decision-making was elaborated on to the interviewee, stating that this is our field of interest and, consequently, that we research the potential of BAIT to enhance decision efficiency. Second, we posed a predefined set of questions to the interviewee. In this regard, the sequence of questions was predefined to ensure a natural storyline throughout the first part of the interview. Once the interviewee answered these questions, we posed an open question to the interviewee in the unstructured interview section. Namely, "Are there any (partly) human-driven decisions within the organisation where expert opinions are gathered to establish an outcome?". Through this question, an investigative conversation was motivated to address the potential applicability of BAIT to one of the organisation's decisions. We chose the semi-structured approach first to compare the outcomes of the structured part of the interview among the different expert interviews. And second, urge the interviewee to start thinking about BAIT's applicability for the unstructured interview part. The predefined list of questions can be found in section A.3.

2.2.1. Logistic Service Providers

We interviewed a total of three LSPs. LSPs facilitate day-to-day operations in the fulfilment network between the E-tailer and the customer. Therefore, we expect most of the LSP's decisions to lie within the operational sphere. The predefined list of questions is twofold. First, we ask several questions regarding the logistical set-up of each LSP. By doing so, we understand what LSPs fulfilment networks look like, which is the result of particular strategic decisions. Is there a distinction between regional and online DCs? Which last-mile delivery methods are used to serve the customer? Do you use regional mobility hubs to facilitate the last mile of deliveries? These are examples of questions that we asked. After the identification of the fulfilment network, we posed multiple operational questions. How are routes and tours scheduled and planned? How is it decided which parcels are fulfilled through which fulfilment centre? And, How is it decided which customers are served by which last-mile transportation mode?

LSP number one is a relatively small but rapidly growing logistic operator in the Dutch e-commerce market that facilitates logistic operations for various Resellers and Webshops. We interviewed the Routing Specialist through Microsoft Teams. Parcels are distributed, first, through one central distribution terminal and second, through several decentralized smaller distribution terminals. The customer's geographical location supports the choice of one of several decentralized distribution terminals. Last-mile delivery mode and delivery time are decided directly by the customer, who chooses either home delivery or pick-up point collection. A distinction is made between different time slots and attended or unattended deliveries within home deliveries. Three modes are used for home deliveries: electric delivery vans, bio-fuel delivery vans and cargo bikes. Pick-up points consist of automated parcel lockers. These points are not offered nationwide; therefore, availability depends on customer location. In the upstream part of the network, only one fulfilment channel exists. Ordered parcels are picked up or delivered from the supplier's DC to the LSP's Centralized Distribution Centres (CDC). In total, this LSP utilizes two different fulfilment channels, being: regional DC to the customer and regional DC to parcel locker to the customer. Figure 2.1 shows the fulfilment set-up of the LSP, accompanied by an operational decision flow.

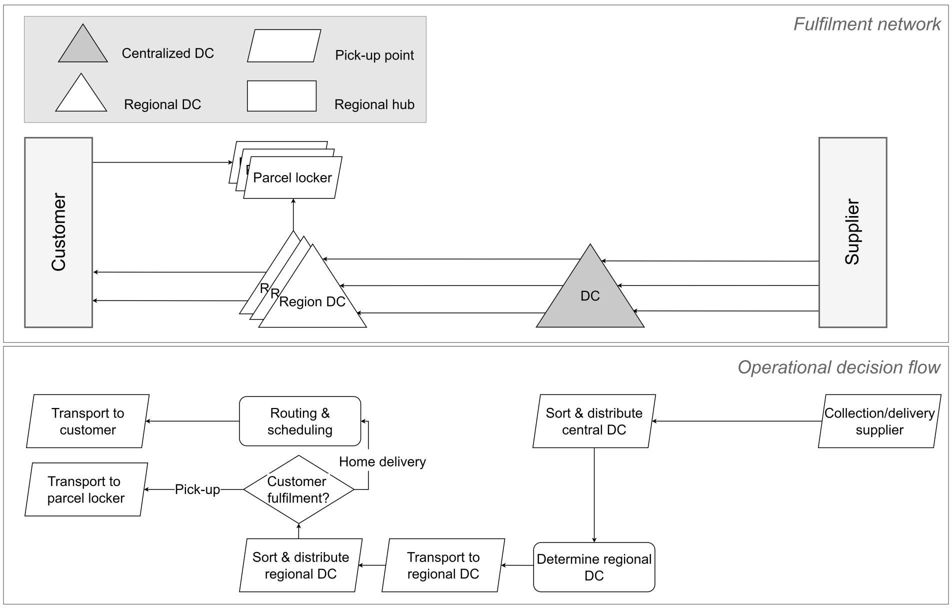


Figure 2.1: Fulfilment network with operational decision flow LSP 1

Regarding the strategic decision sphere, fulfilment network decisions and pick-up point partner contracts are considered. Decisions on regional DC locations are based on customer demand pattern data to serve a nationwide network. Regional DCs are positioned so LSP 1 might serve every customer in the Netherlands. In addition, major Dutch convenience stores are partnered with to establish pick-up points, which increases the network's coverage.

Regarding the tactical sphere, this LSP does not own any product. Hence no assortments and inventory decisions are made.

An operational decision flowchart is added underneath the fulfilment network in Figure 2.1 to address each parcel's decision flow. Only one CDC is identified in the fulfilment network's upstream part. Therefore, no channel allocation decisions are made there. In the downstream part of the network, a regional DC is chosen for fulfilment by the customer's location. This metric is used to reduce last-mile transportation distance and, therefore, transportation costs. When a regional DC is chosen for the fulfilment, parcels are transported to this specific DC and sorted & distributed for customer fulfilment. Whenever the customer selects home delivery, a machine learning algorithm is deployed to determine ordered parcels into delivery routes. This process is almost wholly optimized. Only the characteristics of the order are filled in by the Routing Specialist, which allows the algorithm to generate routes. Also, the operator can adjust the algorithm's performance rate, which is used to build a delay margin during periods of high demand.

LSP number two operates on an international scale and owns multiple service centres across the Netherlands. Since we spent a day at one of this LSP's service centres, we performed no formal interview. During the day, we spent time with the Supervisor Operations, Head Dispatcher and Operations Manager of the LSP. Regarding the fulfilment network, orders are processed iteratively through domestic hubs, service centres and mobility hubs. In the upstream network, domestic hubs are supplied by international suppliers. From there, parcels are transported to regional DCs (i.e. service centres) based on customer location. Each day comprises a morning- and a midday shift in which parcels are unloaded, sorted and distributed within the service centres. After finalization, parcels are transported to a regional hub or delivered directly to the customer by delivery vans. Regional hubs near densely populated city centres deploy cargo bikes to facilitate home deliveries. The decision as to whether a parcel is transported directly or indirectly depends on its size and weight. Regarding the last-mile execution, the LSP offers home delivery and pick-up collection. Home deliveries are executed through a conventional delivery van, electric delivery van or cargo bike. Parcel pick-ups are facilitated by automated parcel lockers or an attended pick-up point. The retailer decides whether customers can choose between sustainable or conventional delivery. Four fulfilment channels are identified: regional hub to customer, service centre to the customer, customer to parcel locker and customer to the pick-up point. These channels establish the fulfilment network, shown in Figure 2.2.

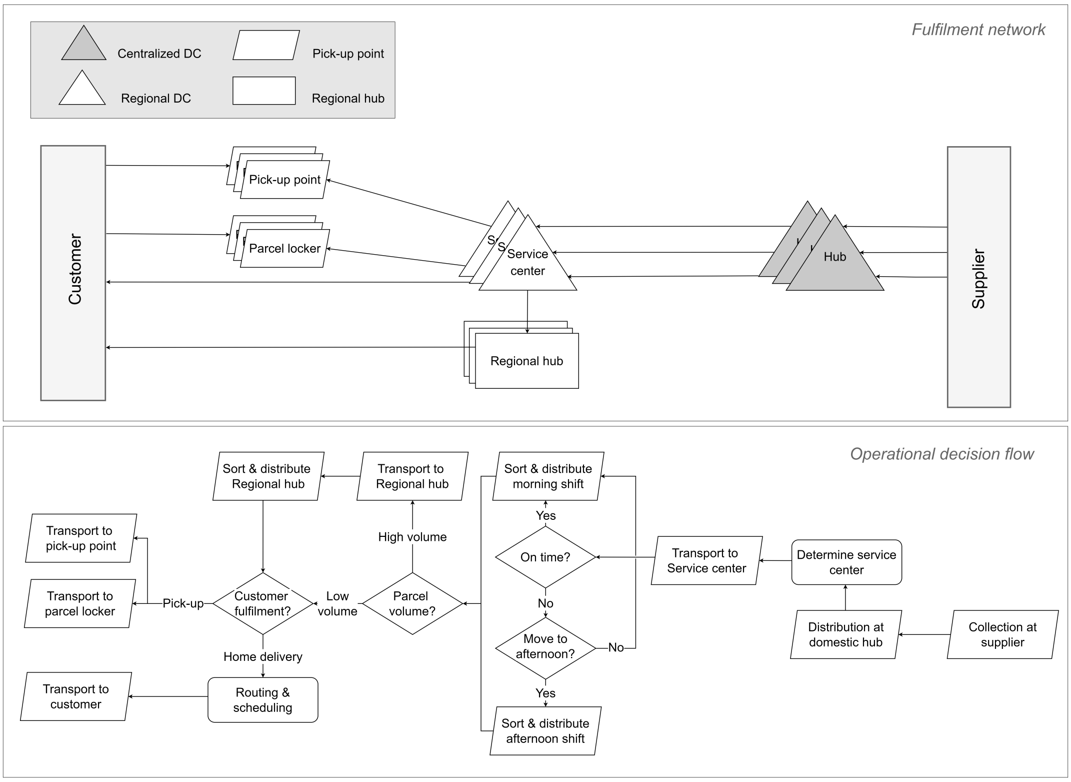


Figure 2.2: Fulfilment network with operational decision flow LSP 2

Regarding the strategic decision sphere, fulfilment network decisions and pick-up point and parcel locker partnerships are considered. Regarding the expansion of the network, new service centres and regional hub locations are decided by multiple factors like property price and employee accessibility. However, customer demand pattern data is decisive. Moreover, contract tenders are posed to conventional stores to establish parcel lockers and attended pick-up locations.

Regarding the tactical sphere, no inventories or assortments are owned, and thus we identified no decisions.

An operational decision flowchart is indicated underneath the fulfilment network in Figure 2.2. In the upstream part of the distribution network, parcels are distributed from a domestic hub to one of the service centres. The specific service centre of fulfilment is decided per parcel based on the postal code of the end customer. After determination of this, parcels are transported to a designated service centre. Every week, incoming service centre deliveries get delayed due to upstream network delays or traffic jams. Whenever this is the case, Supervisors Operations decide to delay the morning shift to be able to include the delayed parcels or move them to the afternoon shift. Next, all packages included in the morning shift are sorted & distributed. During this process, large (and heavy) parcels are sorted for direct transport from the service centre, while low-volume parcels are sorted for indirect transport through regional hubs. From there, parcels are transported to customers by cargo bike.

In the downstream part of the network, the type of customer fulfilment is directly chosen by the customer. Regarding home deliveries, routes are predefined based on historical demand patterns. However, new home delivery orders do not exactly arrive on these predefined routes. Therefore, Supervisor Operations assign new home deliveries to one of the predefined routes. To be able to do so, sufficient knowledge is required of the accessibility of the service area. For instance, a new home delivery order is located outside a city's densely populated inner area. Assigning that order to the route that serves the city's densely populated inner area requires significant egress and back-in access time, which delays existing orders. Instead, it is more efficient to assign the order to another route farther away in distance but with less traffic to overcome. This example indicates that the Supervisor Operations require prior knowledge for successful route assignments.

LSP number three operates a nationwide distribution network of more than 20 fulfilment depots. We interviewed a Senior Logistic Engineer through Microsoft Teams. The fulfilment network consists of depots, a synonym for DCs. These depots are not distinguishable based on process characteristics and thus operate similarly. Each depot has two comprehensive processes, being: sorting of parcels and distributing of parcels. Parcels are collected at suppliers and are forwarded to a particular depot for the first sorting and distributing phase. Packages are distributed directly to the customer or to a second depot, where a new sorting and distribution round is performed. The decision to include parcels in the second sorting and distribution round is based on the supplier's geographical location and the customer's postal code. Between the final depot and customer, parcels are either home delivered to customers or at a pick-up point for collection. Within home deliveries, packages are delivered by delivery vans, electric vans or cargo bikes. Whether the customer has a choice for sustainable delivery depends on what the retailer decides to offer their customers. Customers can choose between attended pick-up points or automated parcel lockers when they decide on a pick-up. Altogether, six fulfilment channels are offered to fulfil customers: depot to the customer, depot to the pick-up point and depot to parcel locker, depot-to-depot to the customer, depot-to-depot to the pick-up point and depot-to-depot to parcel locker. Figure 2.3 shows a visual presentation of the fulfilment network.

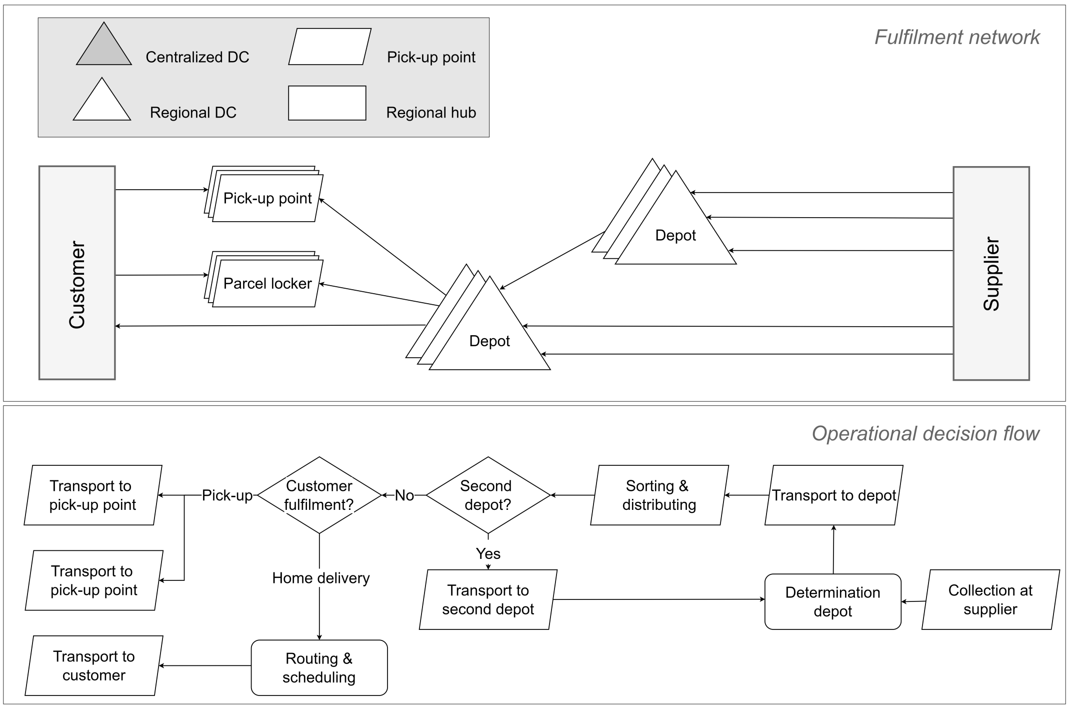


Figure 2.3: Fulfilment network with operational decision flow LSP 3

Regarding the strategic decision sphere, expansion of the fulfilment network through opening new depots and contracting pick-up points and retailers are considered. The opening of new depots is predominantly based on customer demand patterns. By incorporating this key metric, a nationwide network is established and maintained. Although this metric is used chiefly, expanding the network is always a best practice involving many factors. Also, multiple stakeholders are involved, which makes the decision increasingly complex. Moreover, local supermarkets, construction markets and postal offices are contracted to establish pick-up points. Several characteristics like distance to the closest pick-up point, pick-up point capacity and pick-up point opening hours are considered in this regard. In terms of product supply, retailer partnerships are established. These one- or multiple-year contracts entitle the facilitation of logistic operations for a particular retailer. Costs per operated parcel, parcel volumes and the applicability of the retailer's operational resources to own logistical processes are considered.

In regard to the tactical sphere, LSP 3 does not own any inventory or assortments and thus no decisions.

Figure 2.3 indicates the fulfilment network accompanied by an operational decision flow. As can be seen, in the upstream part of the network, parcels are either distributed through one or two depots. This choice is based on geographical locations of both supplier and customer and the available capacity of the depots. Next, parcels are transported to the depot after which sorting & distributing to delivery modes begins. In the downstream network, customer fulfilment is directly chosen by the customer. Whenever a home delivery is selected, routes and schedules are predefined based on historic delivery patterns. However, urban and rural delivery routes require different routing & scheduling constraints, and thus, human control is required. Moreover, particular events might arise that extent delivery times when not identified. For instance, road construction events, car free inner city zones and low speed areas pose risks for meeting delivery times. To cope with these developments, human interference is required, by which new orders are assigned to existing routes.

2.2.2. Retailers

In total, we interviewed two retailers. Retailers can be classified as regular Retailers, offering only offline fulfilment channels or E-tailers, offering both online and offline fulfilment channels. Both the interviewed Retailers classify as E-tailers. E-tailers are known to outsource part of their logistic operations and potentially specify specific assortments and inventories to particular channels. Therefore, we expect E-tailers to encounter strategic, tactical, and operational decisions. The interviews are structured similarly to the ones conducted at LSPs.

Retailer number one is an internationally renowned furniture seller. We interviewed by phone with the Planning Fleet Operations Manager. Regarding the fulfilment network set-up, orders are collected at the supplier, from where products are transported to a CDC or directly to one of the stores. In the downstream part of the network, customers can choose two ways to purchase and collect products. These are collections at the store and home delivery. Collection at the store is distinguished between click & collect and regular store visits. An electric or conventional delivery van performs home deliveries. In this regard, customers cannot pick one of the two, which leaves this decision to the E-tailer. In total, four fulfilment channels are operated: click & collect, store visit and home delivery. However, each channel potentially incorporates cross-docking at the CDC within the upstream network. Thus, a total of six channels are identified.

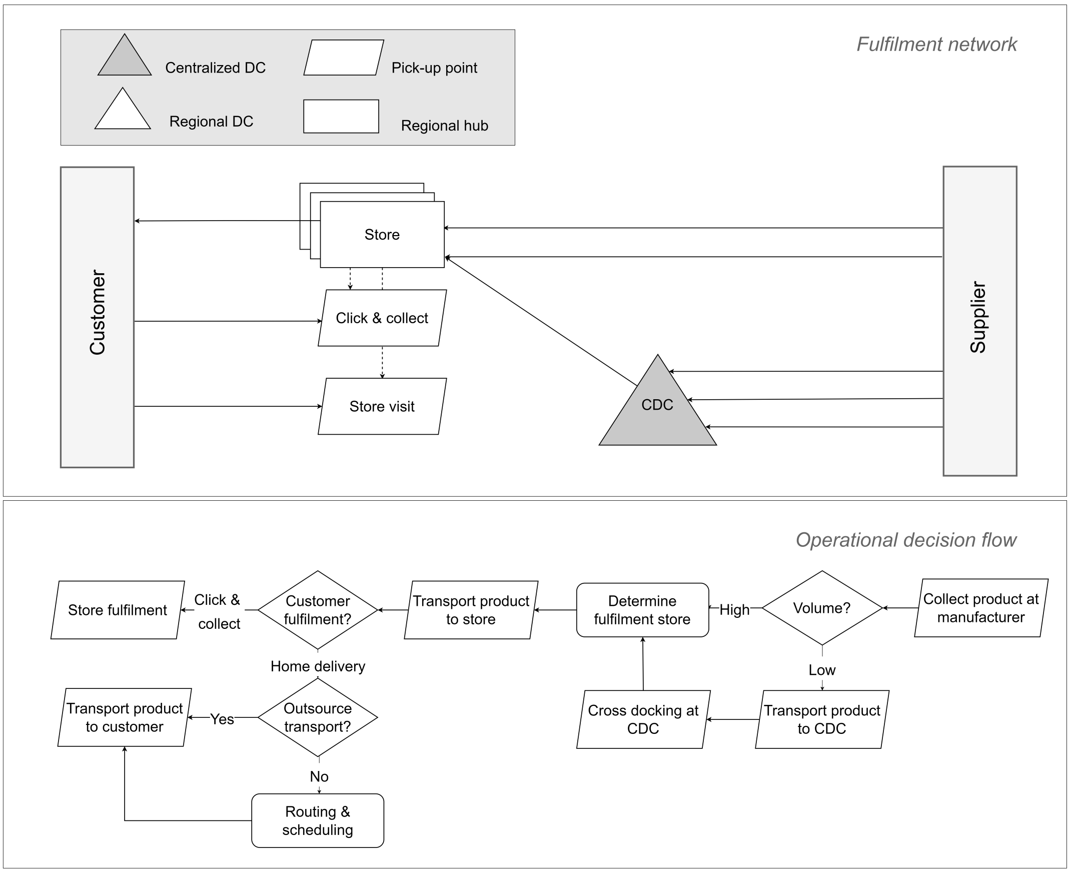


Figure 2.4: Fulfilment network with operational decision flow Retailer 1

Logistic operations are partly organised by owned delivery assets and partially outsourced to a contracted LSP. Tenders are posed for LSP selection to establish logistic partnerships. By trading off several criteria, it is decided which LSP is handed the performance rights to facilitate logistic operations for a designated area. Alignment with the retailer's sustainability goals, transport volume capacity, service area and transport costs are among the decision criteria. Moreover, new store locations are decided based on customer demand patterns in the strategic sphere.

Within the tactical sphere, physical store assortments are constrained according to size. Online channels, on the contrary, offer the entire assortment for home delivery and indicate real-time in-store stock information. Moreover, the interviewee mentioned that a shift to a fragmented showrooming model is being deployed. What is meant by this is that existing stores pivot to a DC purpose for home deliveries, and smaller stores are opened within city centres for showrooming purposes. Customer shopping behaviour implies opening more stores in high shopping demand areas. Moreover, separate inventories are maintained per store, so no inventory pooling strategies are deployed.

Figure 2.4 indicates the fulfilment network accompanied by the operational decision flow. In the upstream network, product volume decides whether a product is handled through the CDC or not. Low-volume products need to be combined into efficient batches to construct store deliveries. The cross-docking process aims to maximize the use of space within delivery batches per store. High-volume products don't need cross-docking and are directly moved to stores. In the downstream network, the fulfilment store is decided through the postal code of the end customer. Products are delivered to a physical store nearest the customer to minimize transport distance. After arrival, parcels are sorted and either collected by customers (click & collect), delivered to their doorstep (home delivery) or purchased through regular store visits. Customer fulfilment is directly chosen by the customer, while outsourcing of transport depends on the service area. Whenever orders are delivered by owned delivery means, routing & scheduling decisions have to be made.

Retailer number two is a nationally renowned E-tailer selling various products. Customers are served through online and offline channels, in which physical stores are merely an extension of the online channels. Although this is the case, customers might use physical stores to buy products directly or as a showroom to try products. We interviewed the Owner Network Efficiency through Microsoft Teams. Regarding the fulfilment network, goods are collected or delivered by the supplier to a centralised DCs. From there, order batches are composed and transported directly to the customer, a depot, or a regional hub. In the downstream network, customers select either home delivery or pick-up collection. Home deliveries are executed by conventional delivery van or by cargo bike. The latter is being deployed from regional hubs. In this regard, E-shoppers cannot select one of the two fulfilment options. The E-tailer selects the mode of transport and depends predominantly on order size. Whenever a pick-up collection is selected, customers might collect parcels in three ways: attended pick-up, pick-up at a parcel locker, and click & collect at a physical store. In total, six fulfilment channels are utilised: CDC directly to customer, CDC to depot to store to customer, CDC to depot to pick-up point to customer, CDC to depot to parcel locker to customer, CDC to regional hub to customer and CDC to depot to regional hub to customer. Figure 2.5 visualizes the fulfilment network of Retailer 2.

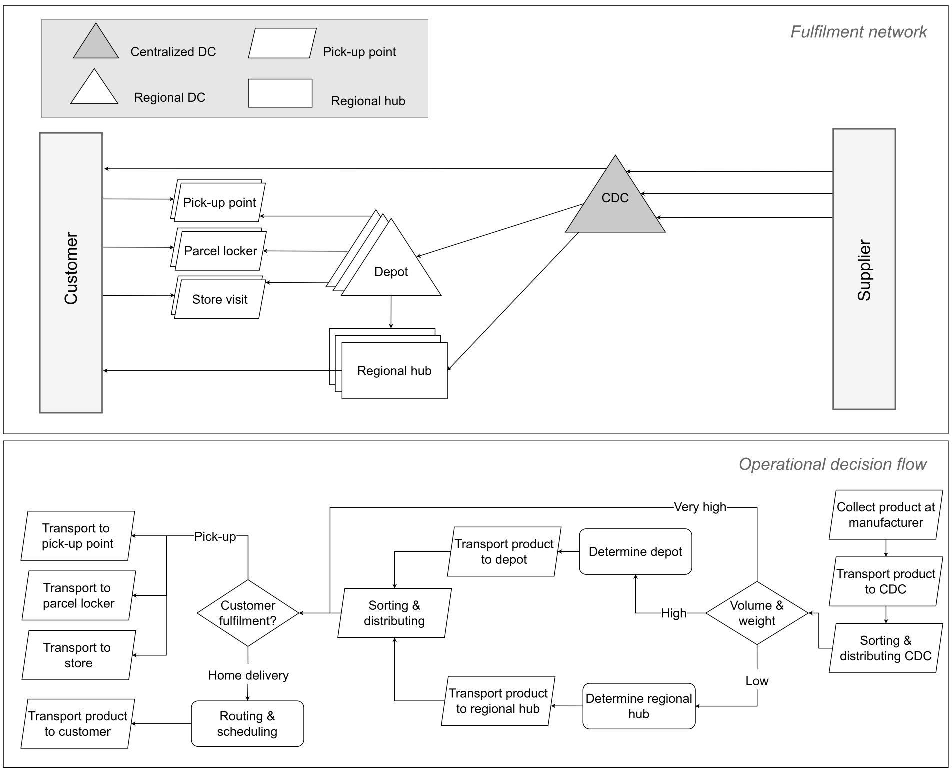


Figure 2.5: Fulfilment network flowchart with decision flow Retailer 2

Regarding the strategic sphere, network expansion decisions are based on several factors. New depots and regional hubs are located based on customer demand patterns, property prices, accessibility to the service area(s) and the accessibility for employees. According to the interviewee, extending the network is always a consensus that depends on the before mentioned factors. In addition to network expansion, several contracting decisions are made. Regarding high-volume packages, a nationwide network is facilitated by its owned logistic operations. However, logistic operations are partly outsourced to a LSP for low-volume parcels, which are delivered by cargo bike. Moreover, contracts are established with suppliers, pick-up points and parcel locker providers.

Within the strategic sphere, the online channel offers a full assortment range. As an extension to this channel, physical stores are solely supplied for click & collect purposes and do not inhabit conventional store shopping. In terms of inventory management, a centralized inventory strategy is followed. Inventories of all products are managed and maintained in the CDC. Product stock levels are decided and maintained based on demand forecasts.

Figure 2.5 indicates an operational decision flow alongside the fulfilment network. At first, all products are gathered at the CDC. After a sorting & distributing round, order batches are composed and either transported to a depot, regional hub or directly to the customer. This distinction is based on product volume and weight. Regional hubs deploy cargo bikes for customer fulfilment, which introduces a specific volume constraint that limits the transport of high-volume parcels. On the contrary, whenever products are heavy and take up a lot of space, they are directly delivered from the CDC to the customer. Parcels of average volume and weight are distributed through depots. The specific depot or regional hub used for fulfilment is decided by customer location. In the downstream network, customer fulfilment is directly chosen by the customer as a pick-up or home delivery. Concerning home deliveries, a routing & scheduling step is performed based on historical data.

2.3. Conceptual framework E-commerce distribution services

In this chapter, first, literature was consulted to identify the different decision hierarchies that comprise E-commerce distribution services, being: *Strategic*, *Tactical* and *Operational*. Second, we interviewed five

E-commerce actors to gather market practices regarding decision-making to identify underlying decisions. These two analyses enable us to answer the **first research question**:

What decisions are included in setting up, managing and executing E-commerce distribution services from an E-tailer and Logistic Service Provider perspective?

In the strategic sphere, decisions are subject to various internal and external factors, which are often difficult to quantify. Therefore, the effectiveness of decisions is hard to track and relies partly on human expertise. Strategic decisions present in e-commerce distribution services are: *Fulfilment network expansion* and *Partner contracting*. In the tactical sphere, decisions serve to organise existing resources to maximise customer satisfaction. These tactical decisions are based on quantifiable factors like customer purchasing data, shopping behaviour and click-stream & browsing data. Decisions are therefore made by humans, under consideration of available data and supported by data analysis. Tactical decisions present in e-commerce distribution services are: *Channel assortments* and *Inventory planning*. The presence of these decisions depends on the degree of assortment ownership. As obtained from the interviews, product assortments are generally owned by the E-tailer; thus, tactical decisions are theirs to make. The operational decision sphere inhabits day-to-day decisions. Either these decisions are made solely by the E-retailer or (partly) outsourced to LSPs. To help decide, human operators are assisted by IT and planning software which provides routing propositions according to historical delivery patterns. Operational decisions in day-to-day operations are: *Channel allocation*, *Routing & scheduling* and *Last-mile mode choice*. Below, we provide the conceptual framework that encapsulates all three decision spheres and the decisions present within them.

