Cognitive Biases and Mitigation Strategies within Multi-Attribute Value Theory

Geqie SUN

Cognitive Biases and Mitigation Strategies within Multi-Attribute Value Theory

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof.dr.ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Wednesday, 13 May 2026 10:00

by

Geqie SUN

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr. J. Rezaei,	Delft University of Technology, promotor
Dr.ir. M. Kroesen,	Delft University of Technology, promotor

Independent Members:

Prof.dr. S. Roeser,	Delft University of Technology, The Netherlands
Prof.dr. M, Abdellaoui,	École des hautes études commerciales de Paris, France
Dr. B. Fasolo,	The London School of Economics and Political Science, United Kingdom
Dr. S.A. Kusumastuti,	University of Twente, The Netherlands
Prof.dr.ir. I.R. van de Poel,	Delft University of Technology, The Netherlands, *reserve member*

Keywords: Cognitive bias; Bias mitigation; Multi-criteria decision making; Behavioral decision making; Multi-attribute value theory
Printed by: Haveka
Cover by: Geqie SUN

ISBN: 978-94-6518-306-0

An electronic version of this dissertation is available at: http://repository.tudelft.nl

"Cried, but did it."

Dedicated to those who never give up.

Acknowledgment

Before I began my PhD journey in the Netherlands, I had many imaginations of what this experience might be like. None of them, however, came close to the richness and complexity of what I have lived through over the past four years. This journey has been filled with excitement and nervousness, anxiety and fulfillment, tears and fears, and many emotions that are difficult to describe. All of these experiences have shaped me, both academically and personally, and have ultimately brought me to this moment.

First and foremost, I would like to express my sincere gratitude to my promotor, Prof.dr. Jafar Rezaei, for his continuous guidance, support, and trust throughout the process. The beginning of my PhD was a particularly difficult period. I doubted my abilities and, at times, almost lost hope. During this challenging phase, you supported me patiently and consistently, helping me step by step and guiding me through moments when I felt completely stuck. Throughout this journey, I have learned an enormous amount from you, not only about research and scientific writing, but also about how to think critically and independently. Beyond academic guidance, you have always shown genuine care for my well-being, for which I am deeply grateful. In the darkest moments of my research, your support meant more to me than words can express. I sincerely appreciate your wisdom, patience, humor, knowledge, and kindness, all of which have played a crucial role in shaping my development as an independent researcher.

I would also like to sincerely thank my other promotor, Dr.ir. Maarten Kroesen. You are the calmest and most relaxed member of the supervisory team, and your attitude has had a positive and lasting influence on my PhD journey. You have always been supportive and approachable, which made it natural and productive to discuss ideas and doubts with you. You consistently raised critical questions about my research topics and concepts, prompting me to reflect more deeply and think carefully about complex issues. I am deeply grateful for the guidance and constructive feedback you provided throughout this research, which helped sharpen my arguments and improve the clarity of my work. I greatly enjoyed our discussions and conversations, and I also appreciate your genuine care for my well-being and your willingness to offer support whenever needed. I truly value the reassuring and positive environment you created during my PhD.

I would like to express my heartfelt thanks to Pelin Gülüm, my dearest research partner. We have shared so many moments throughout this PhD journey that I can hardly imagine going through it without you. We discussed research and life, supported each other through challenges, and celebrated achievements together. Your companionship and support meant more to me than I can express, and I sincerely value our friendship. I am also very grateful to Qiuju Xue. It is truly special that we share so much in common in our backgrounds, which made our connection feel natural from the very beginning. I greatly enjoyed all the conversations we

had about research, relationships, family, and life in general. Those exchanges brought both comfort and joy during my PhD journey. I am grateful that you two can be my paranymphs for my PhD defense.

I would also like to express my gratitude to my colleagues for the cheerful, supportive, and inspiring environment we created and shared: Kailas, Arina, Laura, Mahsa, Francisco, Su Li, Somayeh, Patrick, Julien, Erik, Gabriel, Nicoel, Liubov, Joslyn, Ali, Mohammad, Lukas, Ingrid, Jessie, Sugandha, Thaddaeus, Steven, Esra, Joseph, Jerico, Palok, Nely, Arnoud, Lion, Hong Yan, Kailan Wu, Shahrzad, Mahendra, Benjamin, Mengying Zhou, Qun Wu, Fei Wu, Cherelle, Nicolas, Kateřina, Jan Anne, Sander, Yousef, Eric, Niek, and Baiba. Thank you for your kindness, collaboration, and the positive atmosphere that made working together such a pleasure.