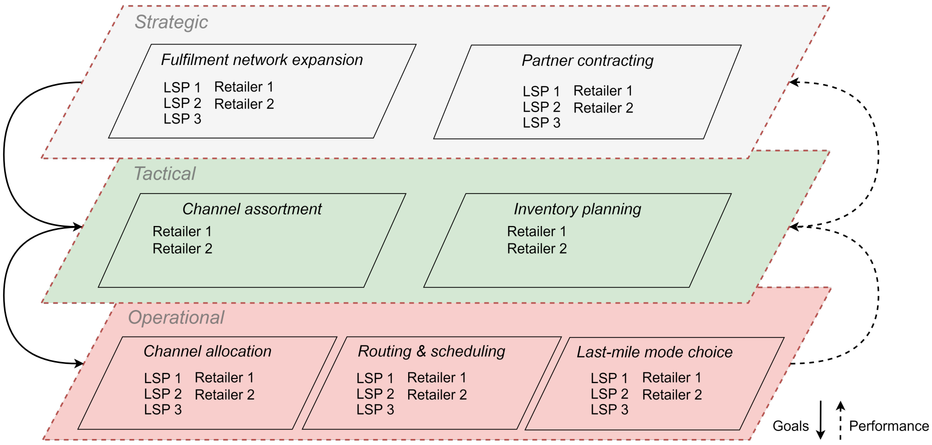


Figure 2.6: E-commerce decision hierarchy comprising decisions from market practices

The three-part decision hierarchy, as bespoken above, is visualized within the conceptual framework. The three planes correspond to the different decision hierarchies. Within these decision hierarchies, the lower-level decisions are indicated. The connecting solid and dashed arrows alongside the decision hierarchies indicate the interdependence among them. The point of interdependence between decisions was elaborately made by Riopel et al. (2005), and although it is not the focus of this research, we discuss it swiftly in the next section.

2.3.1. Causality of decisions

In the figure above, we can identify two streams of arrows next to decision hierarchies and environments. The downward-facing arrows are strategic decision outcomes that seep through the tactical sphere onto the operational sphere. By doing so, lower divisions within the organisation are motivated to organise tactical and operational processes in alignment with higher strategic goals. In the opposite direction, operational performance works its way up from the operational sphere, through the tactical sphere into the strategic sphere. These arrows visualise how operational outcomes influence the strategic decision sphere through the tactical sphere. It shows us that companies use performance in the lowest decision sphere (operational) to adjust goals from the organisation's top. This implies that decision hierarchy and their lower-level decisions can not be regarded as separate, and improvements in the decision quality in one sphere might result in improvements in higher or lower decision spheres. Although this is an important aspect to incorporate when improving the decision process, it is not within the scope of this research and therefore left out in the remainder.

2.4. Behavioural AI Technology to improve decision making

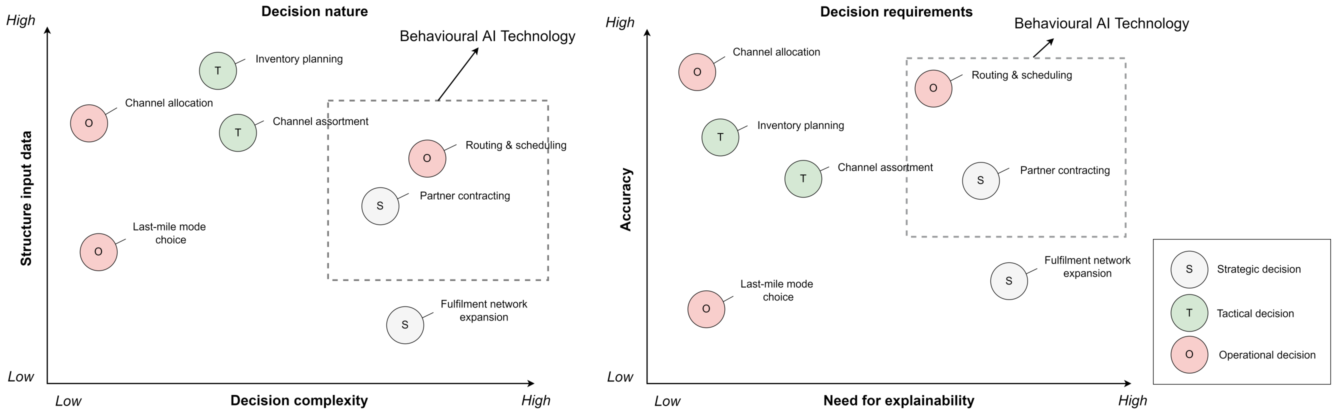
According to our review of existing decision support mechanisms in subsection 1.1.4, decision criteria weights are frequently defined by direct human judgement. Although this is the case, explicit judgements potentially have inherent biases and risk bad decisions. A new method, BAIT, defines criteria weights by decision-makers' choices instead of direct judgement. According to ten Broeke et al. (2021), this method offers a "simple and explainable decision model" that does not require historical data. As described in subsection 1.1.4, BAIT might arguably be seen as a method to increase the reliability of decision criteria weights. Therefore, in this section, we will assess the applicability of BAIT to e-commerce distribution service decision-making. We do this by a two-step classification of the decisions identified by the expert interviews.

2.4.1. Classification of E-commerce distribution services decisions

The conceptual framework of section 2.3 indicates seven decisions in e-commerce distribution services. To classify these about their subjectivity to BAIT, we rank each decision based on its nature and requirements.

A *decision's nature*, on the one hand, inhibits a level of complexity. According to Sintchenko and Coiera (2006), decisions that inherit high complexity require more information processing and, thus, more cognitive effort. In contrast, low decision complexity indicates less information processing and, consequently, a lower mental effort is asked from the decision-maker. Moreover, a certain degree of structured input data is needed to enhance the decision outcome. In this regard, a high degree of structured input data signifies that an accurate decision outcome is merely based on the input data's availability and degree of structure. On the contrary, a low degree of structured input data indicates that the result does not depend solely on structured input data but also on, for example, human expertise.

In terms of a *decision's requirements*, each decision outcome aims for the desired accuracy level, accompanied by a certain level of explainability. This trade-off was described by London (n.d.), who noted that a lack of explainability increases the risk of bias-carryover from the training data into the predictions of the actual decision model. Intended high decision accuracy indicates that decision outcomes must inhabit a high standard of correctness. This is desired for decision outcomes which are highly visible or lead to significant losses when inaccurate and therefore have small margins of error. On the contrary, less visible and less error-sensitive decisions require lower accuracy, for instance, because the outcome is a product of consensus (which makes them sub-optimal by nature), information is incomplete, or the intended decision outcome is difficult to measure. In decision explainability, a high need indicates that decision-makers should be able to explain which factors were decisive for the decision outcome. Whenever this is the case, for instance, a high degree of explainability is desired to convince other employees. On the contrary, low explainability indicates that insight into criteria trade-offs is less critical.



(a) Classification based on decision nature

(b) Classification based on decision requirement

Figure 2.7: Applicability BAIT to decisions

By means of the classifications on decision nature and decision requirement, the **second research question** can be answered:

Where in this decision sphere might Behavioural AI Technology be of added value?

Figure 2.7 entails classifying e-commerce distribution service decisions based on each decision's nature and requirements. Regarding decision nature, BAIT is most applicable to decisions with a medium to a high level of complexity and a medium degree of structured input data. In contrast to knowledge-based (or rule-based) support systems, which apply to low-complexity decisions, the BAIT method collects decision-makers expertise by assessing a multitude of decision-maker choices. As a result, more subtle trade-offs, often inherited in complex decisions, can be captured (ten Broeke et al., 2021). Moreover, BAIT allows the incorporation of criteria constraints, non-linear criteria, interaction effects and different model types corresponding to the choice behaviour of the expert group. These factors all contribute to the applicability of BAIT to complex decisions. In terms of the degree of structured input data, different from non-knowledge-based (or machine learning) support systems, which require training on vast amounts of historical data, BAIT creates its input data through choice experiments. Although this data must be collected and structured, it is created especially for its purpose; therefore, no external input data is required. As a result, BAIT applies to decisions that do not require a high degree of structured input data to achieve the desired outcome.

Figure 2.7b classifies BAIT on its requirements; BAIT allows high decision explainability and a medium to high decision accuracy. Criteria importance is defined by criteria weights, the outcome of analysing the choice experiment results. As a result, model predictions are invariably traceable to single decision criterion valuations, making the decision fully explainable. Besides being explainable, BAIT reaches medium to high levels of accuracy. A validation step is performed in which model predictions are compared to the outcomes of the choice experiment. Several model fit metrics are regarded to assess the model's empirical performance (ten Broeke et al., 2021).

In the first sections of this chapter, seven decisions are concluded that establish e-commerce distribution services. These decisions are indicated above in Figure 2.7. As can be seen, *Partner contracting* decisions and *Routing & scheduling* decisions fall within the spectrum of BAIT. Figure 2.7a indicates that partner contracting decisions depend partly on structured input data. For example, to cope with increased customer demand, E-retailers partner up with new pick-up points based on various criteria like proximity of existing pick-up points, storage capacity and opening hours. These factors are subject to quantification and thus usable as data input. Similarly, E-retailer and LSP partnerships are established by quantifiable factors like transportation price and parcel volumes. However, the compatibility of the E-retailer's operational assets with the LSP's is more difficult to quantify, thus adding to the complexity of the decision. Moreover, negotiations are often part of partner contracting decisions, which signals the incorporation of human judgement and, thus, trade-offs which entail human expertise. This notion indicates that partner contracting decisions inhabit a degree of complexity that might not be captured by (non-)knowledge-based models. Like partner contracting decisions, routing & scheduling decisions require a medium degree of structured input data. A selection of quantifiable input factors is the geographical location of the customer, available transportation assets and the urgency of delivery of parcels. Nevertheless, as seen in practice, these factors only guide the human operators in the process. The importance degree of each criterion is based on human expertise, hence adding to the complexity of the decision.

Figure 2.7b indicates, in terms of decision explainability and accuracy, again, both *Partner contracting* and *Routing & scheduling* fall within the spectrum of BAIT. Especially Routing & scheduling decisions require high accuracy in decision outcomes. Wrong choices result in customer delivery delays and thus are directly visible to the customer. Explainability is also important for Routing & scheduling decisions since they need to be explainable to the delivery drivers. Regarding Partner contracting decisions, accuracy is deemed medium, while explainability needs are high. Decision outcomes of new partnerships are challenging to measure in terms of accuracy because, on the one hand, they have a long-term nature, which makes performance difficult to measure. On the other hand, if another partnership was established instead, alternative performance would be unknown. Additionally, in terms of outcome explainability, contracting new partners is based on the valuation of a set of criteria. This valuation should correspond with company goals. Therefore, the explainability of this criteria valuation is essential to align choices with company goals.

2.4.2. Conclusion on applicability BAIT

Within this chapter, first, we performed a literature review to identify a three-step decision hierarchy. Second, we interviewed experts of industry players to find decisions within these decision hierarchies. Third, a conceptual framework was composed that might be used as a classification tool to link actors to decisions within a fulfilment network. Fourth, we assessed the applicability of BAIT to these decisions by the decision's nature and requirements. In this final section, we conclude the applicability of BAIT to model e-commerce distribution services. By doing so, one particular decision is chosen that we will model with BAIT in the remainder of this research using a pilot case study. Finally, according to the structure of this research, we answer the **third research question**:

What is the most suitable decision to test the application of Behavioural AI Technology in the E-commerce sector?

Altogether it can be concluded that BAIT is most applicable to *Partner contracting* and *Routing & scheduling*. Both decisions rely partly on structured input data but largely on human expertise, contributing to the complexity of the decisions. Out of the two decisions, due to a practical reason, it is chosen to continue this research by modelling delayed line haul decisions, which are a sub-decision of Routing & scheduling decisions. DHL Express gave us the option to model these choices with the cooperation of the firm's expert group. The reader must understand that BAIT's methodology is new in the E-commerce sector, and therefore applying BAIT in a practical sense is essential to assess its purpose and performance, as well as its applicability to E-commerce. In no way do we argue that delayed line haul decisions are the only appropriate application of BAIT in e-commerce or that only the operational decision sphere is subject to modelling with BAIT. However, noting the arbitrary approach of this research and the opportunity to test BAIT and gain practical experience, we decided to model the delayed line haul decision.

3 Methodology

In this chapter, we propose the methodology of this research. The goal is to gain insight into the criteria trade-offs made by DHL's decision-makers in regard to delayed line haul deliveries. To acquire this insight, we use Behavioural Artificial Intelligence Technology (BAIT), which incorporates semi-structured interviews for criteria identification and choice modelling practices to estimate criteria weights. At first, the method BAIT is dissected below into six fundamental steps. By doing so, every aspect of the method is touched upon and elaborated. Each of these six steps is applied on delayed line haul decisions in chapter 4 & chapter 5. Second, essential Discrete Choice Analysis (DCA) theory is discussed, which is a fundamental part of the modelling process of BAIT. A part of DCA is the collection of preferences through a choice experiment. To be able to construct this experiment, thirdly, we discuss particular designs that ensure most efficient choice experiments. Lastly, we elaborate on the choice experiment's composition to indicate the building blocks and methods for composing high trade-off value experiments.

3.1. Behavioural AI Technology

BAIT is a new method that incorporates a behavioural component in criteria weight estimation. As mentioned in subsection 1.1.4, BAIT consists of a *descriptive* and a *prescriptive* part. In the descriptive part, Discrete Choice Analysis (DCA) is utilized to derive decision criteria valuations from Stated Preferences (SP). This is different to existing Multi-Criteria Decision-Analysis (MCDA) methods, which predominantly use a sophisticated method of direct judgement for preference elicitation (Table 1.1). Next, the method mimics decision-makers by incorporating the SP-obtained criteria weights in an interactive prediction model. The second part of the method uses this prediction model to create introspection and open a discussion among the decision-makers to potentially adjust criteria weights. This conversion of descriptive weights to prescriptive ones makes the BAIT methodology complementary to conventional Discrete Choice Models (DCM) as elaborated by ten Broeke et al. (2021). Although the way to convert descriptive criteria weights to prescriptive ones is interesting for further research, we purposely left it out of our research. In this section, we describe the necessary 6 steps to establish the descriptive part of BAIT:

1. **Decision scoping**: A group of (2-3) experts is interviewed once (1,5 hours) to clarify the decision of interest. This process includes decision-maker and decision outcome identification, information availability, decision context and decision time. After identification, boundaries are applied to include and exclude aspects from the modelling process. Thus, the decision is scoped.

2. **Decision characteristics inventarisation**: The same group of (2-3) experts are interviewed twice (1 hour each) to compose a long list of decision attributes (= criteria). The list is deemed complete when the interviewees are satisfied, and a maximum of 15 attributes are identified. After identification, attribute ranges are decided that indicate the maximum and minimum values to which the attribute can vary. Next, the attribute range is divided by 2, 3 or 4 to establish multiple attribute levels. Alongside range determination, knockout levels might be identified. These are levels that overrule other criteria and, thus, determine the decision (e.g. extremely cheap alternative that will overrule all other attribute levels). Next, constraints might be identified that preclude combinations of attributes that are non-existent

in reality or highly unlikely to take place. Lastly, experts rank each attribute as crucial, important or nice-to-have.

3. **Design choice experiment**: By these decision attribute characteristics, a specific choice model is designed that is applicable to this particular situation. For example, the accommodation of non-linear weights (increasing or decreasing marginal importance) and interaction effects (positive or negative addition to a combination of attributes) are particular characteristics that define the choice model design. Depending on these and other attribute characteristics, a particular choice model set-up is chosen. Next, the choice experiment construction software Ngene is used to construct a choice experiment accordingly. In section 3.3, specific choice model designs are elaborated in detail.
4. **Choice experiment execution**: After pinpointing the choice experiment design and generating the actual choice experiment, it is forwarded to a group of approximately ten experts. The choice experiment includes 30 choice scenarios, consisting of a short context description and a combination of the predefined attributes and attribute level ranges. In each situation, the respondent is asked to choose one of the multiple decision alternatives. The choice experiment is filled in online and takes around 30 minutes to complete.
5. **Choice model estimation**: The resulting survey data is processed with the choice modelling Apollo library in the programming language R. In this iterative process, attribute weights are adjusted to match the model's prediction to the actual choices made by the experts. This process finishes when no improvements can be made and various model fit metrics are satisfied. As a result, the most reliable parameter estimates (i.e. criteria weights) are obtained.
6. **Model validation**: Finally, the results are presented to the expert group by visualizing the contribution of each decision attribute to the overall decision. Accompanied by this, a predictive choice model is presented. This model allows probabilistic assessments to be made by adjusting attribute levels to a specific choice situation. An example of such a statement in a medical context is indicated by ten Broeke et al. (2021): "The probability that an expert that is randomly sampled from the expert group would recommend (to the patient's parents) to perform surgery on a patient with this profile equals 18%."

BAIT relies on DCA to estimate criteria weights within the modelling process. DCA is a highly explored academic field with lots of practical applications. Next, we describe the leading theory to estimate choice models.

3.2. Choice model estimation

The concept of choice modelling finds its background in DCA. This quantitative approach assumes that actors make rational choices by maximizing their utility. With DCA, respondents face choice situations which consist of different choice alternatives and underlying choice criteria. By monitoring these choices, Stated Preferences (SP) are derived and consequently used to explain and predict the behaviour of the respondent group. In other words, specific criteria trade-off valuations, which hold at a given moment in time, are made visible by DCA (Chorus, 2015). Besides SP analysis, one can also use Revealed Preferences (RP) to gather criteria valuations. Although this method usually results in high model validity, it is impossible to introduce new alternatives of criteria levels that were not part of previous decisions. We use SP to model Supervisor choices because no historic choices are monitored regarding delayed line hauls.

The utilities of choice alternatives are measured through so-called utility functions. These parametric functions consist of observable independent variables and unknown parameters (Akiva and Lerman, 1985). The parameter values are estimated through samples of observed decisions made by choice respondents in particular choice situations (i.e. the choice experiment). When these values are incorporated into a predictive model, the resulting product is a Discrete Choice Models (DCM). However, it should be noted that due to certain unobserved randomness, the true utility of each alternative is never fully known. Therefore, DCM predictions will never be the same as respondent choices and only pose a probability that a specific alternative is chosen according to the preferences of the respondent group (Chorus, 2015). Most common DCMs follow Random Utility Maximization (RUM) theory. According to this theory, choice respondents act fully rational, thus choosing the alternative that has the highest utility. Equation 3.1 shows the composition of the utility function.

$$U_i = V_i + \epsilon_i = \sum_m \beta_m x_{im} + \epsilon_i \tag{3.1}$$

U_i depicts the total utility that a choice respondent can derive from alternative i. V_i is defined as the structural utility, in other words, the utility that we capture with the observed choices, and thus can explain. ϵ_i is the part of the utility that is not captured by observed choices and is thus referred to as the unobserved utility. β_m is the parameter estimate (i.e. criteria weight) that is attached to a specific attribute m. x_{im} is the attribute level of the specific attribute m for alternative i. A crucial computational property to calculate utilities is that the choice probability calculation is closed form (Chorus, 2015). Therefore, the error term is added to the model. Different choice probabilities will be obtained depending on the specification of the distribution of this error term. Usually, the error term is independently and identically distributed similarly to an Extreme Value Type 1 distribution. By using this distribution, the error term approaches a normal distribution, which is generally assumed to correspond most to a population's distribution. This approach is referred to as the linear-additive Multinomial Logit Model (MNL). Equation 3.2 indicates the formula used to calculate choice probabilities according to the MNL.

$$P_i = \frac{e^{V_i}}{\sum_{j=1...j} e^{V_j}} \tag{3.2}$$

P_i is the probability that alternative i is chosen out of the set of alternatives. V_i is, as mentioned before, the structural utility that the model captures. V_j accounts for all the utilities of the other alternatives in the choice set. Ultimately, the MNL method can predict choices to a certain degree of precision. How good a particular model is in predicting choices can be measured by various model fit metrics, which we discuss next.

3.2.1. Model fit

The MNL model produces three indicative model fit metrics, which can be assessed to derive the prediction precision of the model. First, parameter estimates are given, which indicate how much utility is won or lost by an increase of 1 unit of the attribute. Second, two different metrics are estimated to indicate the model's fit, being the Log-Likelihood (LL) and the Rho-square (ρ^2). The higher the LL (i.e. closer to zero), the better the model can predict the dataset. The ρ^2 measure is calculated by Equation 3.3. It indicates the relative difference between the LL of a model where all estimated parameters are zero and the estimated parameters of the final model. Whenever the ρ^2 is 0, the model is no better than a random model (i.e. throwing a dice). However, if the ρ^2 gets closer to 1, the estimated parameters fit the dataset better, hence approaching a deterministic model. Although this is the case, it is important to note that a higher ρ^2 does not always indicate a better model. In the end, it depends on the goal of the model.

$$\rho^2 = 1 - \frac{LL_{model}}{LL_{zero}} \tag{3.3}$$

Third, a Standard Error (SE) is estimated for each parameter. This SE indicates whether a second sample draw, of the same size, out of the population would lead to a different estimate of the parameter. SEs are calculated by Equation 3.4. The higher the number of respondents included in the choice experiment, the lower the SEs. This conforms to the formula below, where N is the number of respondents, and β is the parameter estimate.

$$SE(\hat{\beta}) = \sqrt{\frac{\left[\frac{1}{-\left\{\frac{\delta^2 LL(\beta)}{\delta\beta\delta\beta}\right\}_{\hat{\beta}}}\right]}{N}} \tag{3.4}$$

In addition to the LL and ρ^2 metric, a third model fit metric is the Mean Absolute Deviation (MAD). This metric indicates a percentage that shows the deviation of the predicted choices compared to the actual choices of the sample group. As the outcome of BAIT is given in a percentage of experts that favour a certain policy, it is easily comparable to the percentage of respondents that favour a particular policy in the actual choice experiment. By doing so, the MAD metric indicates the percentage to which the prediction model is off compared to the actual choices of the respondents. Adding and averaging over all the choice sets of the experiment results in a MAD score for the entire choice experiment.

Choice experiments can be designed through different design methodologies, which ultimately impacts the model fit metrics explained above. Therefore, next, we elaborate on these different design methods.

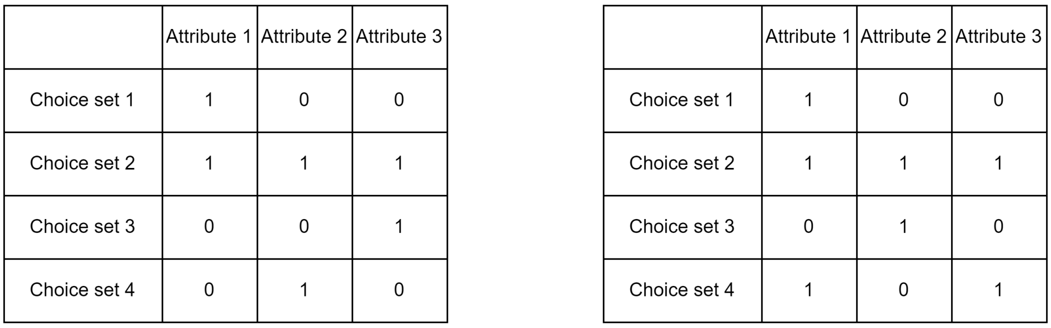
3.3. Choice experiment design

In terms of experimental design, either a *Full Factorial Design* or a *Fractional Factorial Design* might be selected. Full factorial designs inhabit all possible combinations of the selected attributes and attribute levels. The number of choice alternatives is calculated by L^N where L is the number of attribute levels, and N is the number of attributes. Using this experimental design, we can identify interaction effects that indicate dependencies between different attributes (Molin, 2015). Although this is the case, full factorial designs require many choices and thus much effort from the respondent. However, in most cases, not all possible attribute combinations are needed.

Fractional factorial designs inhabit only a share of the total attribute combinations. Fractional factorial designs can be selected in three ways. First, a section might be randomly selected from the complete factorial design. Second, an orthogonal design might be selected in which all correlations between attributes are zero. Last, an efficient design might be selected, which minimizes the SEs of the decision criteria. Each design choice depends on the intended outcome and the means to construct the experiment.

3.3.1. Orthogonal design

Orthogonal designs are a good application when *Multicollinearity* arises within a particular situation. Multicollinearity indicates that attributes are highly correlated with each other and thus interdependent. As a result, estimates are unreliable (i.e. high SEs), which indicates that drawing a different sample of the same size would lead to a differing criteria estimate. As a result, the parameter can not be regarded as significant and thus is not a representative estimation for the sample. To ensure low correlations among attributes (i.e. low SEs), *Attribute level balance* can be preserved in the set-up of choice sets. This means that each attribute level is shown an equal number of times among the attributes. Specifically, this ensures that each attribute level is observed an equal number of times and correlations are kept low. Figure 3.1 shows the difference between a balanced and a non-balanced fractional factorial design.



	Attribute 1	Attribute 2	Attribute 3
Choice set 1	1	0	0
Choice set 2	1	1	1
Choice set 3	0	0	1
Choice set 4	0	1	0

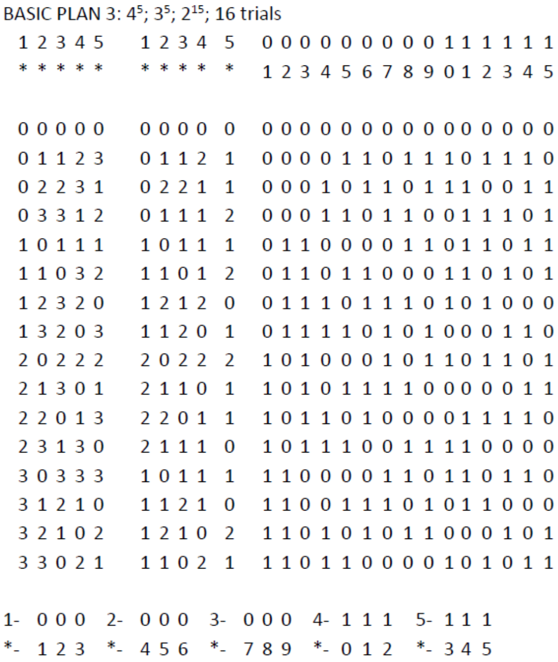
(a) Experimental design with attribute level balance

	Attribute 1	Attribute 2	Attribute 3
Choice set 1	1	0	0
Choice set 2	1	1	1
Choice set 3	0	1	0
Choice set 4	1	0	1

(b) Experimental design without attribute level balance

Figure 3.1: Example attribute level balance

To ensure the orthogonality of experimental designs a *Basic plan* might be used. Basic plans are published in fractional factorial design schemes, combining attribute levels into choice sets to preserve orthogonality and attribute level balance. Figure 3.2 shows an example of a basic plan.



BASIC PLAN 3: 4^5; 3^5; 2^{15}; 16 trials

```
 1 2 3 4 5   1 2 3 4  5   0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
 * * * * *   * * * *  *   1 2 3 4 5 6 7 8 9 0 1 2 3 4 5

 0 0 0 0 0   0 0 0 0  0   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 1 1 2 3   0 1 1 2  1   0 0 0 0 1 1 0 1 1 1 0 1 1 1 0
 0 2 2 3 1   0 2 2 1  1   0 0 0 1 0 1 1 0 1 1 1 0 0 1 1
 0 3 3 1 2   0 1 1 1  2   0 0 0 1 1 0 1 1 0 0 1 1 1 0 1
 1 0 1 1 1   1 0 1 1  1   0 1 1 0 0 0 0 1 1 0 1 1 0 1 1
 1 1 0 3 2   1 1 0 1  2   0 1 1 0 1 1 0 0 0 1 1 0 1 0 1
 1 2 3 2 0   1 2 1 2  0   0 1 1 1 0 1 1 1 0 1 0 1 0 0 0
 1 3 2 0 3   1 1 2 0  1   0 1 1 1 1 0 1 0 1 0 0 0 1 1 0
 2 0 2 2 2   2 0 2 2  2   1 0 1 0 0 0 1 0 1 1 0 1 1 0 1
 2 1 3 0 1   2 1 1 0  1   1 0 1 0 1 1 1 1 0 0 0 0 0 1 1
 2 2 0 1 3   2 2 0 1  1   1 0 1 1 0 1 0 0 0 0 1 1 1 1 0
 2 3 1 3 0   2 1 1 1  0   1 0 1 1 1 0 0 1 1 1 1 0 0 0 0
 3 0 3 3 3   1 0 1 1  1   1 1 0 0 0 0 1 1 0 1 1 0 1 1 0
 3 1 2 1 0   1 1 2 1  0   1 1 0 0 1 1 1 0 1 0 1 1 0 0 0
 3 2 1 0 2   1 2 1 0  2   1 1 0 1 0 1 0 1 1 0 0 0 1 0 1
 3 3 0 2 1   1 1 0 2  1   1 1 0 1 1 0 0 0 0 1 0 1 0 1 1

1-  0 0 0  2-  0 0 0  3-  0 0 0  4- 1 1 1  5- 1 1 1
*-  1 2 3  *-  4 5 6  *-  7 8 9  *- 0 1 2  *- 3 4 5
```

Figure 3.2: Basic plan example

As mentioned, the use of basic plans ensures orthogonality. However, they do not provide a solution for dominance among alternatives. Efficient designs ensure that dominant choice sets are removed from the experiment. We will discuss this design type next.

3.3.2. Efficient design

Efficient designs are experimental designs which are based on preliminary information. Priors are initially estimated parameter values obtained from literature, a pilot survey or any other method. By the use of an efficient design, dominance is avoided. Whenever dominance exists between two choice alternatives, this particular choice set provides no trade-off information on the criteria, and thus, the particular choice set should be removed. However, attribute level balance is violated when this is done, which introduces correlations and an overall less efficient design. Using an efficient design reduces the required number of choice sets, increases the reliability of parameters (i.e. low SEs), and defines the number of required respondents.

Efficient designs are constructed in the software program Ngene by entering several choice experiment characteristics. One of those is the number of required choice sets. In this regard, the reliable estimation of one parameter requires one *Degree of freedom*. Every choice set adds to the *Degrees of freedom*, depending on the number of alternatives in that choice set. In a choice set of n alternatives, one choice adds $n-1$ degrees of freedom. The minimum number of choice sets can then be calculated by dividing the number of parameters by the degrees of freedom and adding one to that total. It is important to note that efficient designs do not preserve attribute-level balance. Therefore, the total number of choice sets should be dividable by all the included attribute levels.

Efficient designs are distinguished into D-efficient and S-efficient designs. A choice between these two is based on the objective of the choice experiment.

D-efficient design

D-efficient designs achieve as much reliability as possible across all included parameters. In practice, this minimizes all SEs so that all parameters benefit. As a result, D-efficient designs are preferred when all parameters are regarded as equally important.

S-efficient design

S-efficient designs are favoured when the objective is to achieve as much reliability as possible on the most challenging parameter to make statistically significant. In practice, this means that one wins reliability on one parameter while losing reliability on all other parameters. Concretely, this means using an S-efficient design is appropriate when there is a special interest in one particular parameter.

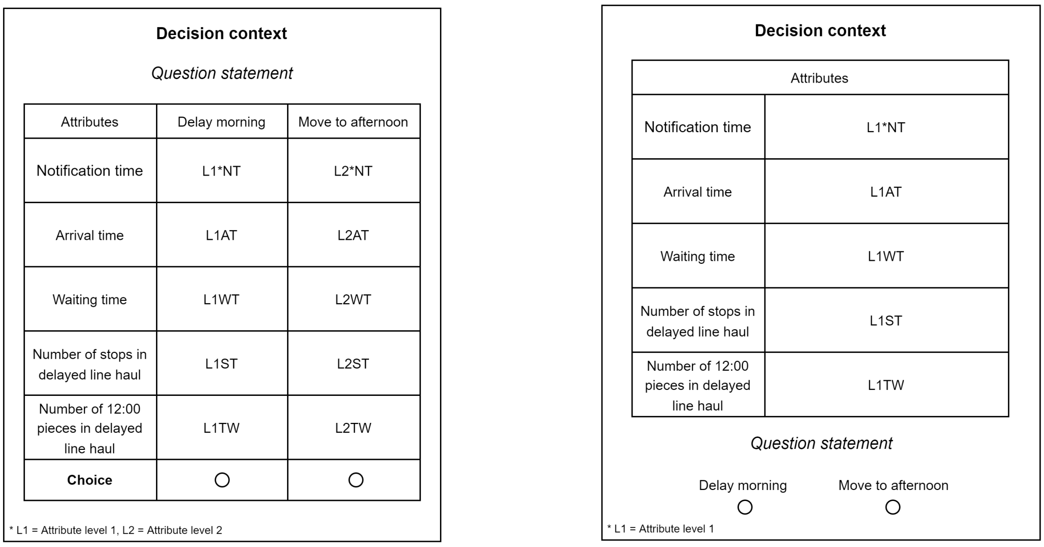
Whenever a specific experiment design is chosen, the composition of the actual choice set, which is the most fundamental part of the experiment, should be decided. Therefore, in the following sections, particular choice set set-ups are discussed.

3.4. Choice set composition

Every choice experiment design consists of a sequence of different choice sets. These choice sets could be indicated to the respondents in different ways. Also, the sequence and definition of choice sets might be constructed differently.

3.4.1. Choice sets

Choice sets are the most fundamental part of choice experiments, consisting of two or more choice alternatives, out of which the respondent chooses one. Each alternative inhabits multiple attributes in a general choice set, which are indicated through attribute levels. Figure 3.3a shows an example of such a choice set. In addition, a binary choice set is shown in figure Figure 3.3b. This particular experiment set-up indicates a set of criteria and corresponding criteria levels that are not directly connected to an alternative but compose a statement regarding the situation. To this extent, decision-makers are asked to indicate whether they favour a particular measure. For example, regarding delayed line hauls, supervisors can either delay the morning shift or move the delayed line haul packages to the afternoon.



Decision context

Question statement

Attributes	Delay morning	Move to afternoon
Notification time	L1*NT	L2*NT
Arrival time	L1AT	L2AT
Waiting time	L1WT	L2WT
Number of stops in delayed line haul	L1ST	L2ST
Number of 12:00 pieces in delayed line haul	L1TW	L2TW
Choice	○	○

* L1 = Attribute level 1, L2 = Attribute level 2

Decision context

Attributes	
Notification time	L1*NT
Arrival time	L1AT
Waiting time	L1WT
Number of stops in delayed line haul	L1ST
Number of 12:00 pieces in delayed line haul	L1TW

Question statement

Delay morning ○ Move to afternoon ○

* L1 = Attribute level 1

(a) General choice set

(b) Binary choice set

Figure 3.3: Choice sets

One of the challenges in deriving criteria trade-offs from choice experiments is to create sufficient variation throughout the choice experiment. Parameter estimates should be reliable (i.e. have small SE), and regarded as valid (i.e. resemble the true world). Choosing a proper experimental design for the choice experiment, and creating the choice tasks in a way that does not exhaust the respondents is essential to establish reliable and valid parameters (Molin, 2015).

3.4.2. Constructing choice sets

Construction of choice sets can be done in two ways: *Sequential construction* or *Simultaneous construction*. Sequential construction is executed by first, using basic plans or Ngene to construct choice alternatives. Second, decide how many alternatives to include per choice set. Third, all alternatives are, figuratively speaking, put in two vases. Fourth, draw one alternative from each vase. Redraw if either, alternatives are the same, or an earlier drawn choice set is replicated. And at last, check for correlations between attributes of options. Sequential construction solely works with generic attributes. Generic attributes have no labels (i.e. Option A & Option B), and have the same attributes and attribute levels across all alternatives.

Whenever it is preferred to use labeled alternatives, *Simultaneous construction* is used to construct choice sets. Different from sequential, one design row represents two choice alternatives simultaneously. Specifically this means that each row is one choice set. Labeled alternatives might have different attributes and attribute levels, which can be integrated in simultaneous construction.

By understanding the fundamental theory of BAIT, we discuss the modelling & survey design process in regard to the delayed line haul decision in the next chapter.

4 Modelling & Survey Design

In this Chapter, we discuss the modelling and survey design of the Behavioural Artificial Intelligence Technology (BAIT). At first, we dissect the delayed line haul decision to identify different aspects of the decision. After that, we scope the decision to decide which aspects will be involved in the choice experiment. After this scoping step, a list of decision criteria is composed. In this modelling step, we attach a criteria range, and importance rating to each criterion. We include the criteria list and criteria-specific characteristics in the final survey design in the third section. Also, a theoretical framework indicates context statements informing the respondent about the situation and demographic questions. At last, we elaborate on the specific design choice for the choice experiment and indicate an overview of the 32 choice sets. Fourth, as we collect personal data through the choice experiment, a section is posed to clarify how this is dealt with. Finally, the actual Multinomial Logit Model (MNL) model estimation step is discussed. In this last step, the identified criteria are combined in a utility function, and the model outcomes are discussed.

4.1. Scoping of the decision

Three conversations were held with a group of two Operational Supervisors to scope the decision. These Operational Supervisors are the end responsible decision-makers in case of a delayed line haul. During the first conversation we elaborated on the goal of this case study, and explained the method BAIT with an example case study. Next, we took an investigative approach by asking the Supervisors a set of structured interview questions. These questions were listed in a predefined sequence to preserve a natural conversation flow. Questions entailed (1) the degree of involvement of decision-makers, (2) the potential outcomes of the decision, (3) the time spent on each decision, (4) the information availability to the decision, (5) measurability of the decision outcome and contextual factors that might influence the decision outcome. The full conversation was supported using a PowerPoint presentation. The second and third conversation were about inventorying decision characteristics through (1) decision criteria identification, (2) criteria range determination, (3) criteria importance, (4) knock-out levels and (5) interaction effects. In the following subsections, we explain the decision scope and characteristics.

4.1.1. Decision-makers

As mentioned by the Supervisors, the decision outcome is a product of a discussion between the Operational Supervisors and the Support Centre employees. During this decision, the Support Centre employees act merely as information providers (i.e. give location updates on the delayed line haul). In addition to the live location of the truck, whenever a line haul delivery is delayed, the Supervisors gather specific truckload information through two information platforms. The Supervisors noted that a delay is encountered on average once every week. However, during periods when service centres are understaffed (i.e. summer holidays), delivery delays are more frequent. Moreover, also in winter periods, when roads tend to be more difficult to access, an increase in traffic jams leads to more delays. In the end, Supervisors assess the final verdict.

4.1.2. Decision outcomes

In total, four different decision outcomes were identified during the conversations: (1) delay the morning shift, (2) move delayed line haul parcels to the afternoon shift, (3) hold up part of the morning shift, and (4) move delayed parcels to the next day. However, during the conversations, it became clear that partly delaying the morning shift outcome is difficult to incorporate as a choice alternative since it needs information on all other morning and midday tours. In addition, the Supervisors mentioned the alternative of moving the delayed line haul to the next day as the worst-case alternative, which is generally only chosen when the midday shift capacity is full. Therefore, we decided to exclude these two alternatives and limit the experiment to the delayed morning shift, and move to the afternoon shift alternatives. Making a decision in regard to one of these decision outcomes results in particular consequences on operational performance. These consequences are measurable through company established KIPs. According to the Supervisors, consequences are measured using three Key Performence Indicators (KPI).

- **AM** = Number of delivered pieces / Hours in distribution
- **SPORH** = Number of stops per route
- **TDX** = On-time deliveries of urgent 10:30 & 12:00 deliveries

It was noted during the first conversation that, by means of these KPI's, decisions are influenced and might be evaluated afterwards. What is meant by influenced, is that real-time KPI standings bear an effect on the decision outcome. For instance, a low TDX might incentivize Supervisors to move the delayed parcels to the afternoon to increase on-time deliveries by departing the rest of the morning shift, and thus increase the standing of the TDX KPI. Although this notion was raised by the Supervisors, in the second conversation where we presented a mock-up choice experiment containing criteria from the first conversation, current KPI standings were not actively overthought by the Supervisors. Therefore, we decided to not include the KPIs as a criteria. In this line of thought, we assume that Supervisors calculate the consequences of their choice through accumulated expertise.

4.1.3. Decision time spent

According to one of DHL's service centre managers, Supervisors spend a significant amount of time to decide on delayed line haul deliveries. According to the manager, this is valuable supervisor time, most preferably used in other processes. According to Supervisors, they spent 10 - 15 minutes between the notification of a delayed linehaul and the actual decision. These 10-15 minutes are used to gather relevant information about the linehaul and subsequent discussion between Supervisors to form a decision.

4.1.4. Information availability

In total, Supervisors consult two information platforms to derive their choice. The NMIV software indicates all planned line haul arrivals, accompanied by a Stated Time of Arrival (STA) and an Actual Time of Arrival (ATA) for each specific day. Of this information, Supervisors are alarmed when a line haul delay is at risk. Whenever this situation arises, secondly, Supervisors might inform the NCG software. This software consists of the load characteristics of each incoming line haul. Altogether, it was noted that the real-time truckload arrival forecast received from the Support centre is perceived as unreliable by Supervisors. The availability of a more accurate forecast would increase the ability to make relevant trade-offs with load characteristics criteria.

4.1.5. Decision context

In a decision context, events occur wherein Supervisors encounter malfunctioning distribution machinery (i.e. assembly lines, scanners). As a result, the distribution process is shut down, which implies an automatic delay in the morning shift. Because this is the case, malfunctioning distribution machinery is not included as a contextual factor. Additionally, the Supervisors indicated that service centres differ in size and therefore cope with varying parcel volumes. As a result, service centres might act differently to similar delayed line haul loads. Noting that the choice experiment will be forwarded to different service centres, and thus Supervisors that work with different parcel volumes, this context factor will be incorporated in the choice experiment.

In terms of context, also demographic factors might affect the decision outcome. Therefore, Supervisors will be asked for their years of work experience at DHL and their respective service centre location of deployment to assess whether these context factors significantly influence the decision outcome.

After dissecting and scoping the different decision aspects incorporated in the modelling stage, decision criteria will be identified in the following sections.

4.2. Decision criteria inventarisation

In the final survey design, first, it is necessary to identify which criteria are overthought during the decision and their attribute levels and importance scores. Second, potential criteria knock-out levels, interaction effects and constraints might be identified.

4.2.1. Criteria identification

Through the group of experts, a list of criteria is established. These are the so-called attributes which the Supervisors value with each decision. Since this list of criteria was established with only three Supervisors, particular interest was put in the wording of each criterion. This is essential since the experiment will, in the end, be executed by the entire Supervisor group. Thus the meaning of each criterion should be derivable just from its name. After criteria identification, each particular criterion gets defined by a criterion range. In total, we identified 8 decision criteria being the *Notification time*, the *Arrival time*, the *Waiting time*, the *Number of stops in the delayed line haul*, the *Number of twelve-hour pieces in the delayed line haul*, the *Number of non-timebound pieces in the delayed line haul*, the *Capacity of the afternoon shift* and the *Closeness of delayed stops to the midday routes*.

The *Notification time* indicates the duration between the notification that a line haul will be delayed and the moment of arrival of that line haul. The *Arrival time* indicates the arrival time of the last delayed line haul. As a delay indicator for other morning shift included parcels, the *Waiting time* indicates the time between finalization and the moment of potential departure with the inclusion of the delayed line haul packages. To address the contents of the delayed line haul, the *Number of stops in the delayed line haul* serves as an indicator of the number of stops which should be added in case of a transfer of delayed parcels to the afternoon shift. The *Number of twelve-hour pieces in the delayed line haul* indicates the urgency of on-time delivery of the delayed parcels. To address the needed storage room for the delayed packages and the volume addition to the afternoon shift, we added the *Number of non-timebound pieces in the delayed line haul*. Lastly, the *Capacity of the afternoon shift* indicates the excess parcel capacity of the afternoon shift, and the *Closeness of delayed stops to the midday routes* the fit of the delayed parcels to the afternoon shift routes.

4.2.2. Criteria range determination

Since criteria attributes must vary per choice alternative or choice situation, it is essential to establish attribute ranges. Again, this process is executed in consultation with two Supervisors. Establishing at least three attribute levels is preferred since this allows for testing for non-linearity. Non-linearity indicates that a marginal increase of the attribute is not uniformly distributed, in terms of utility contribution, over the different attribute levels. Thus, an increase from attribute level 1 to attribute level 2 bears a lower utility contribution than an increase from attribute level 2 to attribute level 3. Regarding our case, we derived the minimum and maximum values of each criterion together with the Supervisors, after which we divided the total range by two to establish a third attribute level.

Table 4.1 shows an overview of the attributes, type, attribute levels and corresponding ranges.

Table 4.1: Overview attributes, type, levels and range

Criteria	Type	Level 1	Level 2	Level 3	Range
Notification time	Numeric	30 minutes	105 minutes	180 minutes	150
Arrival time	Ordinal	09:30	09:45	10:00	n.a.
Waiting time	Numeric	10 minutes	20 minutes	30 minutes	20
Number of stops in delayed line haul	Numeric	20	140	160	140
Number of 12:00 pieces in delayed line haul	Numeric	0	20	40	40
Number of non-timebound pieces in delayed line haul	Numeric	50	275	500	450
Capacity of afternoon shift	Numeric	125%	100%	75%	50%
Closeness of delayed stops to midday routes	Ordinal	Close	All around	Far	n.a.

4.2.3. Criteria importance determination (priors)

After criteria identification and attribute range determination, since we are using an efficient design, priors must be defined. To do so, again, we consulted the Supervisor expertise to estimate priors. Priors consist of two factors: a positive or negative sign and a value that indicates the utility contribution. First, the Supervisors directly chose a positive or negative connotation per criterion. Next, one out of three importance scores - nice-to-have, important, or critical - was assigned to each criterion. These three levels correspond to prior values of 0.5, 1.0 and 1.5. Table 4.2 indicates each attribute's sign, importance score, utility distribution and corresponding prior.

Table 4.2: Overview attribute priors

Criteria	Expected sign	Expected utility distribution	Importance rating	Prior
Notification time	Positive	Concave	Important	n.a.
Arrival time	Negative	Concave	Important	n.a.
Waiting time	Negative	Linear	Critical	-1.5
Number of stops in delayed line haul	Positive	Linear	Critical	1.5
Number of 12:00 pieces in delayed line haul	Positive	Linear	Important	1.0
Number of non-timebound pieces in delayed line haul	Positive	Linear	Important	1.0
Capacity of afternoon shift	Negative	Linear	Important	-1.0
Closeness of delayed stops to midday routes	Positive	Linear	Nice to have	-0.5

The table above indicates no prior for the Notification and Arrival time attributes due to their expected non-linear utility distribution. As mentioned, there is an uneven distribution of expected utility contribution across the attribute ranges. To cope with this, it is decided to dummy code these attributes. Software program Ngene inhabits a fixed internal coding structure to include dummy variables. Table 4.3 & Table 4.4 indicate this structure. As can be seen, two priors instead of one are determined, and attached to newly established attribute levels. By doing so, varying utility contributions can be ensured per attribute level.

Table 4.3: Notification time dummy coding scheme

Notification time	**Attribute level**	**Prior 1 (-2.00)**	**Prior2 (-0.89)**	**Expected utility**
30 minutes	*0*	1	0	-2.00
105 minutes	*1*	0	1	-0.89
180 minutes	*2*	0	0	0

Notification time is the first dummy coded criterion. Concretely, it was observed with the Supervisors, that a decrease between 105 and 30 minutes notification time, contributes more to the choice to delay the morning shift than a similar decrease between 180 and 105 minutes. More specifically, the utility function of notification time follows a concave-like shape, where utility contribution declines with every marginal increase of the attribute.

Table 4.4: Arrival time dummy coding scheme

Arrival time	**Attribute level**	**Prior 1 (2.00)**	**Prior2 (1.78)**	**Expected utility**
09:30	*0*	1	0	2.00
09:45	*1*	0	1	1.78
10:00	*2*	0	0	0

In addition to the notification time, also the arrival time is non-linear. Therefore, we modelled the arrival time according to Table 4.4. The arrival time follows a concave utility distribution similar to the notification time. Namely, the difference between 09:30 and 09:45 weighs less for the choice to delay than the difference between 09:45 and 10:00.

4.2.4. Criteria knock-out levels

Following the definition of priors, if present, knock-out levels must be identified. A knock-out level corresponds to an attribute level that, by itself, defines the decision-maker's choice. Whenever this is the case, no criteria trade-off can be derived from that particular choice set, which makes it irrelevant to the experiment. In our case, the attributes and attribute levels were established so that no knock-out levels exist.

4.2.5. Criteria interaction effects

Although it is desired to have only independent criteria, this is often not entirely possible. Therefore, interaction effects might be added to the model to account for additional utility contributions. Interaction effects exist when a particular combination of criteria provides extra utility besides their independent main effects.

In our case, one interaction effect was included in the model. Being the effect between the *Number of stops in delayed line haul* and the *Closeness of delayed stops to midday routes.* As confirmed by the Supervisors, the reasoning behind this is that more stops in the delayed line haul increase the importance of closeness to midday routes. Because more stops must be added to these midday routes, significantly larger delivery times are at risk. Thus, closeness gets more important. Figure 4.1 shows a conceptual explanation of the interaction effect.

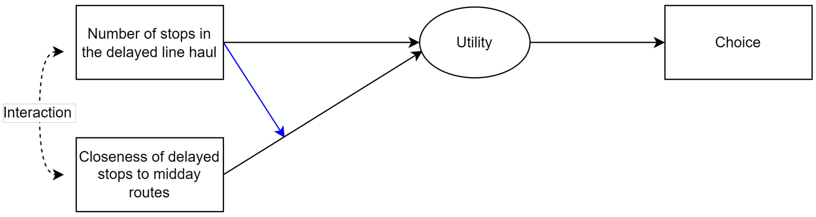

Number of stops in the delayed line haul
Closeness of delayed stops to midday routes
Interaction
Utility
Choice

Figure 4.1: Conceptual explanation interaction effect

4.2.6. Criteria constraints

Lastly, it might be the case that particular criteria levels can not be combined. In this situation, the design is made such that these levels will not be present within one choice set, as combining them would lead to an unrealistic scenario. During the conversations, the Supervisors identified one constraint: between the *Number of stops in delayed line haul* and the *Number of non-timebound pieces in delayed line haul.* Intuitively, the number of stops in the delayed line haul should not exceed the total number of delayed pieces. Whenever this is the case, multiple stops must be made for a singular piece that is faulty by nature. Therefore, a constraint is added to the choice experiment, noting that 50 non-timebound pieces in the delayed line haul can not be combined with 140 or 160 stops in the delayed line haul.

4.3. Final survey design

After scoping the decision and conducting criteria identification, a final survey design can be made. First, we pose a theoretical framework indicating all the survey components. Second, we elaborate on the context statement added to each choice scenario. Third, we discuss the demographic questions added at the end of the survey. Fourth, the number of required choice sets is calculated, and the particular choice of design is elaborated. Lastly, all choice sets of the survey are indicated in a table.

4.3.1. Choice experiment theoretical framework

With the information from the previous sections, a theoretical framework is visualized below, consisting of all the survey components. As can be observed, the survey consists of three components that influence the overall utility. First, the eight decision attributes are indicated. Second, the service centre's size indicates a particular context setting. Lastly, the utility might be affected by the demographic characteristics of the respondents. In this regard, it is chosen to include the service centre of deployment and the job experience at DHL. As seen in Figure 4.2, the utility leads to the respondent's choice.

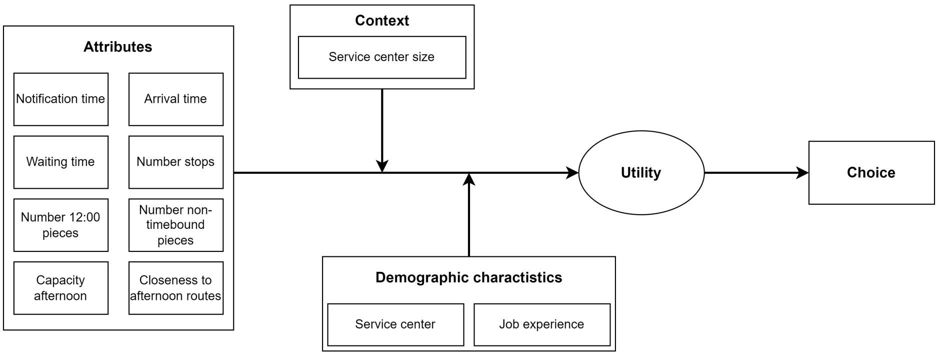


Figure 4.2: Theoretical framework survey design

4.3.2. Survey context statements

A statement is added to each choice situation to address the context in which delayed line haul decisions are made. As mentioned before, the service centre size of the Supervisor's deployment might affect the Supervisor's choice. For example, large service centres process large volumes of parcels and inhabit a large storage capacity.

Therefore, these service centres might experience less burden when delayed line haul deliveries are stored till the afternoon shift. Since the process of decision scoping and criteria characteristic determination was executed with two Supervisors from a medium-sized service centre, the attributes and attribute ranges were determined according to that particular service centre size. Therefore, to establish a common understanding of the attributes and attribute levels among the Supervisors, the following statement was added to each choice set:

Some of the criteria values below may differ per service centre. We are aware of that. Since we cannot design a separate survey for every service centre, it is chosen to take values that correspond to an average service centre size.

That is why we ask you to take the choices, as being a Supervisor from a medium-sized service center

4.3.3. Survey demographic questions

In terms of demographics, two factors are identified. Supervisors are distinguishable by service centre of deployment and years of working experience. These factors are regarded as factors that potentially affect utilities and, consequently, choices. A differing location of deployment might result in different Supervisor training, size of operations, and executive supervision. All two factors might influence the decision process and, thus, the valuation of criteria weights. Additionally, years of working experience might also affect criteria valuation. For example, more experienced Supervisors might be more conservative in their choices and thus have a natural tendency towards morning shift delay.

Table 4.5 indicates the two categories and subcategories in which Supervisors can indicate their characteristics. This is done through three indicative questions at the end of the choice experiment.

Table 4.5: Demographic factor questions

Service centre?	**Years of experience?**
	0 - 5
Den Hoorn	6 - 10
Amersfoort	11 - 15
Breda	16 - 20
	>20

4.3.4. Number of required choice sets

In previous sections, eight criteria were introduced, consisting of 3 levels each. Out of those eight criteria, two are dummy-coded, thus requiring two parameters. Therefore, ten parameters are to be included as criteria indicators. Besides that, also a parameter is estimated for the interaction effect described in subsection 4.2.5, and an Alternative Specific Constant (ASC), which we discuss in chapter 5 in more detail. Because only two alternatives are offered to the respondents, each choice adds 1 degree of freedom according to the calculation described in subsection 3.3.2. As a result, 13 (12+1) choice sets are required in the experiment. However, due to a low number of choice respondents, and two non-linear criteria, it is chosen to increase the number of choice sets to 30. Including 30 respondents aligns with Councyl's standards for choice experiments for small expert groups (10 respondents).

4.3.5. Efficient design construction

With Ngene software, a choice experiment design can be generated. Because the goal is to establish criteria trade-off valuation for all delayed line haul decision criteria, a D-efficient design set-up is chosen. As discussed in subsection 3.3.2, a D-efficient design ensures the minimization of every Standard Error (SE) of all parameters throughout the choice experiment iterations. No special interest is put on one criterion, as would be the case in S-efficient designs. As a result, maximum reliability across all parameters is ensured to achieve the most trade-off valuation. The MNL estimation method is chosen to predict a given set's utilities. As discussed in section 3.2, MNL estimation inhabits a normal-shaped distribution in the error term of the utility function. Because we have no assumptions about the unobserved preferences of the population, this estimation method is regarded most reliable. In section A.4, the full code can be found, which was used in Ngene to generate the choice scenarios.

Moreover, a constant β_0 is added to the utility function. Because a binary choice experiment is modelled, we want the utilities of both most extreme choice situations to equally distribute around zero. If this is not the case, the choice sets within the choice experiment will have a tendency towards high utility (i.e. delay morning shift), or towards a lower utility (i.e. move to afternoon), causing an over representation of choices for a particular alternative. Consequently, less trade-off information can be gathered on the attribute levels attached to the lesser chosen alternative. Most preferably, we want the probability distribution of the choice sets to be around a 30-70% or 40-60% distribution (Molin, 2015). This way, enough trade-off valuation will be captured regarding all attributes. Table 4.6 shows the calculation of the maximum and minimum utilities and indicates the value of the β_0 constant.

Table 4.6: b0 constant calculation

	beta_a. dummy	*beta_b. dummy*	*beta_c*	*beta_d*	*beta_e*	*beta_f*	*beta_g*	*beta_h*	*beta_int_dh*	
Prior	-2	2	-1,5	1,5	1	1	-1	0,5	0,5	
(Prior 2)	-0,89	1,78								
(Prior 3)	0	0								
Attribute levels (max)	2	0	0	2	2	2	0	2	2	**Total**
Utility	0	2	0	3	2	2	0	1	2	12
Attribute levels (min)	0	2	2	0	0	0	2	0	0	**Total**
Utility	-2	0	-3	0	0	0	-2	0	0	-7
Difference										-5
b0										-2,5

As can be seen by calculating the minimum and maximum utilities, these are distributed unevenly in favour of the maximum utility alternative. To prevent this, we divide the difference in half to get a b0 constant of -2.5. Subtracting this utility to both functions achieves a +9.5 or -9.5 utility in maximum and minimum cases. Ngene inhabits a function to gain insight into the probabilities of an alternative to be chosen in each choice set. Table 4.7 shows an overview of these probabilities.

Table 4.7: Choice set probabilities

Choice set	**Alternative 1**	**Alternative 2**	B_s
1	0.445221	0.554779	99%
2	0.131244	0.868756	46%
3	0.5	0.5	100%
4	0.403717	0.596283	96%
5	0.151871	0.848129	52%
6	0.647941	0.352059	91%
7	0.403717	0.596283	96%
8	0.527472	0.472528	100%
9	0.5	0.5	100%
10	0.352059	0.647941	91%
11	0.596283	0.403717	96%
12	0.075858	0.924142	28%
13	0.5	0.5	100%
14	0.075858	0.924142	28%
15	0.268941	0.731059	79%
16	0.622459	0.377541	94%
17	0.622459	0.377541	94%
18	0.622459	0.377541	94%
19	0.622459	0.377541	94%
20	0.569546	0.430454	98%
21	0.445221	0.554779	99%
22	0.268941	0.731059	79%
23	0.569546	0.430454	98%
24	0.377541	0.622459	94%
25	0.268941	0.731059	79%
26	0.596283	0.403717	96%
27	0.68568	0.31432	86%
28	0.352059	0.647941	91%
29	0.527472	0.472528	100%
30	0.622459	0.377541	94%

According to the Ngene user manual, utility balance can be expressed in a percentage, being the utility balance metric B_s. This metric approaches 100% whenever the utility of a choice set is equally distributed over the alternatives. Thus each alternative has an equally big chance of being chosen. Equation 4.1 indicates how B_s can be calculated. J is the total number of alternatives, and j is a set consisting of all alternatives. $P_j s$ is the probability that an alternative of a choice set is chosen.

$$B_s = \prod_{j=1}^{J} \left(\frac{P_{js}}{1/J} \right) * 100 \tag{4.1}$$

In our case, with two alternatives per choice set (J=2), if both alternatives would have the same probability (0.5), then B_s would be 100%. To gather the utility balance of an entire choice experiment, all of the choice set's B_s metrics must be averaged over all choice scenarios:

$$B = \frac{1}{S} \sum_{s=1}^{S} B_s \tag{4.2}$$

By using the formula posed above, with the B_s values of Table 4.7, a choice experiment B value of 86% is achieved. According to the Ngene user manual, an optimal utility balance lies somewhere in the interval of 70-90%.