My heartfelt gratitude goes to my dear friends, Sijia Kong and Yiyuan Zou. We have spent countless days and nights together in Delft, and shared many moments of laughter and tears. Your companionship made the difficult times lighter and the joyful moments even more meaningful. I am deeply grateful for your friendship and unwavering support throughout this journey. I am equally grateful to all the wonderful friends I met during my time in the Netherlands: Yuefan Pan, Sihan Wang, Sidi Liu, Tianyi Deng, Lucia, Zhengang Xia, Yuwei Huang, Jiaxing Fang, Zhenjie Wang, Pengtao Gao, Guangze Qin, Jiarui Zhang, Begum, Zeynep, Tugba, Yuri, Jessie Zhang, Hongyan Wu, Yifei Li, Xiangyu Yang, Hanyu Hu, and Lan Yan. Thank you all for the warmth, laughter, and countless joyful memories we created together, which made life abroad truly special. I would also like to thank my boyfriend, Tijmen. I am grateful for your love, care, and sincerity. I have truly cherished the time we have spent together, through both light-hearted moments and more challenging times. Your kindness, patience, and steady support have been a source of strength for me.

I would also like to thank my friends from China, who have always been supportive and caring, no matter the distance. Wenxin Zhang, you have always been there for me, and I know I can trust you. Thank you for patiently listening to all my worries and for celebrating my achievements with me. Le Du, Ziwei Ren, Xi Yu, Boyu Wan, and Chenglei Diao, we have been the best team for many years, and I am grateful for all the wonderful memories we created together. Beijing will become a part of my life because of you. Jiexin Du and Yuxin Tian, we have known each other since childhood, and I sincerely thank you for your long-lasting friendship, understanding, and support throughout the different stages of my life. Linshuang Tu, my dear friend, thank you for your enduring friendship, and I wish you all the very best as you complete your PhD.

Last but not least, I would like to thank my family, without whom this PhD would not have been possible. To my mother, Suzhen Xiang, and my father, Baohua Sun, I am deeply grateful for your love, support, and understanding. Knowing that I can always rely on you has given me strength throughout this journey. I have inherited and learned so much from you, and these values will always be a part of who I am. To my sister, Xiaoyan Tang, I am deeply grateful for your love and care. You have always been a constant source of support and encouragement. I am truly thankful to have you as my sister.

Geqie Sun
Delft, December 2025

Contents

Acknowledgment **vii**
Summary **xi**
Samenvatting **xv**
Summary in Chinese **xix**
1 Introduction **1**
1.1 Background . . . 1
1.2 Research Gaps . . . 3
1.3 Research Focus . . . 4
1.4 Research Questions . . . 6
1.5 Dissertation Outline . . . 8
2 Anchoring Bias in Value Function Elicitation within Multi-Attribute Value Theory **15**
2.1 Introduction . . . 16
2.2 Anchoring Bias . . . 18
2.3 Multi-Attribute Value Theory . . . 20
2.4 Hypotheses Development . . . 22
2.5 Experiment Design . . . 25
2.5.1 The Experiment Overview . . . 25
2.5.2 Participants . . . 29
2.6 Results and Discussion . . . 31
2.6.1 Hypothesis 1: Impact of Low vs. High Anchors on Midvalue Points and Value Function Shape . . . 31
2.6.2 Hypothesis 2: Effectiveness of Debiasing Strategies . . . 33
2.6.3 Hypothesis 3: Impact of Anchoring on the Overall Value of Alternatives 34
2.6.4 Implications for Similar Procedures . . . 37
2.7 Conclusion . . . 38
3 Anchoring Bias in the Tradeoff Procedure within Multi-Attribute Value Theory **45**
3.1 Introduction . . . 46
3.2 The Anchoring Bias and its Role in Multi-Attribute Decision Making . . . 48
3.3 An Overview of the Tradeoff Procedure . . . 51
3.3.1 Multi-Attribute Value Theory . . . 51
3.3.2 Tradeoff Procedure . . . 53
3.3.3 Best-Worst Tradeoff Method . . . 55
3.4 Hypotheses Development . . . 57
3.5 Experimental Design . . . 61
3.6 Results and Discussion . . . 64
3.7 Conclusion . . . 68
4 Framing and Loss Aversion in Decisions with Multiple Objectives **75**
4.1 Introduction . . . 76
4.2 Gain-Loss Bias and its Role in Multi-Attribute Decision-Making Methods . . . 77
4.3 Multi-Attribute Value Theory . . . 80