4.3.6. Survey choice sets

In total, 32 choice sets are shown to the Supervisors, consisting of the 30 choice sets generated by Ngene plus two additional ones. These additional choice sets are included for calibration purposes. These calibration scenarios test the previously determined attribute signs and corresponding attribute levels on consistency with the Supervisors' choices. By posing an extremely positive utility choice (i.e. delay morning shift) and interpreting the choice of each Supervisor, we can observe whether the attributes are calibrated correctly. Similarly, the second choice set poses an extremely negative utility choice (i.e. move to the afternoon shift), which the Supervisors should mark as such. Table 4.8 shows the final survey design consisting of 32 choice sets accompanied by corresponding attribute levels. The code which was used in Ngene to generate the 30 trade-off revealing scenarios in a most efficient way, can be found in section A.4.

Table 4.8: Ngene choice experiment output

Choice set	**Notification (NO)**	**Arrival (AT)**	**Waiting (WT)**	**Stops (NS)**	**12:00 (TH)**	**Non-timebound (NT)**	**Capacity (CA)**	**Closeness (CL)**
-2	180	09:30:00	10	260	40	500	125%	Far
-1	30	10:00:00	30	20	0	50	75%	Close
1	180	09:45:00	20	260	0	275	75%	Close
2	105	10:00:00	30	140	20	500	100%	All around
3	30	09:30:00	10	140	0	275	100%	All around
4	105	09:30:00	10	20	40	50	75%	Far
5	30	09:45:00	20	140	0	275	100%	All around
6	105	10:00:00	30	260	20	500	75%	Far
7	105	09:30:00	10	20	40	50	100%	Close
8	105	10:00:00	20	260	0	275	75%	Far
9	180	09:30:00	20	20	0	500	125%	Close
10	105	09:45:00	10	20	0	500	100%	Close
11	105	09:45:00	20	140	40	500	75%	Close
12	30	09:30:00	20	20	20	275	100%	All around
13	30	09:30:00	30	140	20	275	125%	Far
14	180	10:00:00	20	20	20	275	100%	All around
15	180	09:30:00	30	140	0	275	100%	All around
16	30	10:00:00	20	140	40	500	125%	All around
17	180	09:30:00	30	260	40	275	75%	Close
18	30	09:30:00	30	260	20	500	125%	Close
19	180	10:00:00	10	20	40	275	125%	Close
20	30	09:45:00	30	260	0	275	100%	Far
21	180	09:45:00	20	20	40	50	125%	Close
22	30	10:00:00	30	260	40	500	75%	All around
23	180	09:45:00	30	20	40	500	100%	Far
24	180	10:00:00	10	20	20	50	125%	Far
25	30	10:00:00	10	140	20	500	75%	All around
26	105	09:45:00	10	20	20	50	125%	Far
27	30	09:45:00	10	140	20	500	75%	All around
28	105	09:45:00	30	20	40	275	125%	Far
29	105	10:00:00	20	260	0	500	125%	Close
30	180	09:30:00	10	20	0	500	75%	Far

Next, these choice experiments were loaded into Councyl's choice experiment software to establish the experiment. In addition to the choice scenarios, an informed consent statement was added to the start and the context statement of subsection 4.3.2 to each scenario. Lastly, the demographic questions of subsection 4.3.3 were established to finalize the experiment. For a visual representation of the components of the experiment, we refer the reader to section A.5.

4.4. Data collection

Data collection is performed through Councyl's decision support software. A question layout is composed through this software, allowing the respondent to complete the experiment online by means of a laptop. In section A.5, several snapshots are indicated to show the actual layout of the experiment. Consisting of an informed consent statement, 32 hypothetical choice scenarios and three demographic questions, the experiment is threefold. Before distributing the experiment to the respondents, the Human Research Ethics Committee (HREC) of the TU Delft was informed of the specific characteristics of the collected data. Using an Informed Consent template, Data Management Plan and HREC Checklist, approval was received to send out the survey on 16 August 2022.

The choice experiment scenarios of figure Table 4.8 were uploaded into the software to generate the choice experiment. The experiment was forwarded on 17 August 2022 to 10 respondents. Each respondent is employed as Supervisor Operations at DHL Express Netherlands and qualified as a decision-maker in case of delayed line hauls. Supervisors were selected in close collaboration with one of DHL's Operations Managers. Supervisors of at least three different service centres were invited to the experiment to decrease the risk of an undifferentiated group causing one-sided results. Cooperation with the choice experiment was completely voluntary. No (monetary) incentives were offered to the respondents to increase the survey completion.

4.5. MNL model estimation

According to the method described in chapter 3, a model was estimated using MNL estimation protocol. With the Apollo library in software program R, iteratively, combinations of parameters estimates (i.e. criteria weights), which make the data most likely, are estimated. The likeliness of these parameters to predict the choices of the dataset can be measured according to three model fit metrics, which we described in subsection 3.2.1.

4.5.1. Model estimation

As a result of the respondents' choices, parameter estimates are derived. These estimates indicate the relative importance of each decision criterion on the total utility of a particular choice scenario. Subsequently, this observed utility defines, according to the MNL principle, the choice of the respondent. In the initial MNL estimation, 12 parameters are estimated. Being, the 8 decision criteria: *Notification Time (NO), Arrival Time (AT), Waiting Time (WT), Number of Stops (NS), Number of Twelve-hour Pieces (TH), Number of Non-timebound Pieces (NT), Capacitiy Afternoon Shift (CA), Closeness to Midday Routes (CL)*; the interaction effect: $\beta_{interactionNSCL}$; and a constant: β_{alt1}, which describes the intrinsic utility for Alternative 1. Out of the eight decision criteria, the Supervisors mentioned the NO and the AT as non-linear. To test for non-linearity, these two criteria are modelled with two dummy parameters as described in subsection 4.2.3. Moreover, the interaction effect described in subsection 4.2.5 between the number of stops and the closeness to afternoon routes is also added to the utility function by a separate parameter. Equation 4.3 shows the utility function of this first model estimation.

$$\begin{aligned} U_{movetoafternoon} = \beta_{Alt1} + \beta_{NOdummy0} * NO_{dummy0} + \beta_{NOdummy1} * NO_{dummy1} + \beta_{ATdummy0} * AT_{dummy0} \\ + \beta_{ATdummy1} * AT_{dummy1} + \beta_{WT} * WT + \beta_{NS} * NS + \beta_{TH} * TH + \beta_{NT} * NT + \beta_{CA} * CA \\ + \beta_{CL} * CL + \beta_{interactionNSCL} * NS * CL \end{aligned} \tag{4.3}$$

In total, two alternatives were included in the experiment. One of these two can be regarded as a do-nothing alternative, or in other words: delay the morning shift. The other alternative, 'c', is fixed to zero, in accordance with the priors and constant b0 that were estimated in subsection 4.2.3 and subsection 4.3.5.

5 Data Analysis & Implications

In this Chapter, we evaluate the results of the model estimation by data analysis and implications. To do so, first, it is essential to pose a sample description. By means of this description, we can analyse whether different segments bearing different preferences might exist within the sample group. Second, we address the outcome of the calibration scenarios, as discussed in subsection 4.3.6. In line with the outcome of these calibration scenarios, third, the parameter estimates of the model are discussed. In total, three models are estimated, out of which one is chosen to analyse in depth. Of this best model, each parameter estimate, as well as the Alternative Specific Constant (ASC) is discussed by relative utility contribution, significance and linearity. Also, per criterion, a comparison is made with its expected (prior) value. In the fourth section, this analysis is highlighted by a comparison of relative importance. Lastly, the outcome of the modelling process is shown by means of a visualization of the decision-tool. Also, prior, sample and population-specific prediction models are highlighted and tested on similar choice scenarios.

5.1. Sample description

Compared to most choice modelling experiments, the sample size in this research is relatively small. Experts are assumed to be busier and therefore possess less time to spend on the choice experiment. Also, there are less total experts that we can reach out to. As a result, finding expert decision-makers to participate in the experiment, and collecting their preferences by doing so, is experienced as more difficult than in population research. Although this is the case, experts are simultaneously more experienced with the decision of interest and thus possess extensive expertise on the valuation of criteria.

The survey was distributed to 10 of DHL's Operational Supervisors, of which 9 completed the choice experiment. To identify different segments within the sample group, supervisors are asked about their service centre of deployment and years of work experience. Table 5.1 shows the presence of these characteristics in terms of percentages of the respondent group. In addition to these characteristics, the completion time is indicated. As can be seen, choice experiment completion time varied largely between supervisors, potentially indicating a differing degree of dedication among the respondent group.

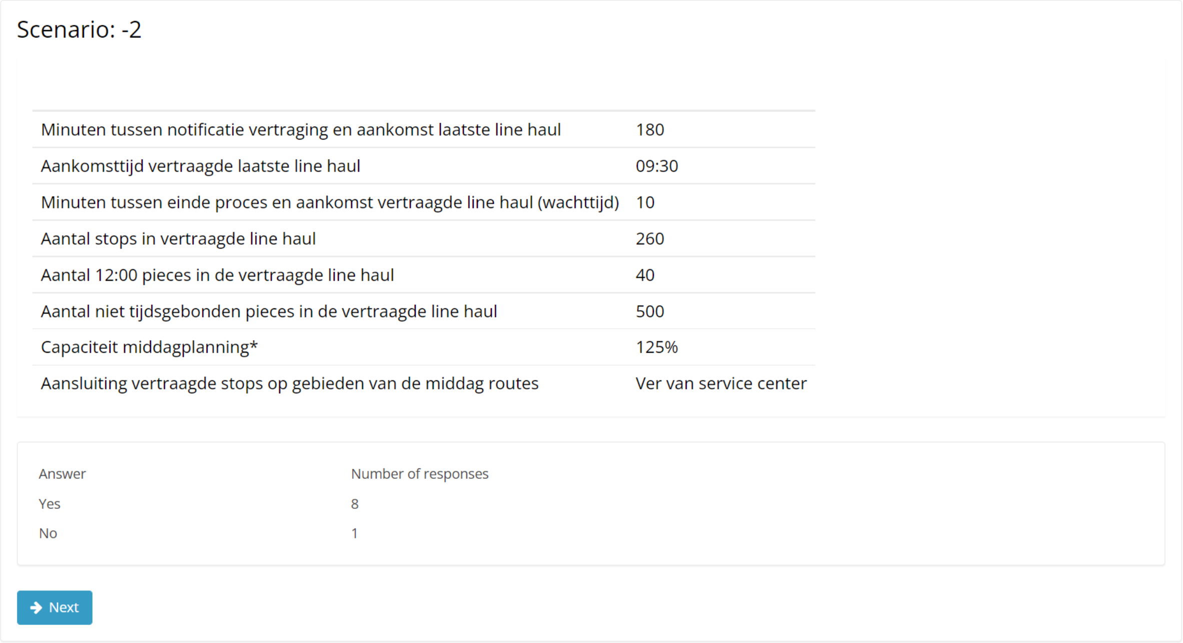
Table 5.1: Demographic characteristics & completion time respondents

Demographic factor	**Category**	**Respondents [%]**
Service center location	Den Hoorn	62.5%
	Breda	25.0%
	Amersfoort	12.5%
Years at DHL	<5 years	11.1%
	6 - 10 years	66.7%
	11 - 15 years	22.2%
	16 - 20 years	-
	>20 years	-
Completion time	<6 minutes	22.2%
	6 - 10 minutes	22.2%
	11 - 15 minutes	-
	16 - 20 minutes	33.3%
	21 - 25 minutes	-
	>25 minutes	22.2%

The abundance of supervisors is stationed in Den Hoorn, indicating that the distribution is quite steeply in favour of Den Hoorn. Regarding employment years at DHL, none of the respondents has been actively employed for more than 15 years. Most respondents are employed for 6 to 10 years, followed by 11 to 15 years and lastly, one respondent is active for less than five years. Again, the distribution of respondents is quite steeply in favour of one of the categories, indicating a skewed information distribution. Lastly, the completion time shows a different distribution.A similar number of respondents completed the survey in either less than 6 minutes, 6 to 10 minutes or more than 25 minutes. The largest share of respondents finished the experiment somewhere between 16 to 20 minutes. Although this last segment category shows the evenest respondent distribution, it might be interesting to investigate why particular Supervisors spend less than 6 minutes on the choice experiment while others more than 25. This notion will be touched upon in the recommendation section of this research.

5.2. Calibration scenarios

As mentioned before in subsection 4.3.6, the two most extreme choice scenarios were added to the choice experiment for calibration purposes. By doing so, criteria formulation and the expected influence of increasing attribute levels on the utility, and thus choices, can be confirmed. More specifically, when the answers to these two initial choice scenarios show great variation, it would be a sign that Supervisors do not understand the formulation of the criteria or that increases in attribute levels might work in the opposite direction than initially understood. The first scenario, denoted as '-2', is the utility maximization scenario. Within this choice scenario, each criterion attribute is valued so that maximum overall utility is the outcome. Thus, supervisors should choose to delay the morning shift, or in other words, do nothing.



Scenario: -2

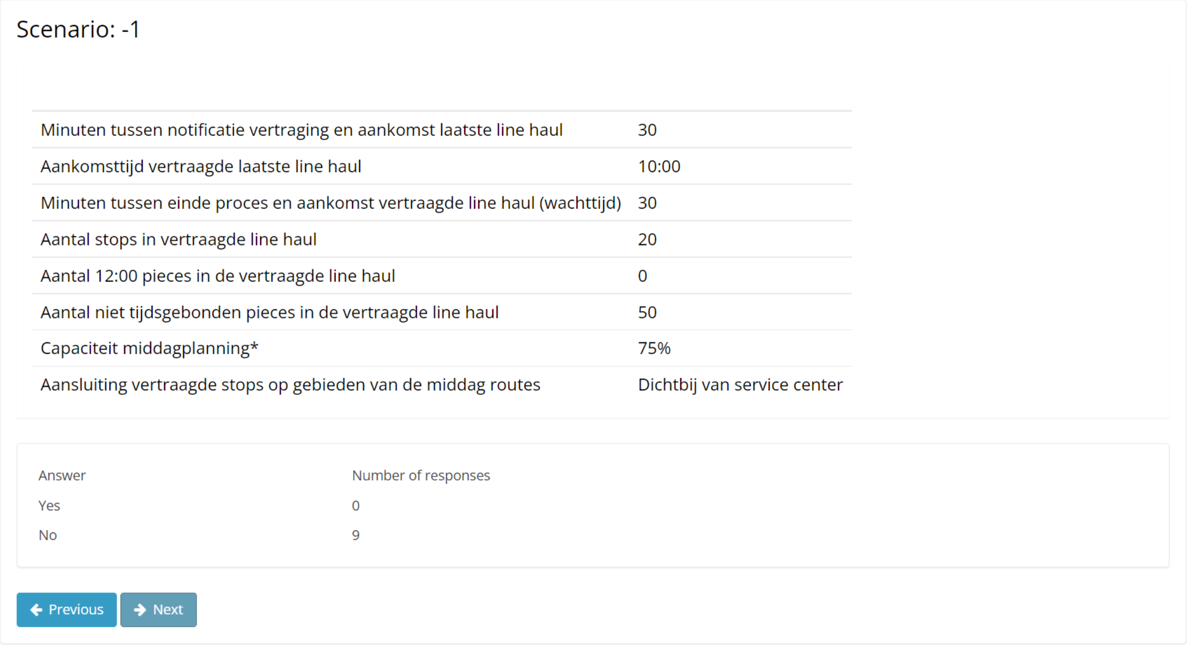
Minuten tussen notificatie vertraging en aankomst laatste line haul	180
Aankomsttijd vertraagde laatste line haul	09:30
Minuten tussen einde proces en aankomst vertraagde line haul (wachttijd)	10
Aantal stops in vertraagde line haul	260
Aantal 12:00 pieces in de vertraagde line haul	40
Aantal niet tijdsgebonden pieces in de vertraagde line haul	500
Capaciteit middagplanning*	125%
Aansluiting vertraagde stops op gebieden van de middag routes	Ver van service center

Answer	Number of responses
Yes	8
No	1

Figure 5.1: Scenario -2 utility maximization scenario

As can be seen in Figure 5.1, out of the nine respondents that participated in the survey, eight chose to delay the morning shift, and 1 Supervisor chose not to. This indicates that the larger share of the respondent group acted in line with our expectations.

Below in Figure 5.2, choice scenario ’-1’ indicates the utility minimization scenario. Each criterion attribute is valued so that minimum overall utility is generated. Therefore, it is expected that supervisors choose to act and move the delayed line haul pieces to the afternoon.



Scenario: -1

Minuten tussen notificatie vertraging en aankomst laatste line haul	30
Aankomsttijd vertraagde laatste line haul	10:00
Minuten tussen einde proces en aankomst vertraagde line haul (wachttijd)	30
Aantal stops in vertraagde line haul	20
Aantal 12:00 pieces in de vertraagde line haul	0
Aantal niet tijdsgebonden pieces in de vertraagde line haul	50
Capaciteit middagplanning*	75%
Aansluiting vertraagde stops op gebieden van de middag routes	Dichtbij van service center

Answer	Number of responses
Yes	0
No	9

Figure 5.2: Scenario -1 utility minimization scenario

As indicated in the picture above, in the utility minimization scenario, all Supervisors choose to move the delayed line haul packages to the afternoon. Again, this confirms our expectations of the overall utility when modifying attribute levels and criteria signs. Overall we can conclude that, although one respondent did not choose in line with expectations, the larger share of the sample did. Thus, the formulation and connotation of the criteria looks to be understood.

5.3. General MNL Parameter estimates

In total, three models were estimated. At first we estimate the model as described in subsection 4.5.1, including two non-linear criteria and one interaction effect. The parameter estimate of the non-linear notification time was found to be non-significant, as well as the interaction effect between the number of stops and the closeness to afternoon routes. Therefore, we decided to estimate a second model, in which only the arrival time is non-linear (this parameter was significant), and the interaction effect withdrawn from the estimation. Lastly, we estimate a third model, in which all non-linear criteria and the interaction effect is excluded. By estimating this third model, the impact of adding non-linear criteria and interaction on the model fit could be assessed. Table 5.2 indicates the different model fit metrics for each model estimation set-up.

Table 5.2: Different prediction model comparisons

Model fit metrics	**Model 1:** **a&b non-linear,** **interaction**	**Model 2:** **b non-linear,** **no interaction**	**Model 3:** **linear,** **no interaction**
LL	-167.29	-167.82	-168.38
$\boldsymbol{\rho^2}$	0.1620	0.1593	0.1565
MAD	7,16%	7.56%	7.55%

According to all three metrics, Model 1 best predicts respondent choices. According to the ρ^2, which compares the model fit of the estimated parameters to parameters of 0, model 1's estimated parameters are best in predicting choices according to the choices in the dataset. This is also made visible by the Mean Absolute Deviation (MAD) scores. As can be seen, the MAD of Model 1 is lowest at 7,16%. Since the choice experiment is performed with a relatively low number of respondents (9 in total), the MAD score, which indicates the difference between the actual percentage of delay choices and the prediction percentage of delay choices, is ambiguous. Because there are only nine respondents, the probability buckets increase by 11.1% (1/9) step at a time. The prediction model gives a probability range of 1 to 100%. Therefore, the difference between the actual probability and the prediction of the model may be larger than would be the case if more respondents had participated in the experiment. Thus a more accurate probability would be determined. Although this is the case, the MAD still gives a good representation of the model's fit.

Taking into account the model fit metrics, it is decided to advance with Model 1. Table 5.5 indicates the parameter estimates of the model, accompanied by their relative importance, the significance of the parameter, prior value and expected curvature of the utility contribution.

Table 5.3: Parameter estimates Model 1

Critera	Estimate	Relative importance	P-value	Significant? p-value <0.05	Prior	Expected curvature
β_{Alt1}	-1.11747	N.A.	0.027513	yes	N.A.	Uniform
$\beta_{N)dummy0}$	-0.08556	1.64%	0.807713	no	-2.0	Concave
$\beta_{NOdummy1}$	0.02705		0.939348	no	-0.88889	
$\beta_{ATdummy0}$	1.21660	17.74%	6.7911e-04	yes	2.0	Concave
$\beta_{ATdummy1}$	0.87721		0.008789	yes	1.77779	
β_{WT}	-0.50049	14.60%	0.010473	yes	-1.5	Linear
β_{NS}	0.69076	20.14%	0.009993	yes	1.5	Linear
β_{TH}	0.82241	23.98%	5.265e-06	yes	1.0	Linear
β_{NT}	0.07766	2.26%	0.707553	no	1.0	Linear
β_{CA}	-0.04132	1.20%	0.821249	no	-1.0	Linear
β_{CL}	0.25938	7.56%	0.239227	no	0.5	Linear
$\beta_{interactionNSCL}$	-0.18626	10.86%	0.305734	no	0.5	Linear

In the next sections we highlight each criterion by touching on the following criteria characteristics:

- **Comparison to priors**: at the start of the choice experiment design, importance scores were attached to each decision criterion in correspondence with the Supervisors. These importance scores were subsequently translated to prior values and used to ensure utility balance between the alternatives of each choice set. Moreover, expected signs were added to each criterion. The parameter estimates can be compared to the predefined priors and signs to analyse whether initial judgement corresponds.

- **Relative importance**: since all attribute levels are of the same magnitude (i.e. 0, 1 or 2), parameter estimations allow us to consider the importance of a single attribute regarding the other attributes. Concretely, a higher parameter estimate indicates that the expert group attaches larger importance to that particular attribute than a lower parameter estimate. In other words, the relative utility contribution of an attribute with a high parameter estimation is higher. It thus explains a more significant portion of the decision than an attribute with low parameter estimation.

- **Significance of parameters**: parameter significance is important in terms of scaling estimations to the population. An important metric that is derived from the SE is the t-value. This value indicates whether the attribute affects the choices in the population. Parameters with a t-value higher than 1.96 are considered statistically significant, thus, affect the choices of the population. Derived from the t-value is the p-value, which should be beneath 0.05 for the parameter to be significant. Concretely, this implies that whenever one would draw a new sample from the population, and observe choices a large number of times to derive new parameter estimates, then 95% of these new intervals contain the true parameter estimate of the population. On the contrary, when an attribute corresponds to a t-value lower than 1.96, the parameter estimate only declares something about the preferences in the sample.

- **Parameter linearity**: parameter estimates might result from the respondent choices as being non-linear. This implies that the difference between attribute levels (i.e. 0, 1 or 2) is different regarding utility contribution. By incorporating three attribute levels, as is done in our choice experiment, each parameter estimate can be tested for non-linearity.

5.3.1. Alternative Specific Constant

To address the unobserved utility, which is associated with factors other than the observed attributes, an ASC was estimated for the alternative to delay the morning shift. This ASC represents the overall utility to delay the morning shift when all other attributes are valued at 0. Therefore, since the ASC in Model 1 is -1.11747, one might argue that the unobserved preference of the sample group is to move delayed pieces to the afternoon. However, this is a one-sided analysis. Because of particular attribute modelling (i.e. linear or non-linear), the total utility contribution of different attribute levels within the same attribute might vary between models.

Based on these absolute differences, the height and sign of the ASC are influenced. To illustrate, we calculate the utilities for all three models to a specific choice scenario with and without the ASC. The results are indicated in Table 5.4.

Table 5.4: Different model utility contribution based on Scenario -1 with and without ASC

	ASC	**Utility contribution (without ASC)**	**Utility contribution (with ASC)**
Model 1 [2 non-linear coded variables]	-1.11747	-1.169	-2.287
Model 2 [1 linear coded variable]	-0.41487	-1.918	-2.333
Model 3 [only linear coded variables]	0.27060	-2.414	-2.143

As seen in the table above, the overall utility contribution of the different models in scenario -1 without incorporating ASCs shows a large variety. This is because of the absolute utility differences between attribute levels within the same attributes described above. As can be seen, when we calculate the actual utility contribution by adding the ASC, the differences in utility contribution become significantly smaller. This indicates that each model's predictions will not vary largely, resulting from a calibration measure, the ASC. This implies that the we must not see the ASC as an independently explaining factor of the unobserved utility of the sample group.

5.3.2. Notification time (NO)

The Notification Time (NO) estimate is relatively small compared to its prior value. This is noteworthy since the Supervisors indicated the NO as important. And thus overestimated in terms of utility contribution. Also, the second dummy parameter's sign contrasts with expectations. These estimates show that the NO follows an n-shape utility distribution as seen in Figure 5.3. The utility contribution between the first and second attribute levels is relatively large compared to the second and third attributes, where utility declines slightly. This differs from the concave utility distribution that was expected. Lastly, the NO parameters are not significant.

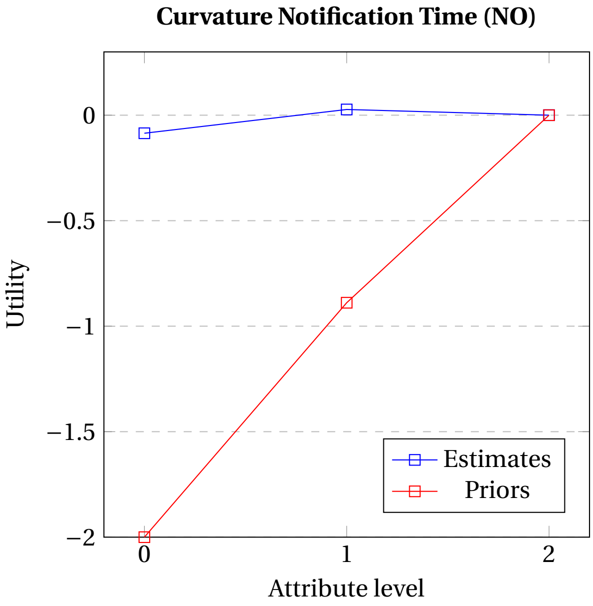


Figure 5.3: Utility contribution Notification Time (NO)

5.3.3. Arrival time (AT)

The utility contribution of the Arrival Time (AT) is high. This is seen in the relative importance of 17.74%. Also, a somewhat similar utility contribution can be seen in its prior, which were labelled as important, and expected to have a concave shaped utility. Figure 5.4 indicates the utility contribution of AT by showing its curvature characteristics. As can be seen, the utility decrease between the first and second attribute levels is more minor than between the second and third attribute levels, following a concave-like curvature. Contrary to the NO, the AT parameters are significant.

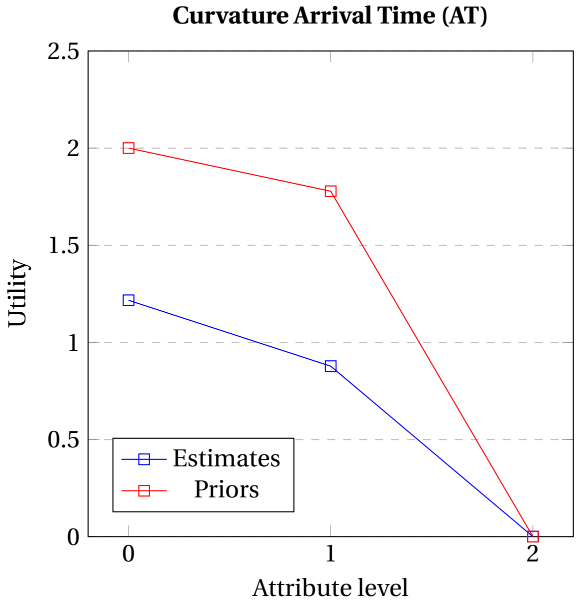


Figure 5.4: Utility contribution Arrival Time (AT)

5.3.4. Waiting time (WT)

The Waiting Time (WT) estimate utility contribution consists of -0.50049 and has a relative importance of 14.60%. Therefore, in contrast to its critically marked prior value, it contributes moderately to the overall utility. Although this is the case, the sign of the estimate corresponds with expectations and thus influences the utility negatively. Moreover, the WT estimate p-value is below 0.05, resulting in a significant parameter. Finally, the expected curvature of the WT is linear, as can be observed in Figure 5.5.

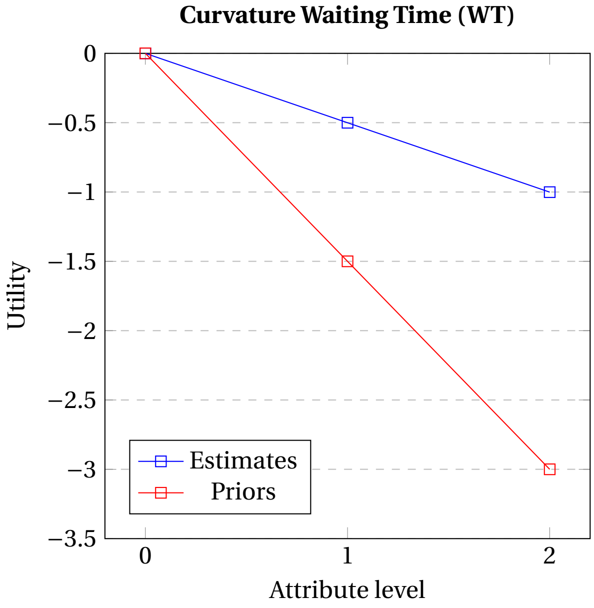


Figure 5.5: Utility contribution Waiting Time (WT)

5.3.5. Number of stops (NS)

Compared to the WT, the Number of Stops (NS) criterion is more important in utility contribution. However, the estimate has a significantly lower utility contribution than its expected utility contribution. Contradicting to the WT, increased stops are expected to increase the utility, which is confirmed by the estimated sign. Moreover, the p-value is below 0.05, indicating that the parameter estimate is significant. Lastly, the expected curvature of the parameter is linear. Figure 5.6 indicates the utility contribution of the NS.