4.4 Hypotheses Development . . . 82
4.4.1 Loss Aversion in the Tradeoff Procedure . . . 82
4.4.2 Framing Effect on Attribute-Specific Value Functions . . . 85
4.4.3 Framing Effect on Weight . . . 87
4.5 Experiment Design . . . 89
4.6 Results and Discussion . . . 91
4.6.1 The Loss Aversion within the Tradeoff Procedure . . . 91
4.6.2 The Framing Effect on Attribute-Specific Value Function . . . 93
4.6.3 The Framing Effect on Weight . . . 96
4.7 Conclusion . . . 97
5 The Mitigation Role of Multi-Attribute Value Theory on Status Quo Bias 103
5.1 Introduction . . . 104
5.2 Theoretical Background and Hypotheses . . . 105
5.3 Multi-Attribute Value Theory . . . 107
5.4 Experiment Design . . . 109
5.5 Results and Discussion . . . 110
5.5.1 Status Quo Bias in Ranking . . . 111
5.5.2 Status Quo Bias in MAVT . . . 113
5.5.3 Status Quo Bias in Weights . . . 115
5.5.4 Status Quo Bias in Attribute-Specific Value Functions . . . 116
5.5.5 The Mechanism of Status Quo Bias in MAVT . . . 117
5.6 Conclusion . . . 119
6 Conclusion 125
6.1 Key Findings . . . 126
6.2 Theoretical and Practical Implications . . . 130
6.3 Limitations and Future Research . . . 132
Appendix 137
A Appendix - Questionnaire design for Chapter 2 . . . 138
About the author 141

Summary

Decision-making permeates everyday life, yet the consequences of poor judgment become particularly severe in important or high-stakes contexts. Many such decisions take the form of multi-attribute decision-making (MADM) problems, where decision-makers must evaluate alternatives described by multiple, often conflicting attributes. MADM methods offer a normative, systematic, and structured framework to support these decisions by decomposing the problem into a sequence of elicitation tasks. Through structured procedures, they elicit attribute-specific preferences, assess tradeoffs, aggregate intermediate judgments, and produce a solution that is optimal under the method's axiomatic foundations.

However, research in psychology has identified numerous cognitive biases that can lead to errors in human judgment when individuals process and interpret information. Cognitive biases occur due to the use of heuristics to reduce task complexity. Such usage is usually effective in arriving at a good enough rather than an optimal solution. Since human judgments are critical inputs of MADM methods, biases can emerge and accumulate across multiple intermediate preference elicitation stages, which can significantly undermine the quality of the final decision. A poor decision in critical areas can result in financial losses, wasted time, misallocated resources, and a significant social impact. Ensuring that MADM methods are not only normatively sound but also behaviorally robust, therefore requires a careful examination of how cognitive biases influence each stage of the elicitation process.

Within behavioral decision research (BDR), researchers have identified numerous biases relevant to decision analysis and distinguished between those that are relatively easy to correct and those deeply embedded in cognition. Yet empirical evidence on how these biases manifest within MADM procedures remains limited. This dissertation addresses this gap by investigating four cognitive biases, anchoring bias, framing effect, loss aversion, and status quo bias, within multi-attribute value theory (MAVT), one of the most widely applied MADM methods. These biases were selected because they arise from different stages of the decision process, are theoretically connected to reference dependence, and thus collectively provide a coherent basis for examination. MAVT structures preference elicitation into two main components: the elicitation of attribute-specific value functions and the elicitation of tradeoff weights. This decomposition offers a transparent and quantitative framework for analyzing where and how biases emerge and for developing more targeted debiasing strategies.

Chapter 1 introduces the research background, identifies the gap in existing literature, formulates the main research questions, and outlines the dissertation structure. It develops three sub-research questions. RQ1 focuses on the effect of anchoring bias within MAVT and develops debiasing strategies. RQ2 investigates the interaction of framing effect and loss aversion across multiple stages of MAVT and explores possible debiasing strategies. RQ3 explores the

mitigation potential of MAVT in reducing status quo bias, which prior BDR research has classified as easy to correct within a decision analysis method. Investigating the three sub-research questions together motivates four empirical studies.

Chapter 2 investigating anchoring bias in the attribute-specific value function elicitation step of MAVT. Anchoring bias refers to the tendency to make insufficient adjustment from an anchor. In Chapter 2, it hypothesizes that the use of starting points in the midvalue splitting procedure can lead to biased attribute-specific value functions and ultimately decision results. It also develops two debiasing strategies, no-anchor and counter-anchor, integrated directly into the midvalue splitting procedure. Although this study focuses on the midvalue splitting procedure, the insights generalize to other value function elicitation methods (e.g., the lock-step procedure, the standard difference procedure, and direct rating). These methods similarly rely on analyst-provided starting points to guide the identification or elicitation of intermediate values and employ iterative steps that can carry and amplify anchoring effects throughout the construction of the value function.

Chapter 3 focuses on anchoring bias in the tradeoff procedure within MAVT. In the traditional tradeoff procedure, the weights are calculated by quantifying the tradeoff a decision-maker is willing to make between the most important attribute and each of the rest attributes. Therefore, the most important attribute is used repeatedly to construct indifference pairs for eliciting weights, which may serve as an implicit anchor. Chapter 3 hypothesizes that this anchoring leads to inconsistencies in the elicited weights across conditions. It also hypothesizes that the Best-Worst Tradeoff method is effective in reducing such bias due to the use of both the most and least important attributes in the procedure. This study highlights the identification and examination of implicit anchors embedded within the elicitation procedures of MADM methods.

Chapter 4 extends the investigation from single biases to examining the combined effect of framing effect and loss aversion across multiple stages of MAVT. It hypothesizes that (i) framing the same attribute in gains or losses can lead to different attribute-specific value functions; (ii) loss aversion in the tradeoff procedure can lead to different weights for the same attribute; (iii) the joint influence of framing and loss aversion can reinforce distortions in the final weights. The study also evaluates several mitigation strategies, such as homogeneous framing, using both gain- and loss-framed tradeoff procedures, and group decision-making. It demonstrates the importance of examining multiple biases simultaneously and illustrates how certain biases can naturally counteract each other, thereby reducing distortions in decision outcomes.

Chapter 5 turns to the mitigation potential of the method itself by examining whether MAVT can reduce status quo bias. While previous studies have focused on how cognitive bias can distort intermediate judgments and the decision outcomes of MAVT, they also offer important prescriptive insights into how to address them within the structured elicitation procedures of MAVT. This study shows that although the status quo bias influences the elicited weights, the structured procedures of MAVT reduce the bias significantly in the final decision outcome. It further demonstrates that MADM methods can be used to mitigate biases.

Chapter 6 concludes the dissertation by summarizing the key findings of each study and answering the research questions. It also discusses several theoretical and practical implications for the BDR and MADM research fields. Limitations and future research directions regarding the experiment design, multiple cognitive bias interactions, and debiasing strategies, as well as

the behavioral robustness of all MADM methods, have been identified.

In conclusion, investigating cognitive biases within MADM, and particularly within the MAVT, provides deeper insight into the decision-making process because these methods decompose complex decisions into simpler elicitation steps. This structure makes it possible to identify where and how specific biases emerge, and to observe whether their effects are carried through, amplified, or attenuated in the final decision. This dissertation presents comprehensive empirical evidence on how anchoring bias, framing effects, loss aversion, and status quo bias influence MAVT across different stages of preference elicitation and in decision outcomes. At the same time, the findings highlight a fundamental duality of MADM methods: while the structured procedures of MADM methods can introduce or reinforce cognitive biases, they also create opportunities to mitigate them. When the sources and mechanisms of bias are understood, the same structure that exposes decision-makers to bias can be redesigned or leveraged to reduce it, thereby enhancing the behavioral robustness of MADM methods.

Samenvatting

Besluitvorming is verweven met het dagelijks leven, maar de gevolgen van een verkeerde inschatting zijn bijzonder ernstig in belangrijke of hoog-risicovolle contexten. Veel van dergelijke beslissingen nemen de vorm aan van multi-attribuut besluitvorming (MADM) vraagstukken, waarbij besluitvormers alternatieven moeten evalueren die beschreven worden door meerdere, vaak tegenstrijdige attributen. MADM-methoden bieden een normatief, systematisch en gestructureerd kader om deze beslissingen te ondersteunen door het probleem te ontleden in een reeks elicitatietaken. Middels gestructureerde procedures vragen zij attribuut-specifieke voorkeuren uit, beoordelen zij afwegingen, aggregeren zij tussentijdse oordelen en produceren zij een oplossing die optimaal is volgens de axiomatische grondslagen van de methode.

Onderzoek in de psychologie heeft echter talrijke cognitieve biases geïdentificeerd die tot fouten in het menselijk oordeel kunnen leiden wanneer individuen informatie verwerken en interpreteren. Cognitieve biases ontstaan door het gebruik van heuristieken om de taakcomplexiteit te verminderen. Het gebruik van heuristieken is doorgaans effectief om tot een oplossing te komen die 'goed genoeg' is, in plaats van een optimale oplossing. Aangezien menselijke oordelen cruciale input vormen voor MADM-methoden, kunnen biases ontstaan en zich opstapelen gedurende meerdere tussentijdse stadia van voorkeurselicitatie, wat de kwaliteit van de uiteindelijke beslissing aanzienlijk kan ondermijnen. Een slechte beslissing op kritieke gebieden kan leiden tot financiële verliezen, verspilde tijd, verkeerd toegewezen middelen en een aanzienlijke maatschappelijke impact. Het waarborgen dat MADM-methoden niet alleen normatief deugdelijk maar ook gedragsmatig robuust zijn, vereist daarom een zorgvuldig onderzoek naar hoe cognitieve biases elke fase van het elicitatieproces beïnvloeden.