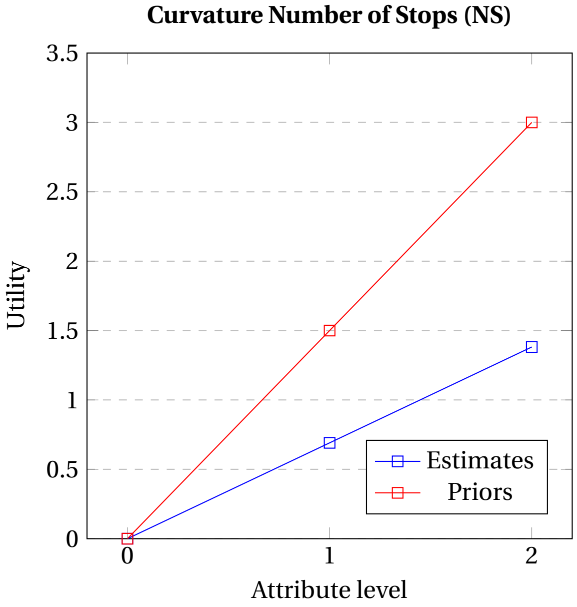


Figure 5.6: Utility contribution Number of stops (NS)

5.3.6. Number of twelve-hour pieces (TH)

Out of all the estimates, the Number of Twelve-hour Pieces (TH) contributes the largest share of the overall utility. 23.98% constitutes a significant portion of the overall utility, making the TH criterion critical for choice determination. Compared to the expected utility contribution, the estimate is very much in line with expectations. Also, the expected sign of the estimate, a positive utility contribution, is in line with expectations. Lastly, the criterion is significant and is expected to follow a linear curvature which can be seen below in Figure 5.7.

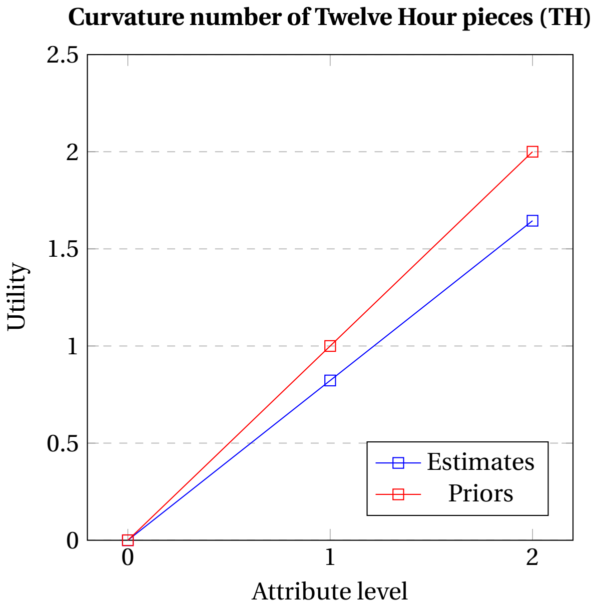


Figure 5.7: Utility contribution number of Twelve Hour pieces (TH)

5.3.7. Number of non-timebound pieces (NT)

The Number of Non-timebound Pieces (NT) is expected to bear a utility contribution in the same range as the TH criterion. However, as we can see from the estimate, this is not the case. 0.07766 signifies significantly less utility contribution than expected. Thus, the NT criterion has a very low impact on the respondent's choice. Moreover, the NT criterion is insignificant and expected to follow a linear utility contribution. Below, in Figure 5.8, the curvature of the number of non-timebound pieces is indicated.

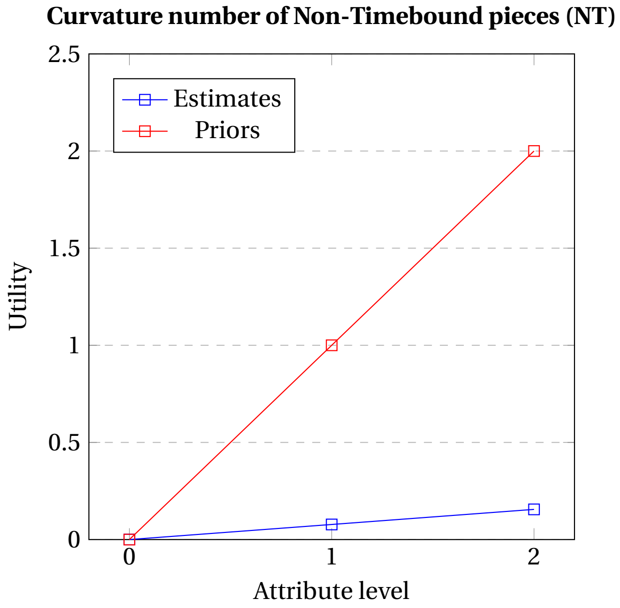


Figure 5.8: Utility contribution number of Non-Timebound pieces (NT)

5.3.8. Capacity afternoon (CA)

Similar to the NT criterion, the Capacitiy Afternoon Shift (CA) contributes very modestly to the overall utility. Bearing a contribution of only 0.04132, it has the most minor utility contribution of all criteria. Compared to the criterion's initial value, which the supervisors expect to be important, we see a significant deviation. Although this is the case, the estimation sign matches with the prior sign, indicating a utility decrease. Moreover, the CA criterion is not significant and is expected to follow a linear curvature, as can be seen below in Figure 5.9.

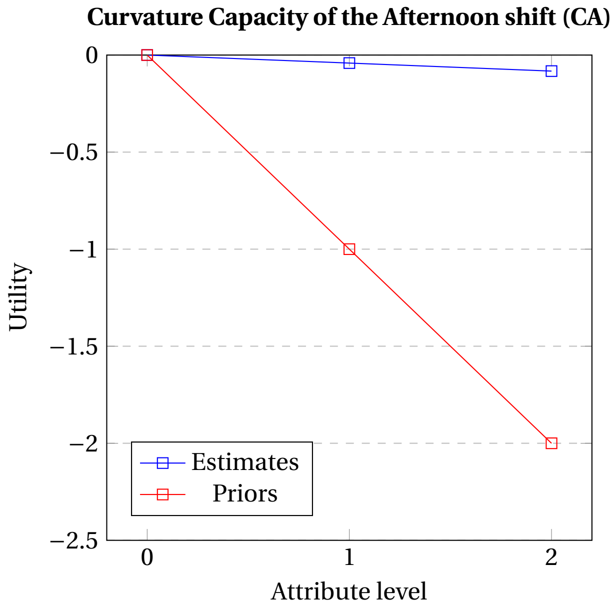


Figure 5.9: Utility contribution Capacity of the Afternoon shift (CA)

5.3.9. Closeness to afternoon routes (CL)

Closeness to Midday Routes (CL) does not contribute much to the utility of delayed line haul choices. Nevertheless, it does contribute somewhat more than the NT and CA criteria. We see an estimate of approximately half in size compared to its initial value. This suggests that the estimation does not derive that much from its prior. Also, the criterion CL is not significant and is expected to follow a linear curvature. In Figure 5.10, the utility contribution of CL is visualized and compared to the prior values.

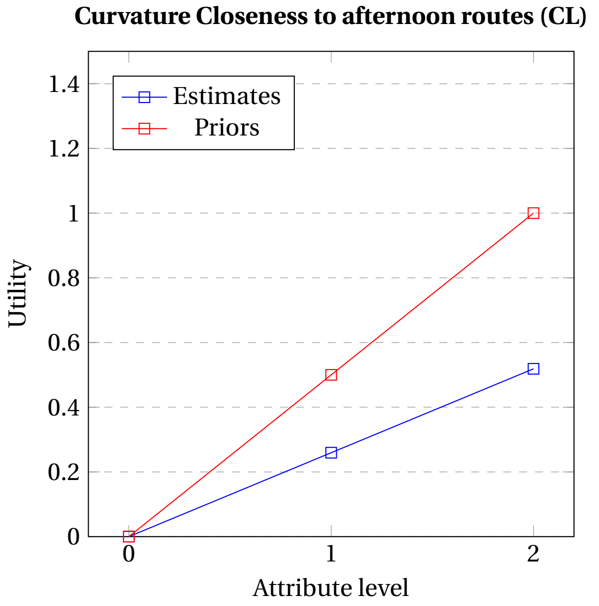


Figure 5.10: Utility contribution Closeness to afternoon routes (CL)

5.3.10. Interaction Number of Stops (NS) & Closeness to afternoon routes (CL)

The number of stops and closeness to afternoon routes interaction effect is expected to contribute positively to the overall utility according to its prior value. However, the parameter estimate indicates a utility decrease when NS and CL attribute levels increase. Here it is seen that the prior value sign differs from the outcome of Model 1, the parameter estimate. Intuitively, the overall utility should decrease whenever NS and CL attribute levels increase. Increases in these two criteria indicate that the number of stops increases while also these stops are closer to the afternoon shift. Whenever this is the case, shifting the delayed line haul packages to the afternoon becomes less troublesome because they fit well with afternoon routes. Thus overall utility must decrease. Therefore, the parameter estimate sign looks more likely than the predefined prior value. Regarding relative importance, the interaction effect only has a moderate effect. Also, it is insignificant, and the utility contribution is expected to follow a linear curvature. Figure 5.11 indicates the curvature and a comparison with the utility contribution of the prior value.

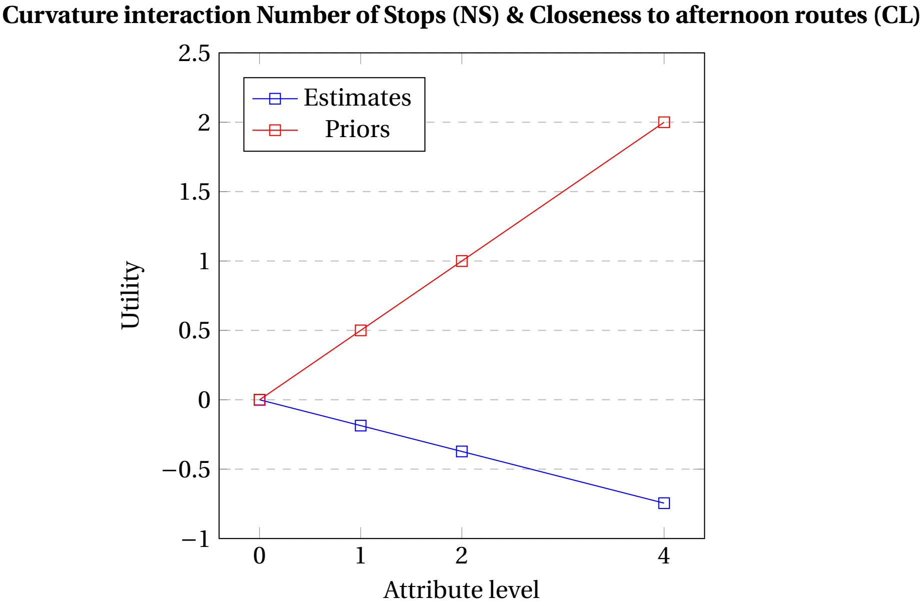


Figure 5.11: Utility contribution interaction Number of Stops (NS) & Closeness to afternoon routes (CL)

5.4. Relative importance comparison

In the previous section, each parameter estimate deviation from its predefined prior values is visualized. In each of these cases, it can be seen that the prior value is an overestimation of the parameter estimate. Therefore, one could argue that Supervisors tend to address more important criteria where overestimation is the largest. However, since the positive overestimations of the priors (i.e. optimistic priors) are also felt in a negative sense (i.e. negative priors), this is a one-sided analysis. As a result, one should not draw any estimate - prior value comparisons based on this one-sided analysis. If we still want to do so, however, it is essential not to regard the absolute differences between the estimate and prior but address their relative importance. Determination of these leaves us with the importance of one parameter, compared to the other ones, as was calculated for parameter estimates in Table 5.5. Comparing this relative importance of each prior with the relative importance of the parameter estimates of Model 1 allows us to compare their relative contribution to the utility. Figure 5.12 shows the difference per decision criteria between the parameter estimate, and it is prior, expressed in relative importance.

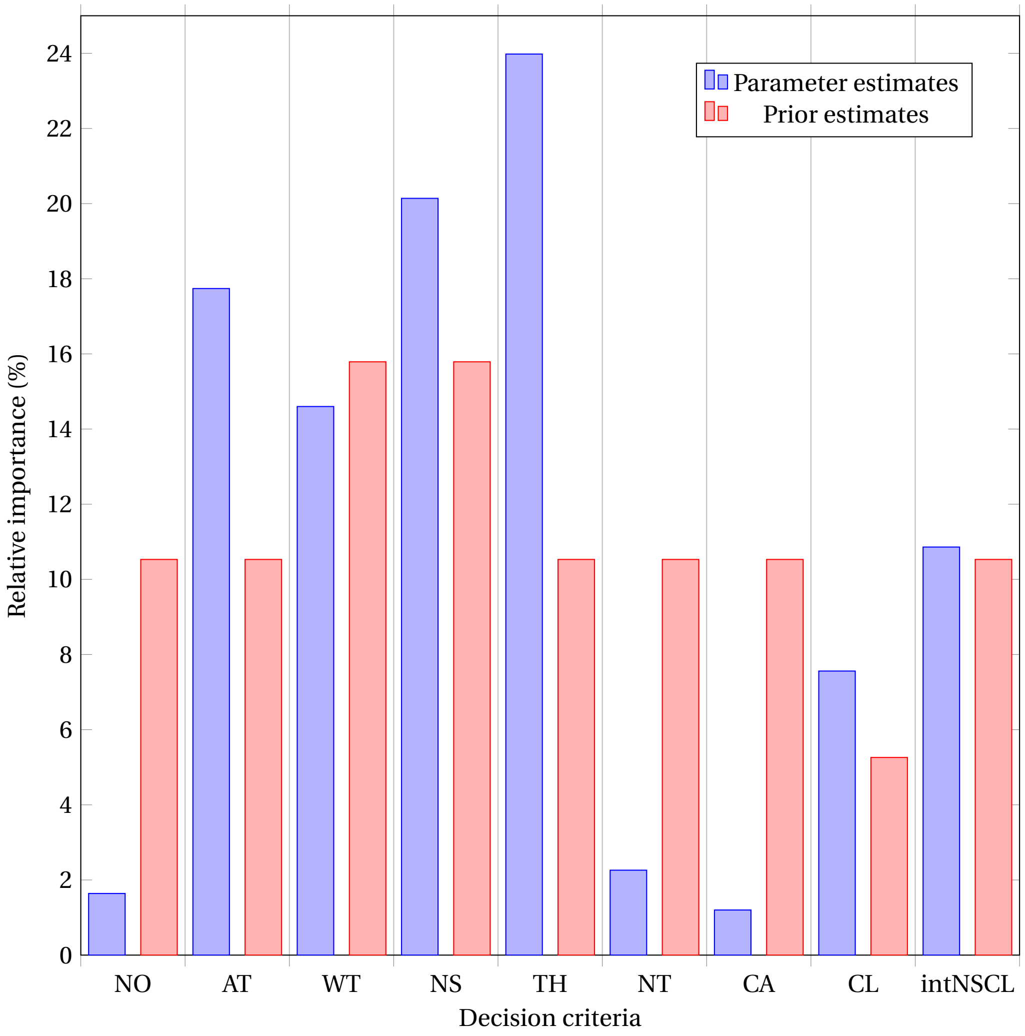


Figure 5.12: Relative importance parameter estimates and priors

The results above allow us to give answer to the **fourth research question** of this research:

How do directly judged criteria trade-offs differ from those obtained by modelling choices with Behavioural AI Technology?

The notification time, non-timebound pieces and capacity of the afternoon shift were overestimated in terms of importance. On the contrary, the arrival time, the number of stops and the twelve-hour pieces were underestimated in terms of importance. No significant deviations are identified concerning the other parameters: waiting time, closeness to the midday routes, the interaction effect between the number of stops and the closeness to the midday routes. Table 5.5 shows a more detailed comparison, indicating the precise relative difference per criterion.

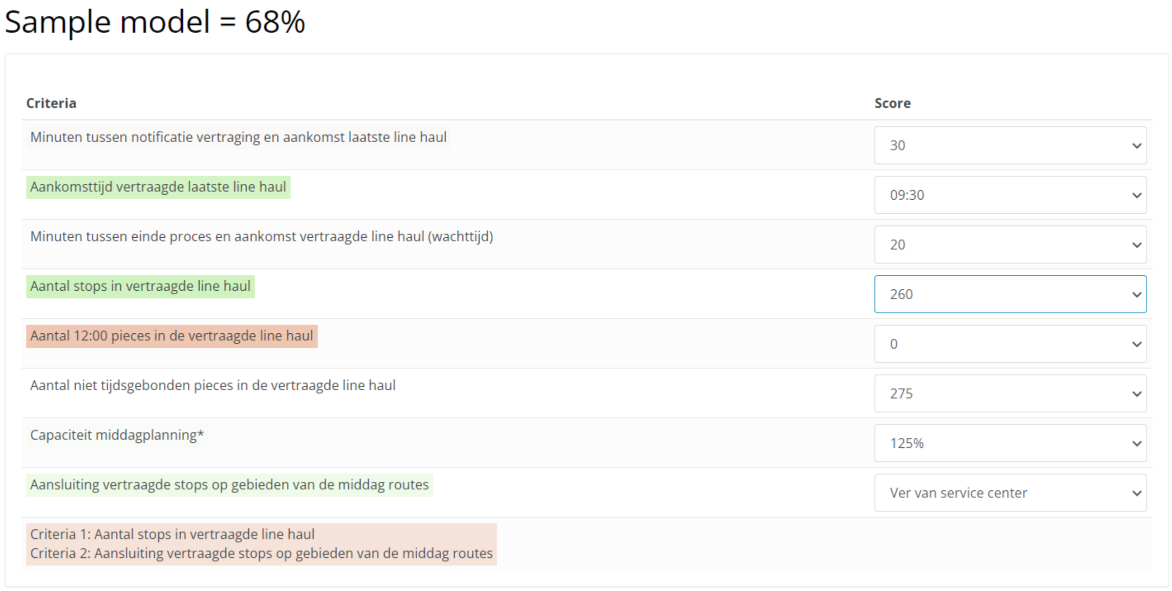
Table 5.5: Comparison relative importance parameter estimates and priors

Critera	Parameter estimate	Relative importance [Estimate]	Prior	Relative importance [Prior]	Difference
$\beta_{NOdummy0}$	-0.08556	1.64%	-2.0	10.53%	-8.89%
$\beta_{NOdummy1}$	0.02705		-0.88889		
$\beta_{ATdummy0}$	1.21660	17.74%	2.0	10.53%	+7.21%
$\beta_{ATdummy1}$	0.87721		1.77779		
β_{WT}	-0.50049	14.60%	-1.5	15.79%	-1.19%
β_{NS}	0.69076	20.14%	1.5	15.79%	+4.53%
β_{TH}	0.82241	23.98%	1.0	10.53%	+13.45%
β_{NT}	0.07766	2.26%	1.0	10.53%	-8.27%
β_{CA}	-0.04132	1.20%	-1.0	10.53%	-9.33%
β_{CL}	0.25938	7.56%	0.5	5.26%	+2.30%
$\beta_{interactionNSCL}$	-0.18626	10.86%	0.5	10.53%	+0.33%

Through modelling delayed line-haul decisions with Behavioural Artificial Intelligence Technology (BAIT), deviations are identified with initial criteria trade-off valuations. As seen above, the relative importance of parameter estimates (BAIT outcome) is compared to the prior estimates' importance stated by the experts before the experiment (direct judgement outcome). Direct judgement trade-offs reveal more evenly distributed relative importance scores than indirectly revealed importance. As can be seen, the relative importance of the parameter estimates is more skewed in favour of individual criteria like the number of stops and the number of twelve-hour pieces. This implies that the sample group valued these criteria heavier than others in the choice experiment. Also, it is noteworthy that the waiting time, closeness to the afternoon routes, and the importance of the interaction effect correspond to direct judgement. Looking at the results, we note a non-negligible difference between the initial and choice experiment-derived set of criteria trade-off scores. Next, we will generate three decision-making tools according to these parameter estimates.

5.5. Decision-making tools

In the end, the parameter estimates can be incorporated into a interactive decision-making tool. DHL may use this choice model freely to support decision-making in future cases. Using colour coding, the importance of each criterion to the choice is indicated. Green indicates a utility contribution, thus in favour of delaying the morning shift. Red, on the contrary, indicates a utility decrease, thus a push in the direction of moving the delayed pieces to the afternoon. The transparency of the colours indicates the degree of utility contribution of the criterion. Thus, high transparency means that the contribution of that criterion is low compared to others. To adjust the model to a new case, a filter can be opened for each criterion containing all attribute levels. Figure 5.13 shows a snapshot of the model's interface. In this case, the model predicts that 68% of the experts would delay the morning shift.

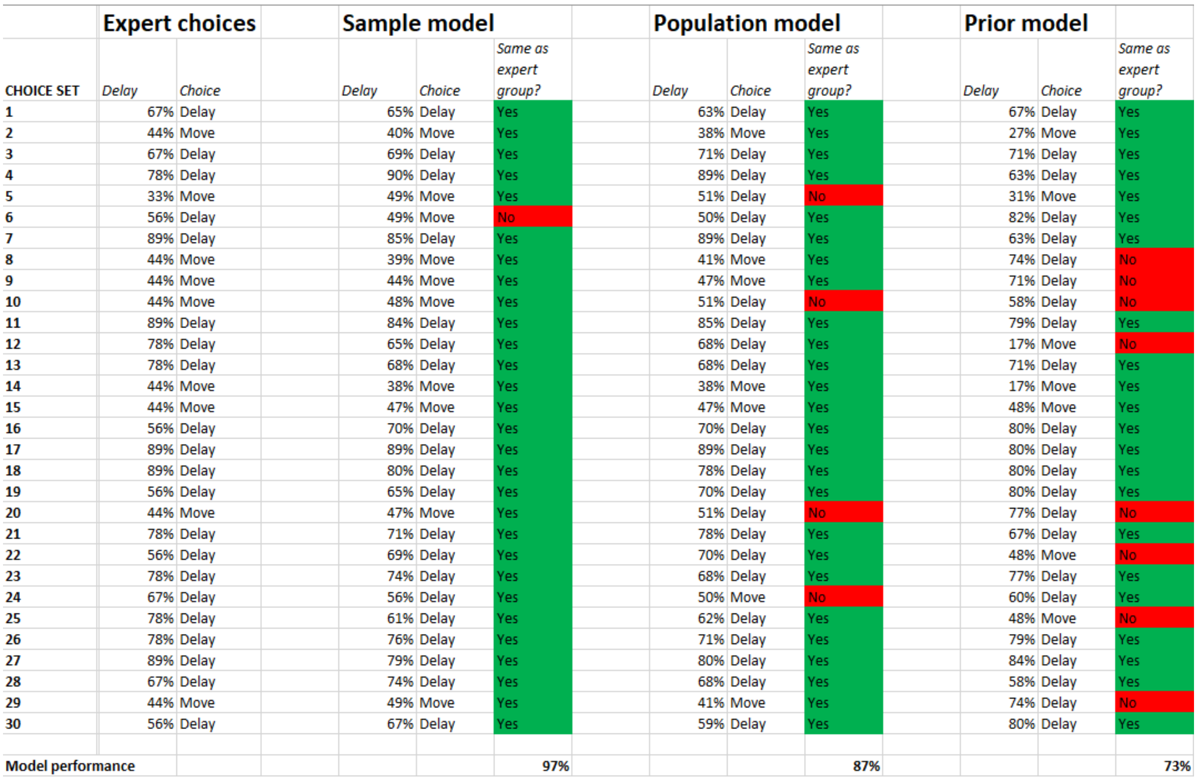


Sample model = 68%

Criteria	Score
Minuten tussen notificatie vertraging en aankomst laatste line haul	30
Aankomsttijd vertraagde laatste line haul	09:30
Minuten tussen einde proces en aankomst vertraagde line haul (wachttijd)	20
Aantal stops in vertraagde line haul	260
Aantal 12:00 pieces in de vertraagde line haul	0
Aantal niet tijdsgebonden pieces in de vertraagde line haul	275
Capaciteit middagplanning*	125%
Aansluiting vertraagde stops op gebieden van de middag routes	Ver van service center
Criteria 1: Aantal stops in vertraagde line haul Criteria 2: Aansluiting vertraagde stops op gebieden van de middag routes	

Figure 5.13: Sample model prediction example

In our case, we can incorporate three different sets of criteria weights into a prediction model. At first, we generate a sample model (i.e. Model 1 from section 5.3), consisting of parameter estimates based on indirect human judgement. Second, we generate a population prediction model. In this model, only parameter estimates that might be called significant (p-value <0.05) and thus scalable to the population of DHL supervisors are included. Third, we generate a prior prediction model to indicate in which parameter estimates are based on direct human judgement. To test performance, each of these models is compared to the actual choices of the expert group in Figure 5.14



	Expert choices		Sample model			Population model			Prior model		
CHOICE SET	*Delay*	*Choice*	*Delay*	*Choice*	*Same as expert group?*	*Delay*	*Choice*	*Same as expert group?*	*Delay*	*Choice*	*Same as expert group?*
1	67%	Delay	65%	Delay	Yes	63%	Delay	Yes	67%	Delay	Yes
2	44%	Move	40%	Move	Yes	38%	Move	Yes	27%	Move	Yes
3	67%	Delay	69%	Delay	Yes	71%	Delay	Yes	71%	Delay	Yes
4	78%	Delay	90%	Delay	Yes	89%	Delay	Yes	63%	Delay	Yes
5	33%	Move	49%	Move	Yes	51%	Delay	No	31%	Move	Yes
6	56%	Delay	49%	Move	No	50%	Delay	Yes	82%	Delay	Yes
7	89%	Delay	85%	Delay	Yes	89%	Delay	Yes	63%	Delay	Yes
8	44%	Move	39%	Move	Yes	41%	Move	Yes	74%	Delay	No
9	44%	Move	44%	Move	Yes	47%	Move	Yes	71%	Delay	No
10	44%	Move	48%	Move	Yes	51%	Delay	No	58%	Delay	No
11	89%	Delay	84%	Delay	Yes	85%	Delay	Yes	79%	Delay	Yes
12	78%	Delay	65%	Delay	Yes	68%	Delay	Yes	17%	Move	No
13	78%	Delay	68%	Delay	Yes	68%	Delay	Yes	71%	Delay	Yes
14	44%	Move	38%	Move	Yes	38%	Move	Yes	17%	Move	Yes
15	44%	Move	47%	Move	Yes	47%	Move	Yes	48%	Move	Yes
16	56%	Delay	70%	Delay	Yes	70%	Delay	Yes	80%	Delay	Yes
17	89%	Delay	89%	Delay	Yes	89%	Delay	Yes	80%	Delay	Yes
18	89%	Delay	80%	Delay	Yes	78%	Delay	Yes	80%	Delay	Yes
19	56%	Delay	65%	Delay	Yes	70%	Delay	Yes	80%	Delay	Yes
20	44%	Move	47%	Move	Yes	51%	Delay	No	77%	Delay	No
21	78%	Delay	71%	Delay	Yes	78%	Delay	Yes	67%	Delay	Yes
22	56%	Delay	69%	Delay	Yes	70%	Delay	Yes	48%	Move	No
23	78%	Delay	74%	Delay	Yes	68%	Delay	Yes	77%	Delay	Yes
24	67%	Delay	56%	Delay	Yes	50%	Move	No	60%	Delay	Yes
25	78%	Delay	61%	Delay	Yes	62%	Delay	Yes	48%	Move	No
26	78%	Delay	76%	Delay	Yes	71%	Delay	Yes	79%	Delay	Yes
27	89%	Delay	79%	Delay	Yes	80%	Delay	Yes	84%	Delay	Yes
28	67%	Delay	74%	Delay	Yes	68%	Delay	Yes	58%	Delay	Yes
29	44%	Move	49%	Move	Yes	41%	Move	Yes	74%	Delay	No
30	56%	Delay	67%	Delay	Yes	59%	Delay	Yes	80%	Delay	Yes
Model performance					**97%**			**87%**			**73%**

Figure 5.14: Choice experiment - Sample - Population - Prior model comparison

By means of the different model's results posed above, we can answer the **fifth research question**:

How do different criteria trade-off valuations affect choice outcomes in a sample, population and prior model?