Binnen het gedragswetenschappelijk beslissingsonderzoek (BDR) hebben onderzoekers talrijke biases geïdentificeerd die relevant zijn voor beslissingsanalyse en een onderscheid gemaakt tussen biases die relatief eenvoudig te corrigeren zijn en biases die diep geworteld zijn in de cognitie. Toch blijft empirisch bewijs over hoe deze biases zich manifesteren binnen MADM-procedures beperkt. Dit proefschrift adresseert dit hiaat door vier cognitieve biases te onderzoeken: het ankereffect (anchoring bias), het framingeffect, verliesaversie (loss aversion) en de status quo bias, binnen de multi-attribuut waardetheorie (MAVT), een van de meest toegepaste MADM-methoden. Deze biases zijn geselecteerd omdat zij voortkomen uit verschillende stadia van het besluitvormingsproces, theoretisch verbonden zijn met referentieafhankelijkheid en zodoende gezamenlijk een coherente basis voor onderzoek vormen. MAVT structureert voorkeurselicitatie in twee hoofdcomponenten: de elicitatie van attribuut-specifieke waardefuncties en de elicitatie van afwegingsgewichten. Deze decompositie biedt een transparant en kwantitatief kader voor het analyseren van waar en hoe biases ontstaan en voor het ontwikkelen van meer gerichte debiasing-strategieën.

Hoofdstuk 1 introduceert de onderzoeksachtergrond, identificeert het gat in de bestaande literatuur, formuleert de hoofdonderzoeksvragen en beschrijft de structuur van het proefschrift. Het werkt drie deelonderzoeksvragen uit. De eerste onderzoeksvraag richt zich op het effect van anchoring bias binnen MAVT en ontwikkelt debiasing-strategieën. De tweede onderzoeksvraag onderzoekt de interactie van het framingeffect en verliesaversie gedurende meerdere fasen van MAVT en verkent mogelijke debiasing-strategieën. De derde onderzoeksvraag verkent het potentieel van MAVT om status quo bias te verminderen, die door eerder BDR-onderzoek is geclassificeerd als eenvoudig te corrigeren binnen een beslissingsanalysemethode. Het gezamenlijk onderzoeken van de drie deelonderzoeksvragen vormt de motivatie voor vier empirische studies.

Hoofdstuk 2 onderzoekt anchoring bias in de stap van de elicitatie van attribuut-specifieke waardefuncties van MAVT. Anchoring bias verwijst naar de neiging om onvoldoende aanpassing te doen ten opzichte van een anker. In Hoofdstuk 2 wordt de hypothese gesteld dat het gebruik van startpunten in de midvalue splitting-procedure kan leiden tot vertekende attribuut-specifieke waardefuncties en uiteindelijk tot vertekende beslissingsuitkomsten. Tevens worden twee debiasing-strategieën ontwikkeld, no-anchor en counter-anchor, die direct in de midvalue splitting-procedure worden geïntegreerd. Hoewel deze studie zich richt op de midvalue splitting-procedure, zijn de inzichten generaliseerbaar naar andere methoden voor waardefunctie-elicitatie (bijv. de lock-step-procedure, de standard difference-procedure en direct rating). Deze methoden vertrouwen eveneens op door de analist aangedragen startpunten om de identificatie of elicitatie van tussenliggende waarden te sturen en maken gebruik van iteratieve stappen die anchoring-effecten kunnen meedragen en versterken gedurende de opbouw van de waardefunctie.

Hoofdstuk 3 richt zich op anchoring bias in de trade-off procedure binnen MAVT. In de traditionele trade-off procedure worden de gewichten berekend door de trade-off te kwantificeren die een besluitvormer bereid is te maken tussen het belangrijkste attribuut en elk van de overige attributen. Daarom wordt het belangrijkste attribuut herhaaldelijk gebruikt om indifferentieparen te construeren voor het eliciteren van gewichten, wat kan fungeren als een impliciet anker. Hoofdstuk 3 veronderstelt dat deze anchoring leidt tot inconsistenties in de geëliciteerde gewichten over verschillende condities heen. Tevens wordt verondersteld dat de Best-Worst Trade-off methode effectief is in het verminderen van dergelijke bias, vanwege het gebruik van zowel de belangrijkste als de minst belangrijke attributen in de procedure. Deze studie belicht de identificatie en het onderzoek naar impliciete ankers die ingebed zijn in de elicitatieprocedures van MADM-methoden.

Hoofdstuk 4 breidt het onderzoek uit van enkelvoudige biases naar het onderzoeken van het gecombineerde effect van het framingeffect en verliesaversie gedurende meerdere fasen van MAVT. Het veronderstelt dat (i) het framen van hetzelfde attribuut in termen van winst of verlies kan leiden tot verschillende attribuut-specifieke waardefuncties; (ii) verliesaversie in de trade-off procedure kan leiden tot verschillende gewichten voor hetzelfde attribuut; (iii) de gezamenlijke invloed van framing en verliesaversie vertekeningen in de uiteindelijke gewichten kan versterken. De studie evalueert tevens diverse mitigatiestrategieën, zoals homogene framing, het gebruik van zowel gain- als loss-framed trade-off procedures en groepsbesluitvorming. Het toont het belang aan van het gelijktijdig onderzoeken van meerdere biases en illustreert hoe bepaalde biases elkaar op natuurlijke wijze kunnen compenseren, waardoor vertekeningen in beslissingsuitkomsten worden verminderd.

Hoofdstuk 5 richt zich op het mitigatiepotentieel van de methode zelf door te onderzoeken of MAVT status quo bias kan verminderen. Hoewel eerdere studies zich hebben gericht op hoe cognitieve bias tussentijdse oordelen en de beslissingsuitkomsten van MAVT kan vertekenen, bieden zij ook belangrijke prescriptieve inzichten in hoe deze binnen de gestructureerde elicitatieprocedures van MAVT kunnen worden aangepakt. Deze studie toont aan dat, hoewel de status quo bias de geëliciteerde gewichten beïnvloedt, de gestructureerde procedures van MAVT de bias in de uiteindelijke beslissingsuitkomst aanzienlijk verminderen. Tevens wordt aangetoond dat MADM-methoden kunnen worden ingezet om biases te mitigeren.

Hoofdstuk 6 besluit het proefschrift door de belangrijkste bevindingen van elke studie samen te vatten en de onderzoeksvragen te beantwoorden. Het bespreekt tevens verscheidene theoretische en praktische implicaties voor de onderzoeksgebieden BDR en MADM. Beperkingen en toekomstige onderzoeksrichtingen met betrekking tot de experimentopzet, interacties tussen meerdere cognitieve biases en debiasing-strategieën, evenals de gedragsmatige robuustheid van alle MADM-methoden, zijn geïdentificeerd.

Concluderend biedt het onderzoeken van cognitieve biases binnen MADM, en in het bijzonder binnen de MAVT, dieper inzicht in het besluitvormingsproces, omdat deze methoden complexe beslissingen ontleden in eenvoudigere elicitatiestappen. Deze structuur maakt het mogelijk te identificeren waar en hoe specifieke biases ontstaan, en waar te nemen of hun effecten worden doorgegeven, versterkt of afgezwakt in de uiteindelijke beslissing. Dit proefschrift presenteert uitgebreid empirisch bewijs over hoe anchoring bias, framingeffecten, verliesaversie en status quo bias MAVT beïnvloeden gedurende verschillende stadia van voorkeurselicitatie en in beslissingsuitkomsten. Tegelijkertijd belichten de bevindingen een fundamentele dualiteit van MADM-methoden: hoewel de gestructureerde procedures van MADM-methoden cognitieve biases kunnen introduceren of versterken, creëren ze ook mogelijkheden om deze te mitigeren. Wanneer de bronnen en mechanismen van bias worden begrepen, kan dezelfde structuur die besluitvormers blootstelt aan bias worden herontworpen of benut om deze te verminderen, waardoor de gedragsmatige robuustheid van MADM-methoden wordt vergroot.

摘要

决策贯穿于日常生活之中，但在重要或高风险情境下，判断失误所带来的后果尤为严重。许多此类决策表现为多属性决策（multi-attribute decision-making，MADM）问题，即决策者需要在多个且往往相互冲突的属性维度上对备选方案进行评估。MADM 方法通过将复杂决策问题分解为一系列用于引出偏好的任务，提供了一种规范性、系统性且结构化的框架来支持决策。借助结构化的程序，这些方法能够引出针对特定属性的偏好，评估不同属性之间的权衡关系，对中间判断进行整合，并在其公理化基础之上生成一个最优解。

然而，心理学研究已经识别出大量认知偏差，这些偏差会在个体加工和解读信息时导致系统性的判断错误。认知偏差源于人们为降低任务复杂度而采用的启发式策略，而这种策略通常只能帮助个体获得“足够好”的解，而非最优解。由于人类判断构成了MADM 方法中的关键输入，认知偏差可能在多个偏好引出的中间阶段产生并逐步累积，从而显著削弱最终决策的质量。在关键领域中做出不当决策，可能导致经济损失、时间浪费、资源错配，甚至引发重大的社会影响。因此，要确保MADM 方法不仅在规范性层面上是合理的，而且在行为层面上具有稳健性，就有必要对认知偏差如何影响偏好引出过程各个阶段进行细致而系统的考察。

在行为决策研究（behavioral decision research，BDR）领域，研究者已经识别出大量与决策分析相关的认知偏差，并区分了其中相对容易纠正的偏差与深植于于认知过程中的偏差。然而，关于这些偏差如何在MADM程序中具体表现的经验证据仍然有限。本论文通过在多属性价值理论（multi-attribute value theory，MAVT）这一应用最为广泛的MADM 方法框架下，系统考察四种认知偏差——锚定偏差（anchoring bias）、框架效应（framing effect）、损失厌恶（loss aversion）和现状偏差（status quo bias）——以弥补这一研究空缺。之所以选择这些偏差，是因为它们产生于决策过程的不同阶段，在理论上均与参照依赖（reference dependence）密切相关，因此能够共同构成一个连贯的分析基础。MAVT 将偏好引出结构化为两个主要组成部分：属性特定价值函数的引出以及权衡权重的引出。这种分解方式为分析认知偏差在何处以及以何种方式产生提供了一个透明且可定量分析的框架，也为开发更具针对性的去偏策略奠定了基础。