As can be seen, the *sample model* is an excellent predictor of expert choices reaching 97% accurateness. This result is not surprising since we include all parameter estimates in this model. Including all parameter estimates ensures that the model captures every subtle utility contribution in the total utility. By looking at the 30 choices, we note that only one choice is predicted wrong by the sample model. In the *population model*, we exclude non-significant parameters, which do not certainly affect the population's choices. As a result, we can see that the population model predicts less accurately than the sample model reaching 87% accuracy. The exclusion of insignificant parameters causes part of the utility, and thus part of the choice explanation, to be removed. Because of the withdrawal of these subtle utility contributions or decreases, it is no surprise that the population model is a worse predictor than the sample model. If we look at the choice scenarios where the population model is wrong, we see that these scenario choices are really near or on the tipping point of 50%. Thus we note that the population model performs worse than the sample model on borderline cases. In the *prior model*, we establish criteria importance by initial criteria valuation. Direct human judgement is used in a three-step linguistic interval scale: nice-to-have, important, and critical. Each attribute was assigned one of these scores by asking for importance compared to the other criteria. Scores were assigned directly by a group of three supervisors. Subsequently, these scores were translated to prior values of 0.5, 1.0 and 1.5 (except for dummy coded criteria). We see that the prior model is the worst predictor of expert choices, reaching 73% accuracy. In 8 of the 30 choice scenarios, the prior model prediction deviates from the initial expert choices. In contrast to the population model, we see that also in less questionable scenarios, the model predicts in the opposite direction.

Altogether, the prediction models indicate a straightforward story. In terms of performance, the sample model makes the least wrong predictions, followed by the population model and the prior model. Although this may not be a big surprise, each model's usability depends on its intention. We discuss model intention and usability in the final chapter.

6 Discussion, Conclusion & Recommendations

6.1. Discussion & Conclusion

Behavioural Artificial Intelligence Technology (BAIT) is a promising method that estimates criteria weights by implicit choices instead of direct judgement. Doing so proposes to capture more realistic criteria trade-offs, which can subsequently be used in an interactive decision-making tool to automate decisions. In our research, multiple sub-questions have been answered to address the applicability of this new methodology to E-commerce distribution services. Through a decision-identification framework, decision nature and requirement classifications and modelling & analysis with BAIT methodology, we can answer the final, **sixth sub question** of this research:

In what ways can Behavioural AI Technology contribute to improve decision-making in E-commerce distribution services?

We argue that BAIT is most applicable to decisions with a high degree of complexity and need for explainability. Compared to decision rules, BAIT allows inclusion of criteria-specific characteristics, like combinatorial constraints, non-linearity and interdependent interaction effects. Moreover, different decision-maker choice behaviour, like regret-minimization or taboo tradeoff-aversion, may be grounded and thus be mimicked in the prediction model. In terms of data input, BAIT generates its data by placing the decision-maker in a most-similar situation as when the decision is made. Utilization of these choice experiments, does not require historical input data for criteria weight definition. According to literature research and interviews with key actors in the E-commerce distribution services industry, we find that BAIT is most applicable to *Routing & Scheduling* and *Partner Contracting* decisions. These decisions are known to inhabit a high degree of complexity, often need to be explainable and are not subject to purely data-driven optimization. To gain practical experience and test BAIT's applicability to a routing & scheduling decision, we modelled delayed line-haul decisions. The modelling process is performed closely with an expert group of Operational Supervisors from DHL Express services. Three different decision-making tools were generated.

First, we estimated a *Sample model*, which inherits all of the criteria weights obtained through the BAIT methodology. Second, a *Population model* was estimated, inheriting only criteria weights that are scalable to the Supervisors' whole population. Third, a *Prior model* was estimated, inheriting Supervisors' initially assigned criteria weights. Next, we let all three models predict the choice scenarios of the choice experiment and compared these predictions to the actual choices of the Operational Supervisor, expert group. We find that the Sample model performs best at 97% prediction accuracy, followed by the Population model, at 87%, and the Prior model, at 73%. This conclusion indicates that the Sample model (thus, BAIT) captures more observed criteria trade-off values than the other two models. Although this is a promising result, we must note that the Sample model is optimized to predict the choices of the, in our case, relatively small (9 respondents) sample group. If the goal is to predict decisions of the entire population group of DHL Supervisors, the Population model will provide more accurate criteria estimates. Moreover, it is not unthinkable that the choice experiment might have missed some decisive criteria, inhabited not understood formulations, or could not imitate the actual choice environment to a high degree. Therefore, the choice experiment may deviate from reality and thus produce non-corresponding decisions. Also, due to a few side notes, the Prior model, based on direct human judgement, might not be as representative as it seems.

First, we determined the directly judged criteria importance scores in discussion with three out of the nine participating supervisors. Therefore, the importance scores might be biased towards the personal importance perceptions of these three supervisors. Second, the importance scores are defined on a three-step scale, being nice-to-have, important and critical. BAIT's parameter estimates, on the contrary, indicate a numeric weight to each criterion ranging between 0 and 1. Because of this difference in step size, supervisors might not be able to initially assign the subtle trade-offs identified by BAIT. Third, due to time constraints, the sample, population and prior model predictions have only been tested on the choices of the choice experiment. To extend the validity of the tool, real-life case testing is essential. Another limitation that we address is that we assume decision-maker rationality by using the MNL method to estimate choices. Also, we assume decision-maker choice consistency by fixing the attribute weights in the prediction model. However, both of these assumptions can be questioned in real-life decision settings. Therefore, a future study might be conducted to break these assumptions and estimate a Discrete Choice Models (DCM) that incorporates non-rational choice behaviour.

Regarding the research question above, we argue that BAIT could open informational black boxes in E-commerce *Routing & Scheduling* and *Partner Contracting* decision-making. Consequently, the decision might be automated, or a safety net is provided to correct the decision-maker in case of a large deviation from the collegial advice given by the decision-making tool. Moreover, we regard Discrete Choice Analysis (DCA) to be a supplementary method to existing Multi-Criteria Decision-Analysis (MCDA) methods when the purpose of the decision tool is to replicate choices. Therefore, regarding academic reflection, we see BAIT not as a competitor of the existing MCDA methods identified in subsection 1.1.2 but as a promising addition when the goal is to reproduce choices. Moreover, by incorporating DCA-established criteria weights into an easy-to-use decision-making tool, BAIT is complementary to conventional Discrete Choice Models (DCM) and increases the automation potential for decisions. Our research shows that decision tool performance differs when modelling criteria weights with DCA compared to direct judgement. This indicates the presence of bias or hidden expertise among the experts. As a second purpose, BAIT proposes to use introspection to provide complementary information by adjusting for biases and translating descriptive criteria weights into prescriptive ones. However, it remains unclear how the latter would specifically work in theory and practice and is therefore subject to further research. In this regard, further research could underpin BAIT's contribution as a competitor to other MCDA preference elicitation methods by using introspection to derive prescriptive criteria weights. As of now, the decision-making tool that we estimated replicates the judgement to DHL's Supervisors' Operations stating: "...% of your colleagues would vote in favour to delay the morning shift". In the next section, we discuss what this practically means for DHL. Herein we give our recommendations to indicate in which directions further research might be conducted, how DHL Express might contribute, and Councyl might advance in the e-commerce sector.

6.2. Recommendations

In line with our conclusion above, we recommend three directions to further assess the applicability of BAIT in E-commerce. First, we discuss future research directions. Second, we discuss specific recommendations for DHL Express as an organisation. Third and finally, we recommend Councyl about the potential of BAIT to support decision-making in the E-commerce sector.

In terms of *future research directions*, our research should be extended with a validation step to more accurately assess the performance of the BAIT decision model. Validation is essential to confirm organisational performance when operationalising a BAIT decision-making tool. In our opinion, there are two directions for validation. From a theoretical perspective, a simulation study could be performed. Using simulation software, the choice situation might be accurately imitated, and rational agents (i.e. decision-makers) added that decide in line with the BAIT decision model. Within this model, DHL's Key Performence Indicators (KPI), as described in subsection 4.1.2, should be included to measure organisational performance. Also, in contrast to the choice experiment, real-life line haul choice scenarios must be used as input instead of most trade-off revealing scenarios. The simulation model should be benchmarked against human-based decision-making to measure the performance of a BAIT decision-making tool. From a practical perspective, a before-and-after study might be performed to assess the performance of the BAIT decision model. In this regard, it is essential to start tracking delayed line haul decisions and note their effect on DHL's KPI measures. Next, the BAIT decision model might be used for a similar period to decide on delayed line haul deliveries. The KPI metrics resulting from this period of time should be tracked and compared to the initial KPI metrics resulting from human choices. With this approach, it is essential that other contextual factors, which might also impact the KPIs, are thoroughly noted and adjusted for. Therefore, we recommend to start testing in a controlled setting.

Besides an additional validation step to assess the performance of the BAIT decision model, Supervisors themselves might also influence model performance. In consequent research, it is essential to determine Supervisor satisfaction regarding the BAIT decision model. If it is determined that user satisfaction is low, this could have consequences for implementation and performance. Also, by assessing user satisfaction, measures could be applied to improve the usability of the tool. Alternatively, the way of using the decision model within the organisation might be changed. To test user satisfaction, we advise the papers of Barki and Huff (1990) and Sprague and Carlson (1982). Both papers offer exact questions and answering formats to assess willingness to participate, participation needs, participation in the implementation process and DSS flexibility. As proven by them, each factor contributes positively to user satisfaction. By presenting these questions after a decision model trial period to a group of experts, user satisfaction can be assessed, and appropriate measures can be taken.

Also, as obtained from Table 5.1, Supervisors spend a different amount of time on the choice experiment. More specifically, a similar amount of Supervisors spend less than 6 minutes and more than 25 minutes on the choice experiment. This raises suspicion as to whether the experiment was filled in attentively by all Supervisors. Answering 32 choice scenarios under 6 minutes indicates approximately 11 seconds per choice scenario which, even for experts, is regarded as very quick. Therefore, we recommend increasing the reliability of the attribute weights by researching whether respondents filled in the experiment attentively. In doing so, we recommend the thesis of de Haas (2022). In chapter 5.3, the author proposes two methods, straightlining and speeding, to assess soft-refusal (i.e. filling in the experiment without thoughtful consideration).

Second, we *recommended to DHL Express* to originate new research projects in line with the recommendations above. By doing so, DHL can validate the positive effect when incorporating a BAIT decision-making tool on organisational performance. Next to validation, more respondents should be involved in the modelling process. By doing so, the parameter estimates become a better indication of the population's preferences, and thus scalability of the model will be improved. Scalability is essential to allow the model to imitate the preferences of all Supervisors and thus use it across all DHL Express service centres. DHL can involve more respondents by start using the model in practice. The BAIT model allows tracking decisions and subsequent adjustment of the weights per these choices. An important notion to consider, however, is that if contextual factors differ largely between service centres (or countries), a service centre (or country) specific sample model might be more accurate for supporting decision-makers. Although these further research directions increase the value of the predictive model, we argue that our model could already be used as it is.

With the insight into hidden expertise and the decision-making tool, we see four potential (future) avenues for application. First, Supervisors might be confronted with their implicit preferences, which might open a discussion among Supervisors as to what is important. Especially the criteria *Notification time (NO)*, *Number of non-timebound pieces (NT)* and *Capacity of the afternoon shift (CA)* were overestimated by the Supervisors in terms of importance. Contrarily, This notion applies to the *Arrival time (AT)*, the *Number of stops (NS)* and the *Number of twelve-hour pieces (TH)*, which the Supervisors underestimated in terms of importance. With this insight, we recommend that DHL internally discuss why Supervisors over or underestimate these criteria and if it is desired to value them as less or more important than stated initially. Consequently, criteria importance can be more accurately defined, which might lead to better decision outcomes.

Second, the BAIT decision tool might replace decision-makers as the decisive factor for the decision. Decision automation, at this stage, is not feasible due to the risk of the incompleteness of choice experiment context and limited knowledge of supervisors' unobserved randomness. However, including more Supervisor preferences by actively starting to use the model improves model performance and, thus, automation. Third, supervisors might be assisted in their decision, which we see as the most likely application at this stage. We note that operational supervisors have various tasks, and E-commerce operational performance depends on small time margins. Therefore, mistakes are easily made, and discussion time is scarce. As a result, a second control check, which provides accurate collegial advice, might support supervisors to make more overthought decisions in less time. In line with this notion, a second check might reduce discussion and assessment time, allowing supervisors to assign more time to other day-to-day tasks. At the origination of this research, this was and still is, seen as one of the main advantages of implementing a BAIT-oriented decision tool at DHL Express. Lastly, we see the model as a method to relieve supervisors from educational tasks. Hypothetical, close-to-realistic choice scenarios could be included in the training process of future Supervisors. This might, subsequently, shorten the overall training period and free up the initial education time of senior Supervisors. In an industry where competition is fierce and time margins are small, relieving the time pressure of senior supervisors by incorporating a BAIT decision-making tool could provide a competitive advantage over other industry players.

Third, we *recommend to Councyl* to continue its partnership with DHL Express. As mentioned, there are multiple directions for further research that new graduate students could discover. The size of DHL's operations, being an internationally oriented e-commerce firm, implies the presence of additional resources to originate future research. Besides that, we argue that the application of BAIT does not need to be limited to regional service centres only. Regional service centres are positioned at the end of the supply chain and are served by domestic and international hubs. Because processes in these service centres are similar, differing only in the scale of operations, we see more potential for BAIT. This notion signifies the company-wide applicability of BAIT to assist and eventually automate decisions at DHL Express. Also, our research focused solely on the routing & scheduling of delayed line hauls. Other decision environments, like partner contracting, might also be investigated. This line of thought was confirmed when we presented the results to DHL. DHL's representatives were enthusiastic about the insights provided by BAIT. However, in an organisation of the size of DHL, they stated that there should be a problem to solve to achieve smooth implementation. Herefore, the BAIT decision-making tool should be wider than delayed line haul decisions and applicability to other decisions where matters are more urgent should be investigated. Again, this indicates the company-wide potential of a BAIT decision tool at DHL. Therefore, further investigation of its applicability is recommended.

Deriving from this, we state that the application of BAIT can extend beyond DHL Express. As was obtained from the expert interviews in section 2.2, other Logistic Service Providers (LSP) and E-tailers operate similarly as DHL. Organisational blocks consist of differently sized distribution centres in which decisions partly rely on human expertise. Therefore, we advise continuing the search for a problem-solving application for BAIT at DHL. When a pressing matter is found, this should motivate Councyl to pursue other e-commerce firms more effectively by providing a solution to a pressing problem. Regarding this search, we recommend inspecting the expert interviews of section 2.2, which can be used as a handout to identify which e-commerce firms encounter similar pressing problems.

A
Appendices

A.1. Scientific paper

Automation of expert decisions in delayed line haul deliveries: an application of the Behavioural Artificial Intelligence Technology

J.A. Smeets[1], A. Nadi Najafabadi[2], A. van Binsbergen[3], L.A. Tavasszy[4] and C.G. Chorus[5]

[1] *Candidate of MSc Transport, Infrastructure & Logistics, Delft University of Technology*
[2] *Faculty of Civil Engineering and Geosciences, Delft University of Technology*
[3] *Faculty of Civil Engineering and Geosciences, Delft University of Technology*
[4] *Faculty of Technology, Policy and Management, Delft University of Technology*
[5] *Co-founder & Scientific advisor, Councyl Behavioural Technologies, Delft*

Reception date of the manuscript: 14/11/2022

Abstract—Making decisions regarding delayed line haul transport is a very demanding and complex process in E-commerce distribution centres. Automating this process can decrease decision-maker discussion and assessment time and, as a result, allow decision-makers to spend more time on other demanding tasks like sorting and distributing. Automation in this domain increases on-time deliveries and strengthens the E-commerce firms' competitive position. However, such decisions involve experts' knowledge, discussions among planners, and complex thought processes. Therefore, it is necessary to involve planners' inclinations and preferences to automate their decisions. This paper proposes the Behavioural Artificial Intelligence Technology (BAIT) to automate expert decisions in delayed line haul deliveries. BAIT uses fundamental Discrete Choice theory under the hood to capture expert preferences effectively, and incorporates these into a decision-making tool. We use this method in a case study to replicate expert decisions in delayed line haul deliveries at DHL Express. The case study results show that BAIT can accurately replicate expert decisions.

Keywords—E-commerce, delayed line haul decisions, automation, Behavioural Artificial Intelligence Technology, Discrete Choice Analysis, decision-making tool

I. Introduction

In the early days of E-commerce, retailers used phone calls and paper letters to communicate with customers. As opposed to current days, customers did not derive convenience from quick deliveries but from the ability to order goods while continuing life as usual. However, this changed with the rise of the internet from the 2000s onwards. E-commerce saw an enormous increase in demand, which incentivised online retailers (E-tailers) to coordinate multiple distribution centres in strategic locations. Consequently, these E-tailers were left with complex fulfilment network set-up and coordination decisions, which proved too challenging to combine with other company operations [1]. As a result, many E-tailers decided to focus on core competencies like research & development and large-scale production. They partnered with Logistic Service Providers (LSPs) to outsource (part of) their logistic services. Simultaneously, unique logistic experience accumulated over the years within LSPs [2] that has become essential to improve the online shopping experience and thus ensure a competitive advantage over other firms [3]. As a result, next to product quality, consumer satisfaction increasingly depends on product distribution services [4]. The quality of these distribution services depends partly on repetitive decisions that rely on human expertise. In the E-commerce context, delayed line haul decisions are one of the most demanding and complex tasks that require human expertise. These decisions take up a large portion of Operational Supervisor time which, in this high-paced industry, is scarce. Therefore, decision automation decreases discussion & assessment time and creates valuable time for more demanding sorting & distributing tasks. As a result, more on-time deliveries lead to improved distribution service quality, which is essential to strengthening the competitive position of E-commerce firms. To support human decision-makers in these often complex decisions, several Decision Support Systems (DSSs) are posed in literature which use sophisticated Multi-Criteria Decision-Analysis (MCDA) methods to derive prescriptive decision criteria valuations: Analytical Network Process (ANP) & Analytical Hierarchy Process (AHP) (Maede Sarkis (2002); Jharkaria & Shankar (2007); Gol & Catay (2007)), Fuzzy Analytical Network Process (FANP) (Saaty & Tran (2007); Zhu (2014); Singh *et al.* (2018)), Interpretive

Structural Modelling (ISM) (Thakkar *et al.* (2005) & Qureshi (2007)), Case Based Reasoning (CBR) & Rule Based Reasoning (RBR) (Isiklar *et al.* (2007)), Decision Making Trial and Evaluation Labatory (DEMATEL) (Govindan & Chauduri (2016)) and the Best-Worst Method (BWM) (Jafar Rezaei (2015)). Using these prescriptive valuations, decision-makers are assisted with complementary information to adjust for decision-maker biases and incomplete information, thus leading to better decisions. Using these methods helps decision-makers to make complex decisions easier by providing recommendations, especially when there are many criteria to trade-off. However, the final decision is yet to be made by decision-makers. Therefore, we argue that MCDA methods are not necessarily the best approach when automation of these decisions matters. To automate the decision-making process of the decision-makers, we need a method that can capture the decision-makers' preferences and accurately replicate their choices. To this end, we propose utilising the Behavioural Artificial Intelligence Technology (BAIT) [16], which uses Discrete Choice Analysis (DCA) to replicate expert choices accurately within a decision-making tool. BAIT has recently been researched only in a medical context to replicate experts' decisions [17]. The question we would like to address in this research is:

How can E-commerce actors use Behavioural Artificial Intelligence Technology to automate delayed line haul decisions and decrease decision-maker discussion & assessment time?

We conduct our research according to the following paper outline:

1. We discuss the method BAIT and fundamental DCA theory.
2. We address the case study of the delayed line haul decision performed at DHL Express.
3. We discuss the chosen survey design set-up to generate the choice experiment.
4. We discuss our results by proposing the delayed line haul criteria trade-offs and the interactive BAIT decision-making tool.
5. We draw up a discussion and several recommendations.

II. Method

BAIT uses DCA to derive expert preferences through statistically designed choice experiments. Concretely, these experiments consist of a set of choice scenarios, each consisting of two or more choice alternatives. By asking respondents to make decisions, Stated Preference (SP) data is derived that contains the criteria importance valuations of the respondent group [18]. Subsequently, the SP data is analysed in the programming language R with the Apollo library to establish concrete decision criteria valuations (i.e. criteria weights). These criteria weights can be entered into utility functions to calculate the utility attached to a particular choice alternative. Equation 1 indicates the different components of the utility function.

$$U_i = V_i + \varepsilon_i = \sum_m \beta_m x_{im} + \varepsilon_i \tag{1}$$

U_i depicts the total utility that a choice respondent can derive from alternative i. V_i is defined as the structural utility, in other words, the utility that we capture with the observed choices, and thus can explain. ε_i is the part of the utility that is not captured by observed choices and is thus referred to as the unobserved utility. The unobserved utility is commonly referred to as the error term. β_m is criteria weight that is attached to a specific attribute m. x_{im} is the attribute level of the specific attribute m for a specific alternative i.

We follow linear-additive MultiNomial Logit (MNL) theory to replicate decisions according to the observed utility. MNL theory assumes decision-maker rationality by defining the error term as a normal-shaped distribution function. Because we have no assumptions about the unobserved preferences of the population, this estimation method is regarded most reliable. Equation 2 indicates the MNL equation used to replicate choices.

$$P_i = \frac{e^{V_i}}{\sum_{j=1\dots j} e^{V_j}} \tag{2}$$

P_i is the probability that alternative i is chosen out of the set of alternatives. V_i is, as mentioned before, the structural utility that the model captures. V_j accounts for all the utilities of the other alternatives in the choice set.

In the end, the criteria weights are incorporated into a decision-making tool replicating expert decisions and can thus be used to automate decision-making. By use of colour coding, individual criterion importance is indicated, contributing to the explainability of the model. Green colour coding indicates a utility contribution, thus favouring a particular policy. Red, on the contrary, indicates a utility decrease, thus not in favour of a particular policy. Also, a filter can be opened for each decision criterion to adjust the model to a specific case. A case study was performed by modelling delayed line haul decisions with BAIT to test the method's applicability.

III. Case study

To assess the potential benefits of BAIT to automate delayed line haul decisions in E-commerce, a case study is performed at DHL Express. This case study is executed in close collaboration with a group of Supervisor Operations, the decision-makers regarding delayed line hauls. In this section, the delayed line haul discussion is elaborated.

1. Decision outcomes

A line haul consists of a truckload of parcels which arrive at a service centre to deliver for the morning or midday shift. In the situation of a delayed line haul, morning shift designated parcels are delayed. As a result, the cargo spaces of the morning shift delivery vans are incomplete, and delivery drivers have to wait for the delayed parcels. Consequently, this extends the delivery times of already loaded morning shift parcels. Therefore, a pressurized situation arises wherein Supervisors choose one out of four choice alternatives:

- Delay the morning shift
- Depart a share of the morning routes
- Move the delayed parcels to the afternoon shift
- Move the delayed parcels to the next day

Of these four choices, we regard the departure of a share of the morning shift as too comprehensive to display in a choice experiment because it is executable in various ways. Moreover, the Supervisors explained that moving parcels to the next day is the worst possible choice alternative. Hence, it is only chosen when the delayed line haul does not include any urgent parcels, which is seldom the case for DHL's Express parcels. Therefore, we decided to limit the choice experiment to two choice alternatives and only include the *Delay the morning shift* and *Move the delayed parcels to the afternoon shift* in the choice experiment.

2. Criteria identification

To identify a list of decision criteria, three group discussions were held with three Operational Supervisors. Besides criteria identification, criteria levels were determined to create unique choice scenarios in the experiment. Table 1 indicates the identified criteria, corresponding criteria levels and the criteria range.

TABLE 1: OVERVIEW ATTRIBUTES, TYPE, LEVELS AND RANGE

Criteria	Type	Level 1	Level 2	Level 3	Range
Notification time	Numeric	30 minutes	105 minutes	180 minutes	150
Arrival time	Ordinal	09:30	09:45	10:00	n.a.
Waiting time	Numeric	10 minutes	20 minutes	30 minutes	20
Number of stops in delayed line haul	Numeric	20	140	160	140
Number of 12:00 pieces in delayed line haul	Numeric	0	20	40	40
Number of non-timebound pieces in delayed line haul	Numeric	50	275	500	450
Capacity of afternoon shift	Numeric	125%	100%	75%	50%
Closeness of delayed stops to midday routes	Ordinal	Close	All around	Far	n.a.

The *Notification time* indicates the duration between the notification that a line haul will be delayed and the moment of arrival of that line haul. The *Arrival time* indicates the arrival time of the last delayed line haul. To indicate the delay of the morning shift included parcels, the *Waiting time* depicts the duration between morning shift finalization and the potential departure with the inclusion of the delayed line haul packages. To address the contents of the delayed line haul, the *Number of stops in the delayed line haul* indicates the number of stops which must be added to the afternoon routes in case of a move of parcels. The *Number of twelve-hour pieces in the delayed line haul* indicates the on-time delivery urgency of the delayed parcels. To address the needed storage room for the delayed parcels and the shifted volume in case of a move to the afternoon shift, we added the *Number of non-timebound pieces in the delayed line haul*. Lastly, the *Capacity of the afternoon shift* indicates the excess parcel capacity of the afternoon shift, and the *Closeness of delayed stops to the midday routes* the fit of the delayed parcels to the afternoon shift routes.

3. Context setting

Besides line haul-specific decision criteria, contextual factors might influence delayed line haul decisions. In this regard, Supervisors mentioned that the size of the Service centre bears influence on the decision to delay or move parcels to the afternoon. Service centres differ in size and, thus, in distributed volumes of parcels. Therefore, Supervisors make different assessments on an identical delayed line haul based on their specific service centre of deployment. To address this notion, a statement is added to each choice scenario in the experiment:

Some of the criteria values below may differ per service centre. We are aware of that. Since we cannot design a separate survey for every service centre, it is chosen to take values that correspond to an average service centre size.

That is why we ask you to make your choices, as being a Supervisor from a medium-sized service center

By pointing out the specific service centre to be medium-sized, we urge Supervisors to make choices accordingly.

4. Criteria importance determination

To establish utility balance in the choice experiment, the Supervisors determined contribution signs & criteria importance scores (i.e. priors) on a linguistic three-step scale being nice-to-have, important or critical. These linguistic scores were translated to numeric criteria weights. By doing so, we can distribute the expected utilities evenly over the choice alternatives in the experiment, which is one of the prerequisites of an efficient choice experiment setup.

5. Non-linear criteria

According to the Supervisors, the notification and arrival time criteria are non-linear. Dummy-coding schemes were designed for these two criteria to account for nonlinearity. Using these schemes, two criteria weights are estimated instead of one to account for two different utility contributions [19]. In Table 2, all criteria' expected utility contributions, prior value and linearity characteristics are indicated. The concave label indicates that the utility contribution follows a concave-like shape.

TABLE 2: OVERVIEW ATTRIBUTE PRIORS

Criteria	**Expected sign**	**Expected utility distribution**	**Importance rating**	**Prior**
Notification time	Positive	Concave	Important	n.a.
Arrival time	Negative	Concave	Important	n.a.
Waiting time	Negative	Linear	Critical	-1.5
Number of stops in delayed line haul	Positive	Linear	Critical	1.5
Number of 12:00 pieces in delayed line haul	Positive	Linear	Important	1.0
Number of non-timebound pieces in delayed line haul	Positive	Linear	Important	1.0
Capacity of afternoon shift	Negative	Linear	Important	-1.0
Closeness of delayed stops to midday routes	Positive	Linear	Nice to have	-0.5

6. Interaction effect

Besides direct utility effects, interdependent interaction effects might exist between decision criteria [19]. The Supervisors identified one interaction effect between the *Number of stops in delayed line haul* and the *Closeness of delayed stops to midday routes*. An increased number of stops increases the importance of these stops to the closeness of the midday routes. Because more stops must be added to these midday routes, significantly larger delivery times are at risk. Thus, closeness gets more important. Figure 1 shows a visual explanation of the interaction effect.

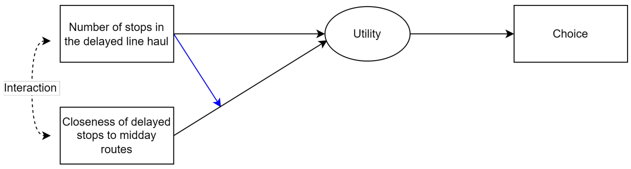

Number of stops in the delayed line haul
Utility
Choice
Interaction
Closeness of delayed stops to midday routes

Fig. 1: Interaction effect Number of stops & Closeness of stops to afternoon

7. Criteria constraint

Criteria constraints might be introduced to exclude non-realistic choice scenarios from the experiment. Referring to the criteria levels of Table 1, a criteria level combination of *"140"* or *"160"* stops in the delayed line haul with *"50"* non-timebound pieces in the delayed line haul would create an unrealistic scenario. By exceeding the total number of pieces, the delayed line haul would consist of more delivery stops than parcels contained, which is physically not possible. By excluding these criteria level combinations in the choice experiment, Supervisors are not encountered with non-realistic choice scenarios. Thus, the choice experiment will rule out these two combinations of these criteria levels.

IV. SURVEY DESIGN

Using Ngene software, we generated a choice experiment design most efficiently. Because our goal is to establish criteria trade-off valuation for all delayed line haul decision criteria, a *D-efficient design* set-up is chosen. D-efficient designs ensure the minimization of every parameter's Standard Error (SE) without putting a special focus on particular criteria [19].

As input for Ngene, 12 criteria weights are defined. The notification and arrival time are modelled with two criteria weights to test for non-linearity. Moreover, a separate parameter adds the interaction effect to the utility function. Lastly, a constant parameter is added as a calibration measure to ensure utility balance. The input parameters for Ngene are defined as follows:

- *Notification time* $\beta_{NO_{dummy0}}$ & $\beta_{NO_{dummy1}}$
- *Arrival time* $\beta_{AT_{dummy0}}$ & $\beta_{AT_{dummy1}}$
- *Waiting time* β_{WT}
- *Number of stops in the delayed line haul* β_{NS}
- *Number of twelve-hour pieces in the delayed line haul* β_{TH}
- *Number of non-timebound pieces in the delayed line haul* β_{NT}
- *Capacity of the afternoon shift* β_{CA}
- *Closeness of delayed stops to the midday routes* β_{CL}
- The interaction effect $\beta_{interactionNSCL}$
- Constant β_{alt1}

Because a binary choice experiment is constructed (i.e. two choices: delay the morning shift or move to the afternoon), we want the utilities of both most extreme choice situations to distribute around zero equally. If this is not the case, the choice sets within the choice experiment will have a tendency towards high utility (i.e. delay morning shift) or towards a lower utility (i.e. move to afternoon), causing an over-representation of choices for a particular alternative. Consequently, less trade-off information can be gathered on the attribute levels attached to the lesser-chosen alternative. To account for this, we add a constant β_0 to the utility function.

The input parameters, accompanied by their corresponding attribute levels (indicated by criteria abbreviation) and the constant β_0, form the observable part of the delayed line haul utility in Equation 3.

$$\begin{aligned} V_{linehaul} = \beta_{Alt1} + \beta_{NOdummy0} * NO_{dummy0} + \beta_{NOdummy1} \\ * NO_{dummy1} + \beta_{ATdummy0} * AT_{dummy0} + \beta_{ATdummy1} * AT_{dummy1} \\ + \beta_{WT} * WT + \beta_{NS} * NS + \beta_{TH} * TH + \beta_{NT} * NT + \beta_{CA} * CA \\ + \beta_{CL} * CL + \beta_{interactionNSCL} * NS * CL + \beta_0 \end{aligned} \tag{3}$$

In total, 30 choice sets were generated by Ngene.

V. RESULTS

1. Sample description

The survey was distributed to 10 Supervisors, of which 9 completed the choice experiment. Table 4 shows the respondent characteristics. In addition to these characteristics, the completion time is indicated. As can be seen, choice experiment completion time varied largely between Supervisors, potentially indicating a differing degree of dedication among the respondent group.

TABLE 3: PARAMETER ESTIMATES

Critera	Estimate	Relative importance [Estimate]	P-value	Significant?	Relative importance [Prior]	Expected curvature
β_{Alt1}	-1.11747	N.A.	0.027513	yes	N.A.	Uniform
$\beta_{NTdummy0}$	-0.08556	1.64%	0.807713	no	10.53%	Concave
$\beta_{NTdummy1}$	0.02705		0.939348	no		
$\beta_{ATdummy0}$	1.21660	17.74%	6.7911e-04	yes	10.53%	Concave
$\beta_{ATdummy1}$	0.87721		0.008789	yes		
β_{WT}	-0.50049	14.60%	0.010473	yes	15.79%	Linear
β_{NS}	0.69076	20.14%	0.009993	yes	15.79%	Linear
β_{TH}	0.82241	23.98%	5.265e-06	yes	10.53%	Linear
β_{NT}	0.07766	2.26%	0.707553	no	10.53%	Linear
β_{CA}	-0.04132	1.20%	0.821249	no	10.53%	Linear
β_{CL}	0.25938	7.56%	0.239227	no	5.26%	Linear
$\beta_{interactionNSCL}$	-0.18626	10.86%	0.305734	no	10.53%	Linear

2. Model fit

The MNL model produces three model fit metrics. First, *Log-Likelihood (LL)* and *Rho-square* ρ^2 metrics are calculated. The higher the LL (i.e. closer to zero), the better the model can predict the dataset. The ρ^2 indicates the relative difference between the LL of a model where all estimated parameters are zero and the estimated parameters of the final model. Whenever the ρ^2 is 0, the model is no better than a random model (i.e. rolling a dice). However, if the ρ^2 gets closer to 1, the estimated parameters fit the dataset better, hence approaching a deterministic model. Table 5 indicates the model fit metrics for the estimated MNL model.

TABLE 4: DEMOGRAPHIC CHARACTERISTICS & COMPLETION TIME RESPONDENTS

Demographic factor	Category	Respondents [%]
Gender	Male	100%
	Female	-
Service center location	Den Hoorn	62.5%
	Breda	25.0%
	Amersfoort	12.5%
Years at DHL	<5 years	11.1%
	6 - 10 years	66.7%
	11 - 15 years	22.2%
	16 - 20 years	-
	>20 years	-
Completion time	<6 minutes	22.2%
	6 - 10 minutes	22.2%
	11 - 15 minutes	-
	16 - 20 minutes	33.3%
	21 - 25 minutes	-
	>25 minutes	22.2%

Because of the small population sample (i.e. nine respondents), we do not expect a very high ρ^2. Nevertheless, according to a ρ^2 of 0.1620, the estimated parameters perform much better than a model with parameters of zero would. In addition to the two model fit metrics mentioned above, we calculate the Mean Absolute Deviation (MAD). The MAD score is calculated by summing the differences between the actual percentage of delay choices and the model-estimated delays for each choice scenario and averaging over the number of scenarios. The decision-making tool indicates a MAD of 7,16%, which indicates that, on average, the decision model deviates 7,16% from the respondent choice.

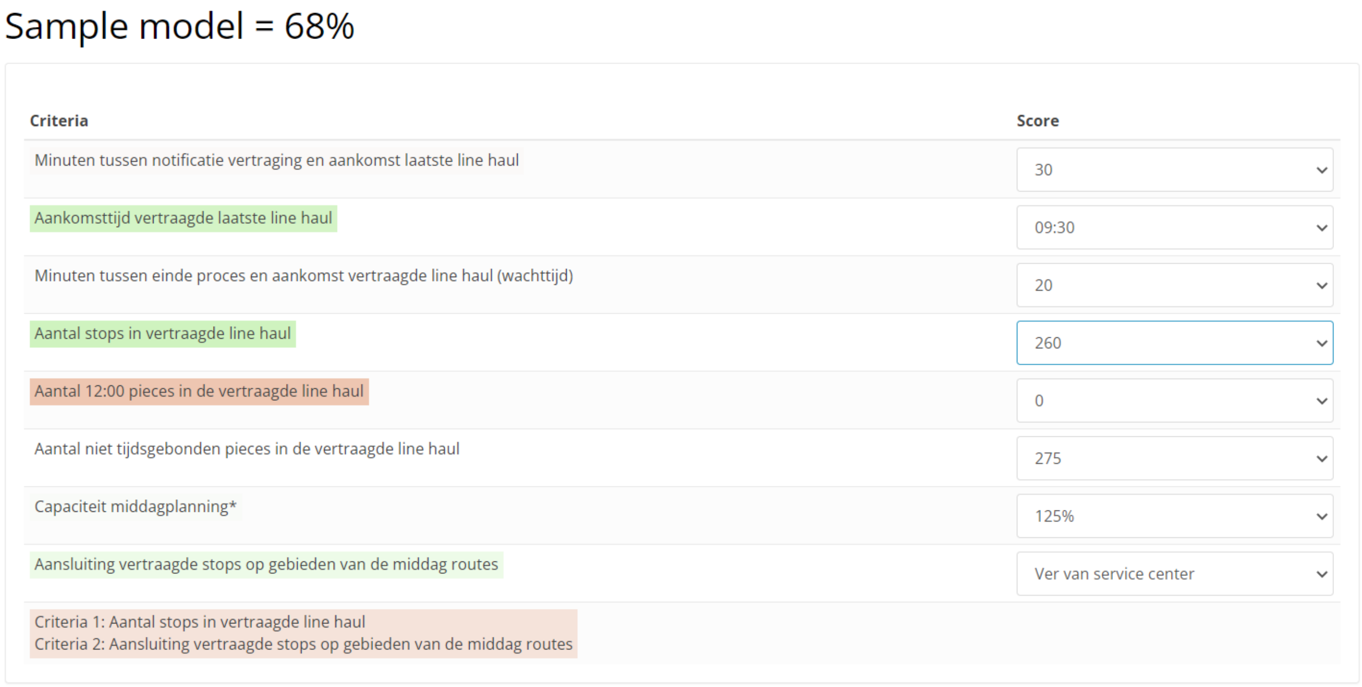
TABLE 5: MODEL FIT METRICS

Model fit metrics	
LL	-167.29
ρ^2	0.1620
MAD	7,16%

3. Parameter estimates

Table 3 indicates all criteria accompanied by their estimate, relative importance, p-value, significance, prior relative importance value and expected curvature. The estimate depicts the utility contribution per 1 unit increase or decrease of the attribute level. In addition, the relative importance compares the utility contribution of each criterion by indicating its contribution to the decision. The p-value is posed to address the transferability of each criterion estimate to the population of Supervisors. A p-value beneath 0.05 labels the criterion as statistically significant, thus scalable to the population.

Regarding the relative importance, we note that the Arrival time, Number of stops and the Number of twelve-hour pieces *contribute most* to the decision. In contrast, the Notification time, Number of non-timebound pieces and the Capacity of the midday shift *contribute least* to the decision. Of all criteria estimates, we note that the Arrival time, Waiting time, Number of stops and Number of twelve-hour pieces are statistically significant and thus scalable to the population. In terms of prior importance comparison, the results show that the Notification time, Number of non-timebound pieces



Sample model = 68%

Criteria	Score
Minuten tussen notificatie vertraging en aankomst laatste line haul	30
Aankomsttijd vertraagde laatste line haul	09:30
Minuten tussen einde proces en aankomst vertraagde line haul (wachttijd)	20
Aantal stops in vertraagde line haul	260
Aantal 12:00 pieces in de vertraagde line haul	0
Aantal niet tijdsgebonden pieces in de vertraagde line haul	275
Capaciteit middagplanning*	125%
Aansluiting vertraagde stops op gebieden van de middag routes	Ver van service center
Criteria 1: Aantal stops in vertraagde line haul Criteria 2: Aansluiting vertraagde stops op gebieden van de middag routes	

Fig. 2: Decision-making tool for delayed line haul decisions

and Capacity of the midday shift were largely *overestimated* by the Supervisors. In contrast, the Arrival time & Number of twelve-hour pieces were largely *underestimated*.

4. Decision-making tool

The delayed line haul decision can be automated by loading the parameter estimates into the decision-making tool. Figure 2 shows a screenshot of the decision-making tool. In this case, we generated a sample decision model, incorporating all criteria estimates from the experiment. Filters are indicated to adjust the criteria levels to the specific choice scenario. Also, the colour coding is added to indicate the utility contribution of each criterion. Above the list of decision criteria, probabilistic advice is indicated by the tool that should be interpreted as: "68% of your colleagues would delay the morning shift in this particular situation."

By applying the Sample model to estimate the choice scenarios of the choice experiment, we found that in 29 of 30 scenarios, the sample model can replicate the choice of the Supervisors corresponding to a hit rate of *97%*. We also generated a population model, which only incorporates the statistically significant criteria estimates. This decision tool correctly predicts *87%* of the choice scenarios. Lastly, we generated a prior model in which the Supervisor initially determined criteria importance scores were included. This final model can correctly predict *73%* of the choice scenarios. We finalize our paper with a discussion and recommendations to address the applicability of BAIT to automate delayed line haul decisions.

VI. Discussion

BAIT is a promising method that estimates accurate criteria weights by implicit choices instead of asking decision-makers directly what is important. In addition, these criteria weights can be loaded into a flexible decision-making tool to accurately replicate decision-makers choices. We modelled delayed line haul decisions and generated a decision-making tool with the cooperation of a group of Operational Supervisors at DHL Express. This decision-making tool can accurately reproduce choices in a controlled setting and thus shows potential to automate the delayed line haul decision in real-life. Automating this decision would save Operational Supervisors a significant amount of discussion & assessment time, which can be assigned to more demanding sorting & distribution tasks.

Regarding the criteria weight estimation process, our results show that decisive decision criteria were initially under or overestimated by DHL's Supervisors. This signifies the presence of decision-maker cognitive bias or hidden expertise. It also confirms that people do not make accurate judgements when assigning criteria importance; thus, defining criteria weights for automation purposes should be based on decision-makers' choices. By leveraging the BAIT method, decision-maker choices are collected in a safe environment and might be subsequently used to open informational black boxes. Although this is the case, we note a few limitations of this research. First, we determined the prior importance scores in discussion with three out of the nine participating Supervisors. Therefore, the importance scores might be biased towards the personal importance perceptions of these three Supervisors. Second, the decision-making tool has only been tested on the choices of the choice experiment. To increase model validity, real-life case testing is essential. Third and finally, we assume decision-maker rationality by using the MNL method to estimate choices.

Also, we assume decision-maker choice consistency by fixing the attribute weights in the prediction model. In future research, these assumptions can be questioned by utilizing the decision tool in real-life decision settings.

VII. Recommendations

Regarding the automation of delayed line haul decisions, we see three directions for implementing the decision-making tool. First, Supervisors might be confronted with their implicit preferences, which might open a discussion among Supervisors as to what is important. Consequently, criteria importance can be more accurately defined, which might lead to better decision outcomes. Second, the BAIT decision tool might replace decision-makers as the decisive factor for the decision. Although the decision tool shows automation potential, it is not yet feasible to fully automate decisions due to incomplete Supervisor preferences and unobserved decision-maker randomness. Therefore, we strongly recommend including more Supervisor preferences to increase the model's validity. DHL can do this by putting the decision-making model into practice since it allows the tracking and adjusting of criteria weights according to new decisions. Third and finally, the decision tool can assist Supervisors in their decision. As noted before, operational Supervisors have various tasks, and E-commerce operational performance depends on small time margins. Therefore, discussion time is scarce, and mistakes are easily made. As a result, a second control check, which provides accurate collegial advice, supports Supervisors to make more overthought decisions in less time.

References

[1] A. Robinson, "Reuters events: Supply chain amp; logistics business intelligence," 2014. [Online]. Available: https://www.reutersevents.com/supplychain/3pl/evolution-e-commerce-logistics

[2] Li, L. Li, C. Jin, R. Wang, H. Wang, and L. Yang, "A 3pl supplier selection model based on fuzzy sets," *Computers Operations Research*, vol. 39, no. 8, pp. 1879–1884, 2012.

[3] Y. Vakulenko, P. Shams, D. Hellström, and K. Hjort, "Service innovation in e-commerce last mile delivery: Mapping the e-customer journey," *Journal of Business Research*, vol. 101, pp. 461–468, 2019.

[4] C. C. Luk, K. L. Choy, and H. Lam, "Design of an enhanced logistics service provider selection model for e-commerce application," in *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*, 2018.

[5] L. Meade and J. Sarkis, "A conceptual model for selecting and evaluating third-party reverse logistics providers," *Supply Chain Management*, vol. 7, no. 5, pp. 283–295, 12 2002.

[6] S. Jharkharia and R. Shankar, "Selection of logistics service provider: An analytic network process (anp) approach," *Omega*, vol. 35, no. 3, pp. 274–289, 2007.

[7] H. Gol and B. Catay, "Third-party logistics provider selection: insights from a turkish automotive company," *Supply Chain Management*, vol. 12, no. 6, pp. 379–384, 2007.

[8] T. L. Saaty and L. T. Tran, "On the invalidity of fuzzifying numerical judgments in the analytic hierarchy process," *Mathematical and Computer Modelling*, vol. 46, no. 7, pp. 962–975, 2007.

[9] K. Zhü, "Fuzzy analytic hierarchy process: Fallacy of the popular methods," *European Journal of Operational Research*, vol. 236, no. 1, pp. 209–217, 2014.

[10] R. Singh, A. Gunasekaran, and P. Kumar, "Third party logistics (3pl) selection for cold chain management: a fuzzy ahp and fuzzy topsis approach," *Annals of Operations Research*, vol. 267, no. 1-2, pp. 531–553, 2018.

[11] J. Thakkar, S. G. Deshmukh, A. D. Gupta, and R. Shankar, "Selection of third-party logistics (3pl): A hybrid approach using interpretive structural modeling (ism) and analytic network process (anp)," *Supply Chain Forum: An International Journal*, vol. 6, no. 1, pp. 32–46, 2005.

[12] M. Qureshi, D. Kumar, and P. Kumar, "Modeling the logistics outsourcing relationship variables to enhance shippers' productivity and competitiveness in logistical supply chain," *International Journal of Productivity and Performance Management*, vol. 56, no. 8, pp. 689–714, 2007.

[13] G. Isiklar, E. Alptekin, and G. Buyukozkan, "Application of a hybrid intelligent decision support model in logistics outsourcing," *Computers Operations Research*, vol. 34, no. 12, pp. 3701–3714, 2007, operations Research and Outsourcing.

[14] K. Govindan, R. Khodaverdi, and A. Vafadarnikjoo, "A grey dematel approach to develop third-party logistics provider selection criteria," *Industrial Management Data Systems*, vol. 116, no. 4, pp. 690–722, 2016.

[15] J. Rezaei, "Best-worst multi-criteria decision-making method," *Omega*, vol. 53, pp. 49–57, 2015.

[16] Councyl, 2020. [Online]. Available: https://councyl.ai/en/councyl-makes-decision-support-available-for-your-organisations/

[17] A. Broeke, J. Hulscher, N. Heyning, E. Kooi, and C. Chorus, "Bait: A new medical decision support technology based on discrete choice theory," *Medical Decision Making*, vol. 41, no. 5, pp. 614–619, 2021.

[18] M. E. Akiva and S. R. Lerman, *Discrete choice analysis: Theory and application to travel demand.* The MIT Press, 1985.

[19] E. Molin, "Introduction to experimental designs," Sep 2015.

A.2. Logistic Biases

To address the presence of cognitive biases in E-commerce deicision making this appendice discusses the work of Tversky and Kahneman (1974) and Knapp et al. (2021). The insight that BAIT has given in differing delayed line haul criteria valuation through direct judgement and implicit choices, indicates the presence of biases or hidden expertise among Operational Supervisors. Clarification on judgement biases in logistics may be used as a start to pinpoint the presence of bias or hidden expertise.

A.2.1. Judgement biases in logistics

Cognitive biases were described by Tversky and Kahneman (1974) in their work Judgement under Uncertainty: Heuristics and Biases. Their work would become the groundwork of Behavioural Economics. According to the authors, in complex situations of incomplete information, decisions are often based on "beliefs concerning the likelihood of uncertain events" to take place. Examples are electoral outcomes, the dollar value in a year and a defendant's guilt. In general, people do not derive these likelihood outcomes for every situation in detail but rely on heuristic principles which are established through experience combined with available information. Although these heuristic principles help the decision-maker in most cases, they sometimes lead to systematic errors, as is proven by the authors in multiple instances. Whenever this happens, we refer to a cognitive bias. In a purely logistic context, Knapp et al. (2021) identified three cognitive decision-maker biases within each of the three before-mentioned decision spheres: strategic, tactical and operational.

Strategic biases

Within the strategic decision sphere, *Imaginability bias*, *Conservatism bias* and *Confirmation bias* are found. Imaginability bias indicates that decision-makers assume that more imaginable events are more probable to happen. An example of this is implementing a new IT system, which is valued more probable by a decision-maker with IT experience. Second is the Conservatism bias. According to this bias, decision-makers assign less value to new information than old information, even when new information proves process improvement. Implementing a new decision support method to select LSPs could encounter Conservatism bias from decision-makers. Third, Confirmation bias could arise in the strategic decision sphere. Confirmation bias indicates that decision outcome are (partly) a reflection of personal decision-maker attitudes. Regarding LSP selection, decision-makers might assign a larger weight to the emission criterion because they value the emission factor as more important than, let's say, transport costs.

Tactical biases

Within the tactical decision sphere, *Presentation bias*, *Confidence bias* and *Situational bias* are identified by Knapp et al. (2021). Presentation bias leads to ambiguity, which states that decision-makers favour simple-looking options over complex ones. An example of this is to extend lead times to customers when the due dates of suppliers become less reliable. Looking for a more complex solution might be a better fix, but it requires more effort from the decision-maker who disregards looking into these. Moreover, Confidence bias leads to the illusion of control. Whenever there's an illusion of control, decision-makers overestimate their ability-solving competencies. Especially in incomplete information, monitoring decision-makers is troublesome, which increases the risk of confidence bias. Lastly, Situational bias could initiate the Ostrich effect. This effect dictates that decision-makers disregard negative information to serve their interests. For instance, an LSP that allows higher transportation volume, thus more productivity for the retailer, is chosen over a more sustainable LSP, which might be better for the long horizon.

Operational biases

Regarding to the operational decision sphere, *Adjustment bias*, *Correlation bias* and *Situational bias* were mentioned by the authors. In practice, Adjustment bias is similar to Conservatism bias. Initial information is anchored by decision-makers who exclude new information. The Correlation bias advocates that decision-makers regard one out of two risks as more probable because it has occurred and been solved recently. This inhibits the misconception that old solutions are a fit for new difficulties, thus, a misplaced correlation between solutions. Lastly, Situational bias indicates that decision-makers encounter bias when they are overloaded by information or time pressurised. These situations lead to emotional decision-making, which is obscure.

Incorrect outcomes are the result. An example is unexpected supply chain congestion which requires ad hoc decision-making while considering loads of information and thus become prone to situational bias.

A.3. Expert interview questions

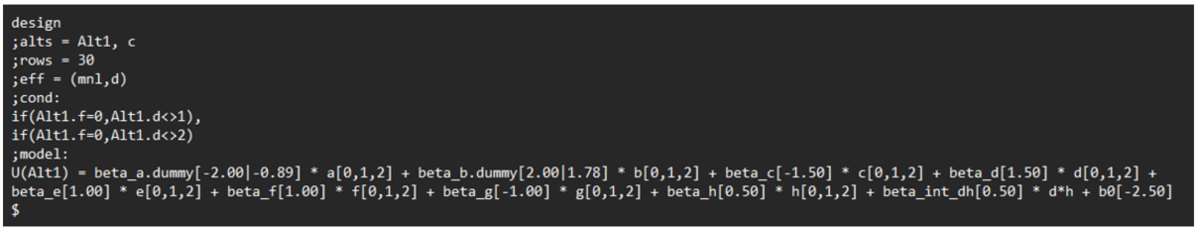
By means of a set of predefined questions, the first part of the decision identification interviews is structured. Lastly, an open question is posed to identify potential decisions that are not captured in the first part. The following questions were subsequently posed to the experts of several E-tailers and Logistic Service Providers (LSPs):

1. **Does [enter name firm] in- or outsource their logistic operations from producer towards delivery to the end-customer?**
    - (a) How is this decided?
        - i. Particular decision criteria?
        - ii. Decision support?
2. **How does the fulfilment network of [enter name firm] look like?**
    - (a) Are customers/stores directly served from the customer?
    - (b) Are customers/stores served through Distribution Centres(DC's)?
    - (c) Is there a distinction between DC's used for online orders and/or replenishment of stores?
        - i. How is this set-up decided?
        - ii. Decision support?
    - (d) How is it decided which products go to which DC?
        - i. Particular decision criteria?
        - ii. Decision support?
    - (e) Do DC's hold certain inventory levels?
        - i. How is this decided?
        - ii. Decision support?
    - (f) Are inventory levels pooled, meaning that inventories of DC's or stores are centralized?
        - i. How is this decided?
        - ii. Decision support?
3. **Which channels are offered to the customer for last-mile deliveries?**
    - (a) How are these channels decided on?
        - i. Particular decision criteria?
        - ii. Decision support?
    - (b) Are delivery channels always directly chosen by the customer?
    - (c) Are all delivery channels always offered to the customer, or connected to existing routes to improve route efficiency?
        - i. How is it decided which channels to offer to the customer?
        - ii. Decision support?
    - (d) Are customer nudged to choose certain channels by e.g. discounts?
        - i. How is this decided?
        - ii. Decision support?
    - (e) How are delivery routes composed and scheduled?
        - i. Particular decision criteria?
        - ii. Decision support?

4. **Does [enter name firm] partner up with one or multiple Logistic Service Providers (LSPs) to facilitate logistic services?**
   (a) How does [enter firm name] partner up with other LSP's?
      i. Particular decision criteria?
      ii. Decision support?
   (b) When partnered up with multiple LSP's besides own logistic services, how is it decided which party fulfills which customers?
      i. Particular decision criteria?
      ii. Decision support?
5. **Are you aware of any human-driven, repetitive decisions that are not yet discussed in this interview, and might be subject to improvement by decision support?**

A.4. Ngene code for choice experiment generation

To generate a most efficient choice experiment design, Ngene software was used. By means of Ngene, a choice experiment can be generated which reveals the most criteria trade-off valuation with minimum required choice sets. As can be seen in the code below, two alternatives were creates, being to delay the morning shift (Alt1) or move to the afternoon (c). The alternative c was fixed to a constant utility of zero. In total, 30 choice sets were produced according to D-efficient design standards. To address the error term, choices were estimated according to the MNL estimation theory. Additionally, the two criteria constrains are indicated as well as the utility function. The utility function is ended with the *b*0 constant of -2.5 to ensure utility balance throughout the experiment.

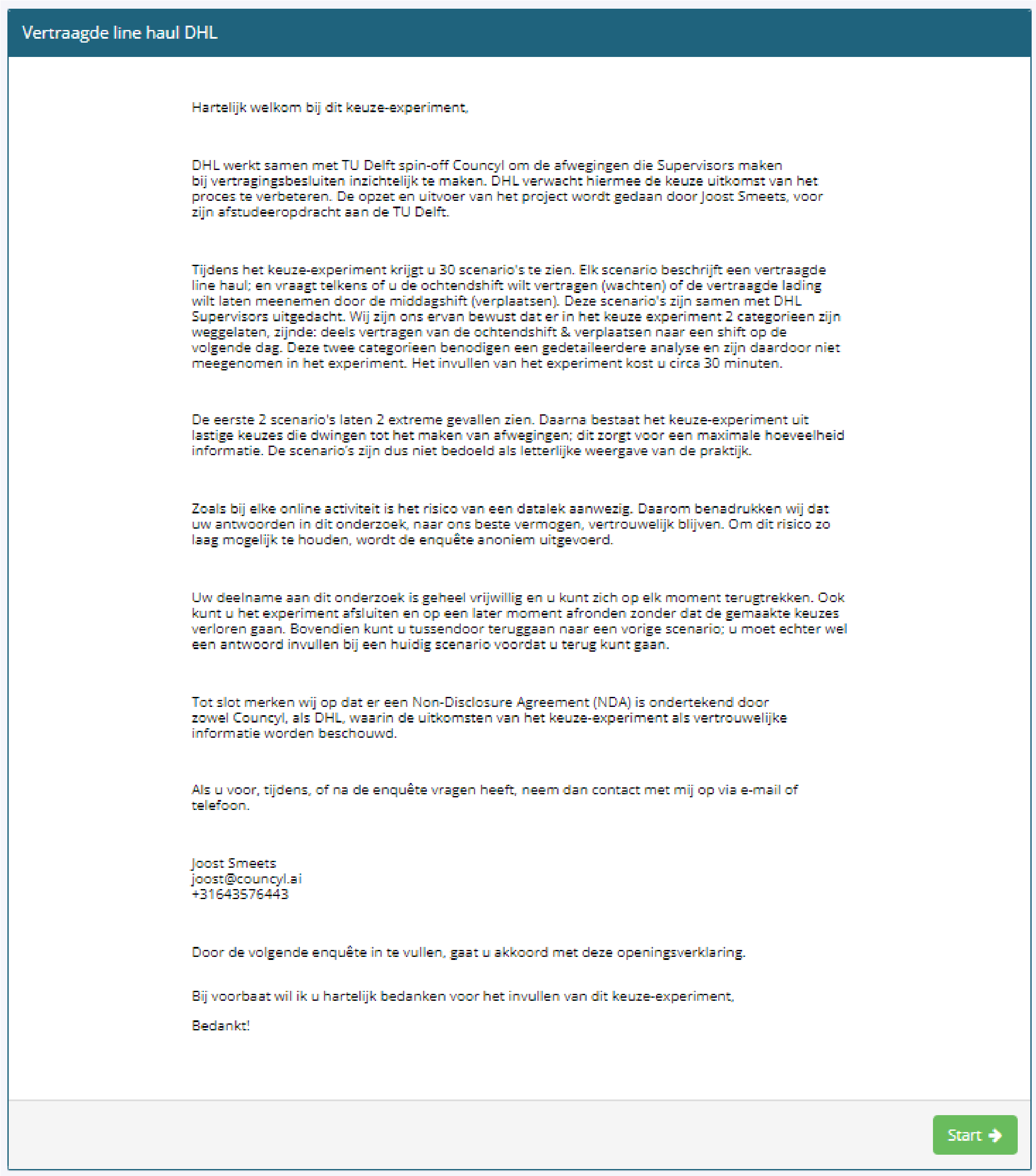


```
design
;alts = Alt1, c
;rows = 30
;eff = (mnl,d)
;cond:
if(Alt1.f=0,Alt1.d<>1),
if(Alt1.f=0,Alt1.d<>2)
;model:
U(Alt1) = beta_a.dummy[-2.00|-0.89] * a[0,1,2] + beta_b.dummy[2.00|1.78] * b[0,1,2] + beta_c[-1.50] * c[0,1,2] + beta_d[1.50] * d[0,1,2] +
beta_e[1.00] * e[0,1,2] + beta_f[1.00] * f[0,1,2] + beta_g[-1.00] * g[0,1,2] + beta_h[0.50] * h[0,1,2] + beta_int_dh[0.50] * d*h + b0[-2.50]
$
```

Figure A.1: Ngene code for generating choice experiment

A.5. Choice experiment snapshots

This appendix section contains several snapshots from the actual choice experiment. First, a statement of informed consent was posed to the respondents. By means of this statement, the respondents are informed on the structure of the choice experiment, agree to us processing their personal data and we note to them that an NDA has been signed between DHL and councyl. Also, contact credentials are posed in case there are any questions or unclearities.