第一章介绍了研究背景，识别了现有文献中的研究空缺，提出了主要研究问题，并概述了论文结构。本章提出了三个子研究问题。研究问题一（RQ1）关注锚定偏差在MAVT 中的作用，并探讨相应的去偏策略。研究问题二（RQ2）考察框架效应与损失厌恶在MAVT 多个阶段中的交互作用，并探索可能的去偏方法。研究问题三（RQ3）探讨MAVT 在缓解现状偏差方面的潜力，由于既有BDR研究将现状偏差归类为在决策分析方法中相对容易纠正的一类偏差。对三个子研究问题的综合考察促成了四项实证研究的开展。

第二章研究了MAVT 中属性特定价值函数引出阶段的锚定偏差。锚定偏差是指个体在判断过程中从初始锚点出发进行调整不足的倾向。本章假设，在中值分割（midvalue

splitting）程序中使用起始点，会导致属性特定价值函数出现系统性偏差，并最终影响决策结果。与此同时，本章还提出并检验了两种去偏策略, 无锚策略（no-anchor）和反向锚策略（counter-anchor），并将其直接整合进中值分割程序中。尽管本研究聚焦于中值分割程序，其所得洞见同样可以推广至其他价值函数引出方法（如锁步程序、标准差分程序和直接评分法）。这些方法同样依赖分析者提供的起始点来引导中间值的识别或引出，并采用迭代步骤，这些步骤可能在价值函数构建过程中传递并放大锚定效应。

第三章关注MAVT 中权衡程序（tradeoff procedure）里的锚定偏差。在传统的权衡程序中，权重是通过量化决策者在最重要属性与其余各属性之间愿意进行的权衡来计算的。因此，最重要属性在构建无差异配对以引出权重的过程中被反复使用，从而可能充当一种隐性锚点。本章假设，这种锚定会导致在不同条件下所引出的权重存在不一致性。同时，本章还假设，由于在程序中同时使用最重要属性和最不重要属性，最佳–最差权衡（Best-Worst Tradeoff）方法能够有效降低此类偏差。本研究强调了识别并系统检验嵌入于MADM 方法引出程序中的隐性锚点的重要性。

第四章将研究视角从单一认知偏差拓展至MAVT多个阶段中框架效应与损失厌恶的联合作用。本章提出以下假设：（i）对同一属性采用收益框架或损失框架进行表述，会导致不同的属性特定价值函数；（ii）在权衡程序中，损失厌恶会使同一属性对应的权重产生差异；（iii）框架效应与损失厌恶的共同作用会在最终权重中相互强化，从而加剧决策结果的扭曲。研究还评估了多种去偏或缓解策略，包括统一框架表述、同时采用收益框架与损失框架的权衡程序，以及群体决策。研究结果表明，同时考察多种认知偏差具有重要意义，并揭示了某些偏差在特定条件下可能相互抵消，从而降低对决策结果的扭曲程度。

第五章转而考察方法本身的去偏潜力，重点分析MAVT 是否能够缓解现状偏差。尽管既有研究主要关注认知偏差如何扭曲MAVT 中的中间判断和决策结果，但这些研究同样为如何在MAVT 的结构化偏好引出程序中应对认知偏差提供了重要的规范性启示。本章研究表明，尽管现状偏差会影响权重的引出结果，但MAVT 的结构化程序能够在最终决策结果中显著降低该偏差的影响。研究进一步表明，MADM方法本身也可以被用于缓解认知偏差。

第六章作为本论文的结尾，总结了各项研究的主要发现，并回答了研究问题。该章还探讨了这些发现对BDR和MADM研究领域的若干理论与实践意义。文中指出了在实验设计、多种认知偏见交互作用、去偏见策略以及所有MADM方法的行为稳健性等方面存在的局限性及未来研究方向。

总而言之，在MADM，尤其是MAVT 框架下考察认知偏差，有助于深化对决策过程的理解，因为这些方法能够将复杂决策分解为若干更为简化的偏好引出步骤。这种结构化特征使研究者能够识别特定认知偏差在何处、以何种方式产生，并观察其影响在最终决策中是被传递、放大，还是被削弱。本论文系统地提供了锚定偏差、框架效应、损失厌恶与现状偏差在MAVT 不同偏好引出阶段及决策结果中作用机制的实证证据。同时，研究结果凸显了MADM 方法所具有的一种根本性二重性：一方面，其结构化程序可能引入或强化认知偏差；另一方面，这种结构本身也为缓解偏差创造了条件。当认知偏差的来源与作用机制得到充分理解后，原本可能使决策者暴露于偏差之中的结构，也可以被重新设计或加以利用，以降低偏差影响，从而提升MADM 方法的行为稳健性。