Hartelijk welkom bij dit keuze-experiment,

DHL werkt samen met TU Delft spin-off Councyl om de afwegingen die Supervisors maken bij vertragingsbesluiten inzichtelijk te maken. DHL verwacht hiermee de keuze uitkomst van het proces te verbeteren. De opzet en uitvoer van het project wordt gedaan door Joost Smeets, voor zijn afstudeeropdracht aan de TU Delft.

Tijdens het keuze-experiment krijgt u 30 scenario's te zien. Elk scenario beschrijft een vertraagde line haul; en vraagt telkens of u de ochtendshift wilt vertragen (wachten) of de vertraagde lading wilt laten meenemen door de middagshift (verplaatsen). Deze scenario's zijn samen met DHL Supervisors uitgedacht. Wij zijn ons ervan bewust dat er in het keuze experiment 2 categorieen zijn weggelaten, zijnde: deels vertragen van de ochtendshift & verplaatsen naar een shift op de volgende dag. Deze twee categorieen benodigen een gedetaileerdere analyse en zijn daardoor niet meegenomen in het experiment. Het invullen van het experiment kost u circa 30 minuten.

De eerste 2 scenario's laten 2 extreme gevallen zien. Daarna bestaat het keuze-experiment uit lastige keuzes die dwingen tot het maken van afwegingen; dit zorgt voor een maximale hoeveelheid informatie. De scenario's zijn dus niet bedoeld als letterlijke weergave van de praktijk.

Zoals bij elke online activiteit is het risico van een datalek aanwezig. Daarom benadrukken wij dat uw antwoorden in dit onderzoek, naar ons beste vermogen, vertrouwelijk blijven. Om dit risico zo laag mogelijk te houden, wordt de enquête anoniem uitgevoerd.

Uw deelname aan dit onderzoek is geheel vrijwillig en u kunt zich op elk moment terugtrekken. Ook kunt u het experiment afsluiten en op een later moment afronden zonder dat de gemaakte keuzes verloren gaan. Bovendien kunt u tussendoor teruggaan naar een vorige scenario; u moet echter wel een antwoord invullen bij een huidig scenario voordat u terug kunt gaan.

Tot slot merken wij op dat er een Non-Disclosure Agreement (NDA) is ondertekend door zowel Councyl, als DHL, waarin de uitkomsten van het keuze-experiment als vertrouwelijke informatie worden beschouwd.

Als u voor, tijdens, of na de enquête vragen heeft, neem dan contact met mij op via e-mail of telefoon.

Joost Smeets
joost@councyl.ai
+31643576443

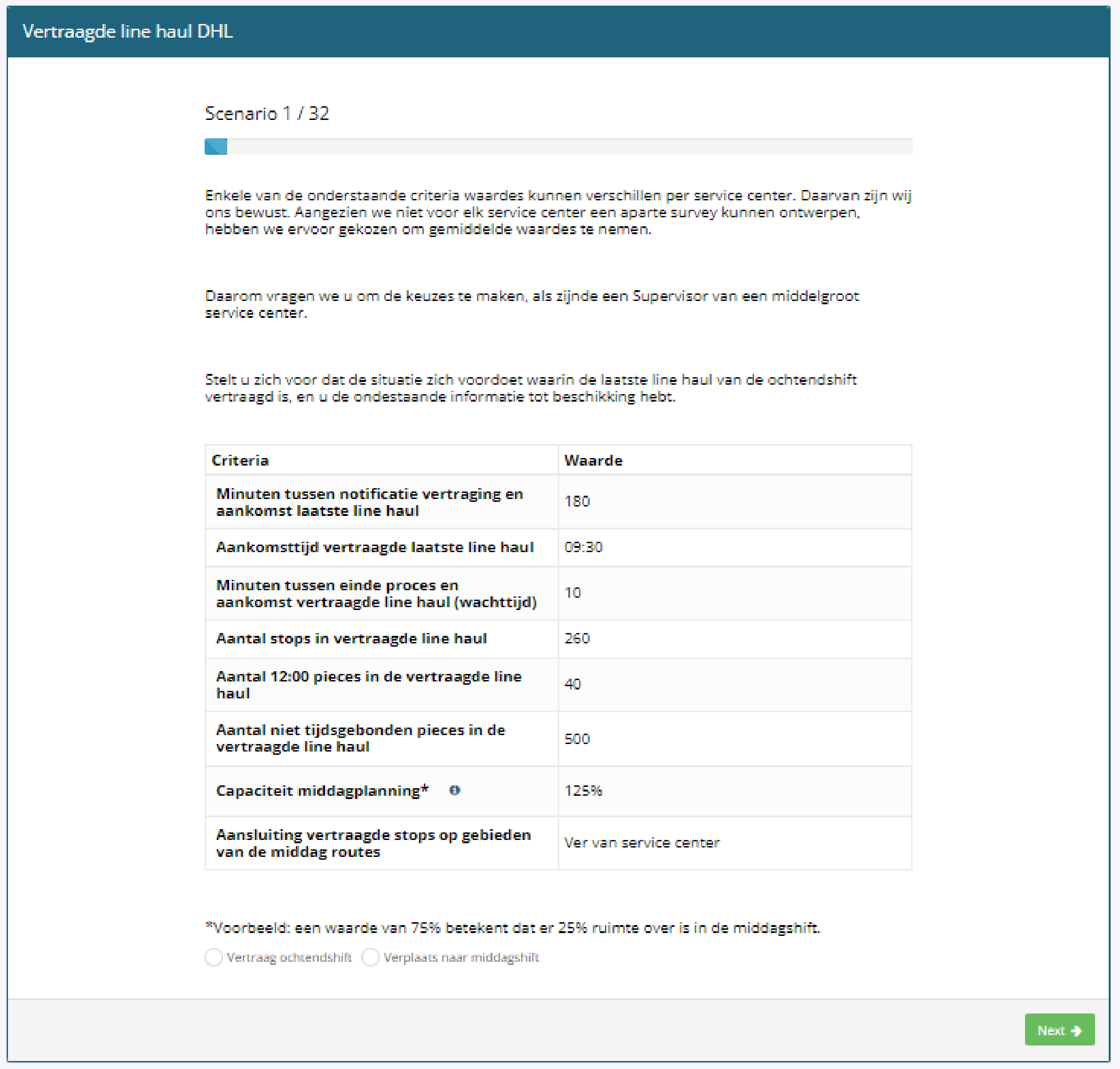
Door de volgende enquête in te vullen, gaat u akkoord met deze openingsverklaring.

Bij voorbaat wil ik u hartelijk bedanken voor het invullen van dit keuze-experiment,

Bedankt!

Figure A.2: Informed consent statement choice experiment

Second, one choice scenario is posed below. As can be seen on the top of the screen, a glider is shown to inform the respondent on their progress. Then a statement is given to inform the respondent that criteria values might differ to the quantities they are used to. Therefore, averages criteria values are included, and the Supervisor is asked to make the choice as being a medium sized service centre Supervisor. Next, the criteria and scenario-specific criteria levels are indicated. Also an explanation of the capacity criteria is given, since this attribute is more difficult to quantify. Last, two boxes are indicated where the Supervisor can make the choice.



Vertraagde line haul DHL

Scenario 1 / 32

Enkele van de onderstaande criteria waardes kunnen verschillen per service center. Daarvan zijn wij ons bewust. Aangezien we niet voor elk service center een aparte survey kunnen ontwerpen, hebben we ervoor gekozen om gemiddelde waardes te nemen.

Daarom vragen we u om de keuzes te maken, als zijnde een Supervisor van een middelgroot service center.

Stelt u zich voor dat de situatie zich voordoet waarin de laatste line haul van de ochtendshift vertraagd is, en u de ondestaande informatie tot beschikking hebt.

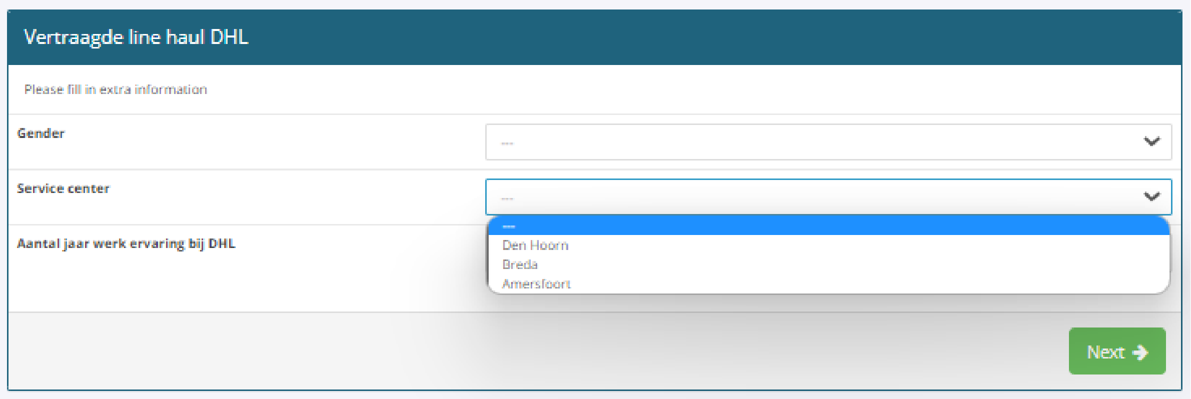
Criteria	Waarde
Minuten tussen notificatie vertraging en aankomst laatste line haul	180
Aankomsttijd vertraagde laatste line haul	09:30
Minuten tussen einde proces en aankomst vertraagde line haul (wachttijd)	10
Aantal stops in vertraagde line haul	260
Aantal 12:00 pieces in de vertraagde line haul	40
Aantal niet tijdsgebonden pieces in de vertraagde line haul	500
Capaciteit middagplanning*	125%
Aansluiting vertraagde stops op gebieden van de middag routes	Ver van service center

*Voorbeeld: een waarde van 75% betekent dat er 25% ruimte over is in de middagshift.

○ Vertraag ochtendshift ○ Verplaats naar middagshift

Figure A.3: Example choice scenario experiment

At last, two questions are posed in regard to the demographic characteristics of the respondents. As mentioned earlier, service centre of deployment and number of years work experience at DHL are asked:

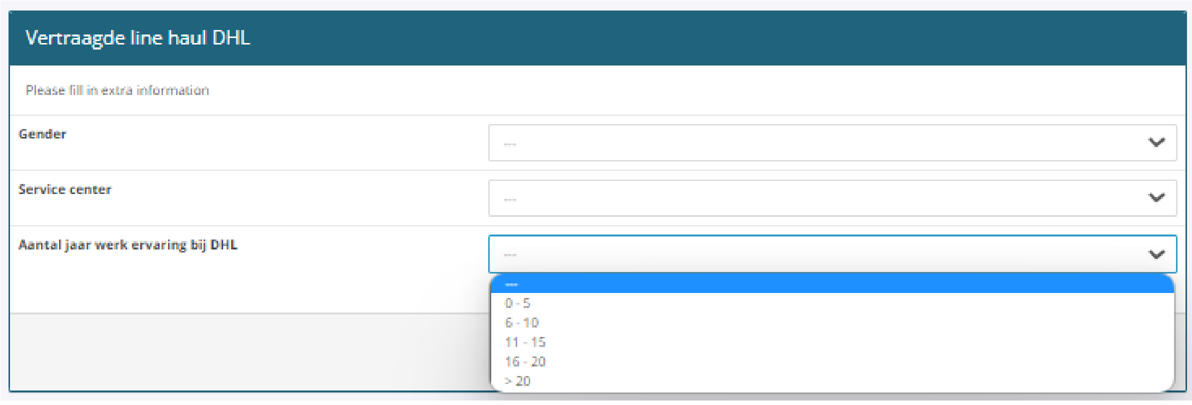


Vertraagde line haul DHL

Please fill in extra information

Gender	---
Service center	--- --- Den Hoorn Breda Amersfoort
Aantal jaar werk ervaring bij DHL	

Figure A.4: Service centre location question choice experiment



Vertraagde line haul DHL

Please fill in extra information

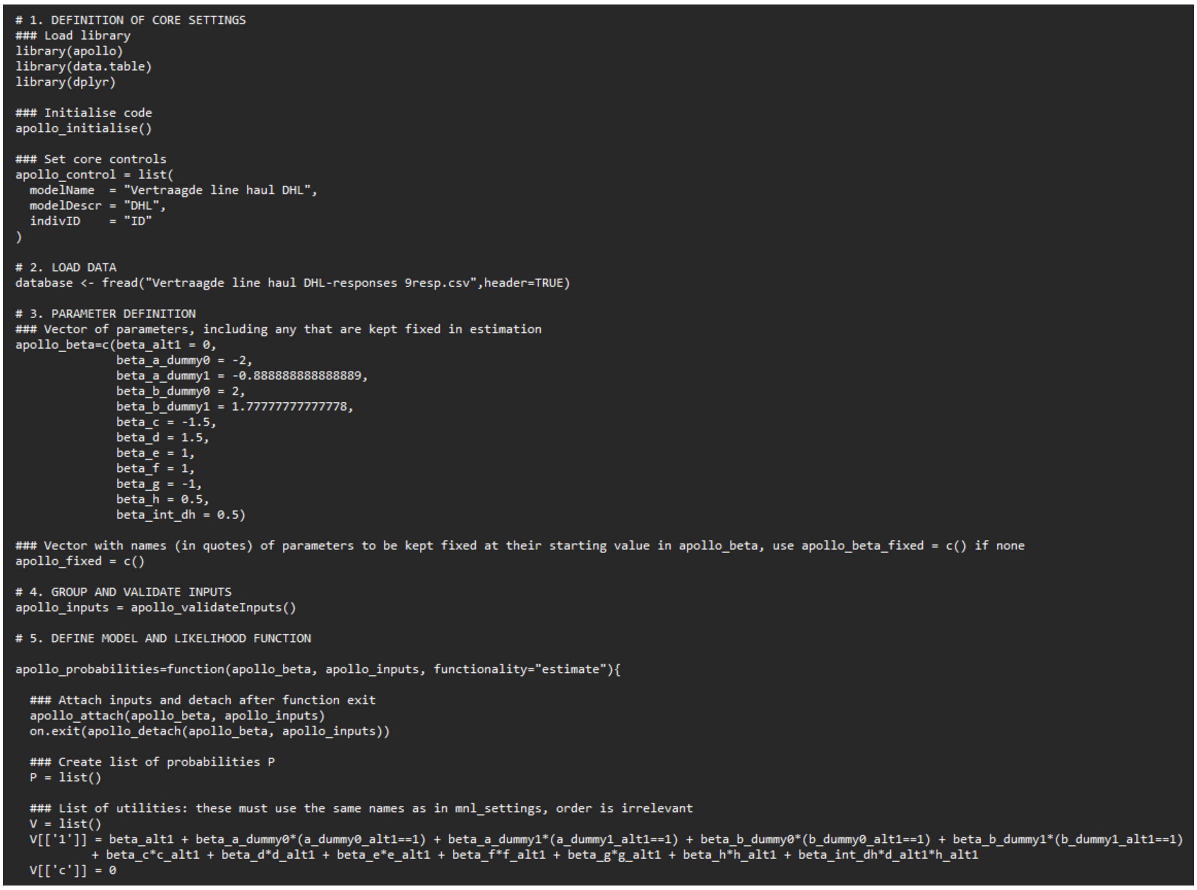
Gender	---
Service center	---
Aantal jaar werk ervaring bij DHL	---

Figure A.5: Deployment years question choice experiment

A.6. Apollo code

MNL estimation was performed in the programming language R with the Apollo library to derive parameter estimates. The two snapshots below show the code used to derive the parameter estimates from the choices captured in the dataset.

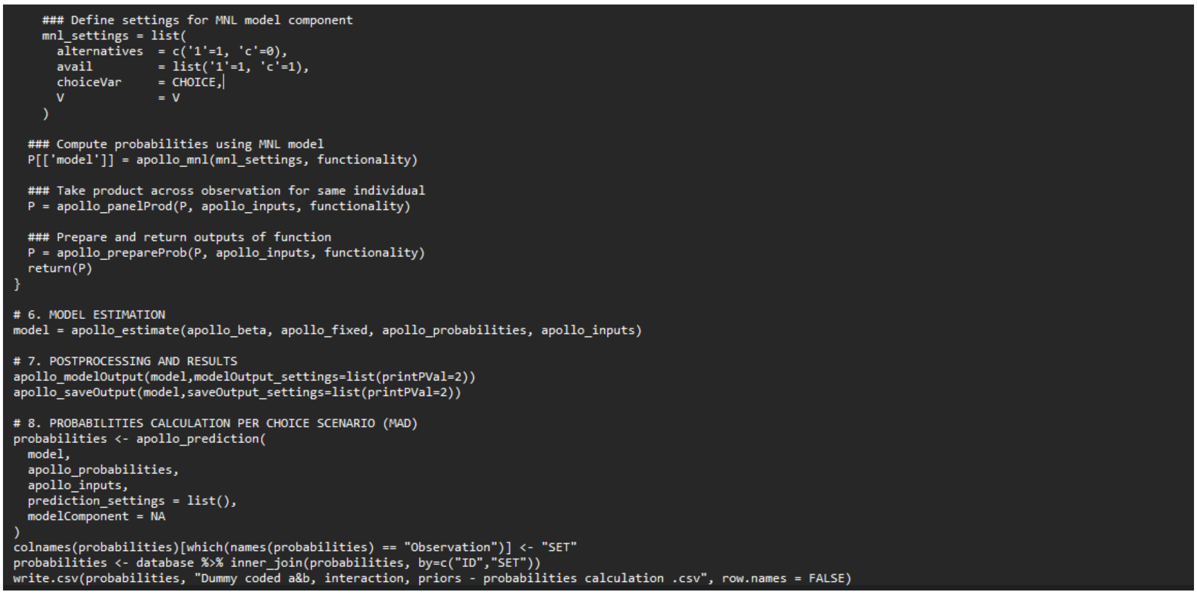
In Figure A.6, we can see the first part of the code where the criteria weights were originated underneath header #3. PARAMETER DEFINITION. As seen, the prior values are indicated as start values. At the bottom of the screenshot, under header #5. DEFINE MODEL AND LIKELIHOOD FUNCTION; we see that a list is created for the choice probabilities, and the two utility functions are defined. Besides, the standard properties of the MNL model estimation are specified.



```
# 1. DEFINITION OF CORE SETTINGS
### Load library
library(apollo)
library(data.table)
library(dplyr)

### Initialise code
apollo_initialise()

### Set core controls
apollo_control = list(
  modelName  = "Vertraagde line haul DHL",
  modelDescr = "DHL",
  indivID    = "ID"
)

# 2. LOAD DATA
database <- fread("Vertraagde line haul DHL-responses 9resp.csv",header=TRUE)

# 3. PARAMETER DEFINITION
### Vector of parameters, including any that are kept fixed in estimation
apollo_beta=c(beta_alt1 = 0,
              beta_a_dummy0 = -2,
              beta_a_dummy1 = -0.888888888888889,
              beta_b_dummy0 = 2,
              beta_b_dummy1 = 1.77777777777778,
              beta_c = -1.5,
              beta_d = 1.5,
              beta_e = 1,
              beta_f = 1,
              beta_g = -1,
              beta_h = 0.5,
              beta_int_dh = 0.5)

### Vector with names (in quotes) of parameters to be kept fixed at their starting value in apollo_beta, use apollo_beta_fixed = c() if none
apollo_fixed = c()

# 4. GROUP AND VALIDATE INPUTS
apollo_inputs = apollo_validateInputs()

# 5. DEFINE MODEL AND LIKELIHOOD FUNCTION

apollo_probabilities=function(apollo_beta, apollo_inputs, functionality="estimate"){

  ### Attach inputs and detach after function exit
  apollo_attach(apollo_beta, apollo_inputs)
  on.exit(apollo_detach(apollo_beta, apollo_inputs))

  ### Create list of probabilities P
  P = list()

  ### List of utilities: these must use the same names as in mnl_settings, order is irrelevant
  V = list()
  V[['1']] = beta_alt1 + beta_a_dummy0*(a_dummy0_alt1==1) + beta_a_dummy1*(a_dummy1_alt1==1) + beta_b_dummy0*(b_dummy0_alt1==1) + beta_b_dummy1*(b_dummy1_alt1==1)
           + beta_c*c_alt1 + beta_d*d_alt1 + beta_e*e_alt1 + beta_f*f_alt1 + beta_g*g_alt1 + beta_h*h_alt1 + beta_int_dh*d_alt1*h_alt1
  V[['c']] = 0
```

Figure A.6: Apollo code part 1. MNL model estimation

In the second part of the code, the actual model is estimated under #6. MODEL ESTIMATION. An important addition to the code is underneath header #8. PROBABILITIES CALCULATION PER CHOICE SCENARIO (MAD). These lines of code allow us to calculate the MAD score for all choice scenarios immediately. This is necessary to calculate the deviation of the decision-tool predicted probabilities and the ones of the actual choice experiment.



```
    ### Define settings for MNL model component
    mnl_settings = list(
      alternatives  = c('1'=1, 'c'=0),
      avail         = list('1'=1, 'c'=1),
      choiceVar     = CHOICE,
      V             = V
    )

  ### Compute probabilities using MNL model
  P[['model']] = apollo_mnl(mnl_settings, functionality)

  ### Take product across observation for same individual
  P = apollo_panelProd(P, apollo_inputs, functionality)

  ### Prepare and return outputs of function
  P = apollo_prepareProb(P, apollo_inputs, functionality)
  return(P)
}

# 6. MODEL ESTIMATION
model = apollo_estimate(apollo_beta, apollo_fixed, apollo_probabilities, apollo_inputs)

# 7. POSTPROCESSING AND RESULTS
apollo_modelOutput(model,modelOutput_settings=list(printPVal=2))
apollo_saveOutput(model,saveOutput_settings=list(printPVal=2))

# 8. PROBABILITIES CALCULATION PER CHOICE SCENARIO (MAD)
probabilities <- apollo_prediction(
  model,
  apollo_probabilities,
  apollo_inputs,
  prediction_settings = list(),
  modelComponent = NA
)
colnames(probabilities)[which(names(probabilities) == "Observation")] <- "SET"
probabilities <- database %>% inner_join(probabilities, by=c("ID","SET"))
write.csv(probabilities, "Dummy coded a&b, interaction, priors - probabilities calculation .csv", row.names = FALSE)
```

Figure A.7: Apollo code part 2. MNL model estimation

Bibliography

Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. The MIT Press.

Akter, S., Wamba, S. F., Mariani, M., & Hani, U. (2021). How to build an ai climate-driven service analytics capability for innovation and performance in industrial markets? *Industrial Marketing Management*, *97*, 258–273. https://doi.org/10.1016/j.indmarman.2021.07.014

Barki, H., & Huff, S. L. (1990). Implementing decision support systems: Correlates of user satisfaction and system usage. *INFOR: Information Systems and Operational Research*, *28*(2), 89–101. https://doi.org/10.1080/03155986.1990.11732123

Chorus, C. (2015). Choice behaviour modelling and the logit model.

Councyl. (2020). https://councyl.ai/en/councyl-makes-decision-support-available-for-your-organisations/

Cui, T. H., & Zhang, Y. (2018). Cognitive hierarchy in capacity allocation games. *Management Science*, *64*(3), 1250–1270. https://doi.org/10.1287/mnsc.2016.2655

de Haas, M. C. (2022). *Longitudinal studies in travel behaviour research* (Doctoral dissertation).

Gol, H., & Catay, B. (2007). Third-party logistics provider selection: Insights from a turkish automotive company. *Supply Chain Management*, *12*(6), 379–384. https://doi.org/10.1108/13598540710826290

Govindan, K., Khodaverdi, R., & Vafadarnikjoo, A. (2016). A grey dematel approach to develop third-party logistics provider selection criteria. *Industrial Management Data Systems*, *116*(4), 690–722. https://doi.org/10.1108/IMDS-05-2015-0180

Govindan, K., & Chaudhuri, A. (2016). Interrelationships of risks faced by third party logistics service providers: A dematel based approach. *Transportation Research Part E: Logistics and Transportation Review*, *90*, 177–195. https://doi.org/10.1016/j.tre.2015.11.010

Hofstra, N., & Spiliotopoulou, E. (2022). Behavior in rationing inventory across retail channels. *European Journal of Operational Research*, *299*(1), 208–222. https://doi.org/10.1016/j.ejor.2021.08.020

Isiklar, G., Alptekin, E., & Buyukozkan, G. (2007). Application of a hybrid intelligent decision support model in logistics outsourcing [Operations Research and Outsourcing]. *Computers Operations Research*, *34*(12), 3701–3714. https://doi.org/10.1016/j.cor.2006.01.011

Jharkharia, S., & Shankar, R. (2007). Selection of logistics service provider: An analytic network process (anp) approach. *Omega*, *35*(3), 274–289. https://doi.org/10.1016/j.omega.2005.06.005

Kaplan, A., & Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Karthikeyan, R., Venkatesan, K., & Chandrasekar, A. (2019). A comparison of strengths and weaknesses for analytical hierarchy process. *Journal of Chemical and Pharmaceutical Sciences*, *9*, S–12 S.

Knapp, F., Kessler, M., & Arlinghaus, J. C. (2021). The influence of cognitive biases in production logistics. In M. Freitag, H. Kotzab, & N. Megow (Eds.), *Dynamics in logistics: Twenty-five years of interdisciplinary logistics research in bremen, germany* (pp. 183–193). Springer International Publishing. https://doi.org/10.1007/978-3-030-88662-2_9

Li, Li, L., Jin, C., Wang, R., Wang, H., & Yang, L. (2012). A 3pl supplier selection model based on fuzzy sets. *Computers Operations Research*, *39*(8), 1879–1884. https://doi.org/10.1016/j.cor.2011.06.022

London, A. J. (n.d.). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/doi.org/10.1002/hast.973

Luk, C. C., Choy, K. L., & Lam, H. (2018). Design of an enhanced logistics service provider selection model for e-commerce application. *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*.

Meade, L., & Sarkis, J. (2002). A conceptual model for selecting and evaluating third-party reverse logistics providers. *Supply Chain Management*, *7*(5), 283–295. https://doi.org/10.1108/13598540210447728

Melacini, M., Perotti, S., Rasini, M., & Tappia, E. (2018). E-fulfilment and distribution in omni-channel retailing: A systematic literature review. *International Journal of Physical Distribution amp; Logistics Management*, *48*(4), 391–414. https://doi.org/10.1108/ijpdlm-02-2017-0101

Moerman, G. (2022). Typologies of interviews. https://www.coursera.org/lecture/qualitative-methods/4-2-typologies-of-interviews-zrBlF

Molin, E. (2015). Introduction to experimental designs.

Qureshi, M., Kumar, D., & Kumar, P. (2007). Modeling the logistics outsourcing relationship variables to enhance shippers' productivity and competitiveness in logistical supply chain. *International Journal of Productivity and Performance Management*, *56*(8), 689–714. https://doi.org/10.1108/17410400710833001

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57. https://doi.org/10.1016/j.omega.2014.11.009

Riopel, D., Langevin, A., & Campbell, J. F. (2005). The network of logistics decisions. In A. Langevin & D. Riopel (Eds.), *Logistics systems: Design and optimization* (pp. 1–38). Springer US. https://doi.org/10.1007/0-387-24977-X_1

Robinson, A. (2014). Reuters events: Supply chain amp; logistics business intelligence. https://www.reutersevents.com/supplychain/3pl/evolution-e-commerce-logistics

Rooderkerk, R. P., & Kök, A. G. (2019). Omnichannel assortment planning. In S. Gallino & A. Moreno (Eds.), *Operations in an omnichannel world* (pp. 51–86). Springer International Publishing. https://doi.org/10.1007/978-3-030-20119-7_4

Saaty, T. L., & Tran, L. T. (2007). On the invalidity of fuzzifying numerical judgments in the analytic hierarchy process. *Mathematical and Computer Modelling*, *46*(7), 962–975. https://doi.org/10.1016/j.mcm.2007.03.022

Singh, R., Gunasekaran, A., & Kumar, P. (2018). Third party logistics (3pl) selection for cold chain management: A fuzzy ahp and fuzzy topsis approach. *Annals of Operations Research*, *267*(1-2), 531–553. https://doi.org/10.1007/s10479-017-2591-3

Sintchenko, V., & Coiera, E. (2006). Decision complexity affects the extent and type of decision support use. *AMIA Annu Symp Proceedings Archive*, 724–728.

Sprague, R. H., & Carlson, E. D. (1982). Building effective decision support systems.

ten Broeke, A., Hulscher, J., Heyning, N., Kooi, E., & Chorus, C. (2021). Bait: A new medical decision support technology based on discrete choice theory. *Medical Decision Making*, *41*(5), 614–619. https://doi.org/10.1177/0272989X21100132

Thakkar, J., Deshmukh, S. G., Gupta, A. D., & Shankar, R. (2005). Selection of third-party logistics (3pl): A hybrid approach using interpretive structural modeling (ism) and analytic network process (anp). *Supply Chain Forum: An International Journal*, *6*(1), 32–46. https://doi.org/10.1080/16258312.2005.11517137

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Vakulenko, Y., Shams, P., Hellström, D., & Hjort, K. (2019). Service innovation in e-commerce last mile delivery: Mapping the e-customer journey. *Journal of Business Research*, *101*, 461–468. https://doi.org/10.1016/j.jbusres.2019.01.016

Zhü, K. (2014). Fuzzy analytic hierarchy process: Fallacy of the popular methods. *European Journal of Operational Research*, *236*(1), 209–217. https://doi.org/10.1016/j.ejor.2013.10.034