Chapter 1

Introduction

1.1 Background

Decision-making plays a central role in numerous real-world contexts, ranging from everyday choices to high-stakes decisions in the corporate and public sectors. Whereas simple decisions can often be made intuitively without major consequences, poor decisions in critical contexts, such as investment decisions, strategic corporate decisions, or public policy making, can lead to substantial financial losses, wasted resources, and wide-ranging societal impacts. Because such decisions typically involve multiple criteria, heterogeneous stakeholders, and complex tradeoffs, unaided intuitive judgment is often insufficient. As a result, decision analysis methods are widely employed to support more systematic, transparent, and informed decision-making.

Multi-criteria decision-making (MCDM) problems refer to decision problems described by multiple, often conflicting criteria. To support decision-makers addressing such problems, the field of multi-criteria decision analysis (MCDA) develops formal methods for structuring decision problems, eliciting preferences, and evaluating alternatives in a systematic manner. MCDM can be broadly classified as (Hwang & Yoon, 1981; Tzeng & Huang, 2011; Zanakis et al., 1998): (i) multi-attribute decision-making (MADM), which concerns evaluating a finite set of alternatives on multiple, potentially conflicting attributes; and (ii) multi-objective decision-making (MODM), which typically involves optimizing continuous objectives under constraints. This dissertation focuses on MADM.

MADM methods provide a normative framework for representing and aggregating decision-makers' preferences (Belton & Stewart, 2012; Figueira et al., 2005; Howard, 2007). Their mathematical foundations ensure that, under their respective axiomatic assumptions, the decision outcomes are normatively optimal (Fishburn, 1970, 1989). In practice, the application of MADM methods involves a series of elicitation procedures, such as attribute identification, value function elicitation, and tradeoff assessment, that decompose the holistic judgment task into simpler components. A decision analyst typically guides the decision-maker through these steps using clear, non-technical language (Keeney, 1977). This decomposed structure encourages reflection on attribute-level evaluations, makes tradeoffs explicit, and allows the decision-maker to examine how their preferences are translated into an overall ranking. As a result, MADM enhances process transparency, improves the justification of decisions, and supports more constructive engagement among stakeholders with diverse perspectives (Keeney, 1982;

French, 1989; Roy & Vanderpooten, 1996).

Despite the normative rigor of MADM methods, their performance depends critically on the quality of the judgment provided by human decision-makers. Decision analysis traditionally assumes that individuals can offer stable, unbiased, and coherent judgments (Edwards, 1954; Shafir & LeBoeuf, 2002; Simon, 1978). However, extensive empirical research in psychology and behavioral decision research (BDR) shows that this assumption is frequently violated (Tversky & Kahneman, 1973, 1989, 1974; Kahneman et al., 1979; Slovic et al., 1977; Einhorn & Hogarth, 1981; Gilovich, 1981; Dunning et al., 1990; Hogarth, 1981; Cohen, 1981). Human decision-makers operate under bounded rationality, meaning their cognitive resources, attention, and information-processing capabilities are limited relative to the complexity of the decision task (Simon, 1955; Jones, 1999; Miller, 1956). To cope with complexity, individuals often rely on heuristics (Tversky & Kahneman, 1974), which are simplifying strategies people use to reduce cognitive effort. These strategies are typically based on satisfying rather than maximizing utilities; therefore, they can lead to cognitive bias and flawed judgments (Kahneman, 2011; Griffin et al., 2001; Peer & Gamliel, 2013). Cognitive bias refers to the systematic error in judgment that occurs when individuals process and interpret information. Over the past few decades, hundreds of cognitive biases have been identified and have been shown to be pervasive across laboratory and field studies (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1981, 1991, 1983; Kahneman et al., 1979; Kahneman, 2011). As human judgments are the significant input for decision analysis methods, a flawed judgment can lead to biased, sub-optimal decision results.

The pervasiveness of cognitive biases has made them a central topic in BDR (Einhorn & Hogarth, 1981; Payne et al., 1992; Fischhoff & Broomell, 2020; Federspiel et al., 2024; von Winterfeldt, 1999), which encompasses both descriptive and prescriptive perspectives (Montibeller & von Winterfeldt, 2024). Descriptive BDR aims to develop theories or models that describe biases in judgments. Prescriptive BDR focuses on improving judgments and decision quality, typically by developing debiasing strategies or designing elicitation procedures that help individuals adhere more closely to normative standards (Milkman et al., 2009). In this prescriptive context, cognitive bias is defined as "a systematic discrepancy between the 'correct' answer in a judgmental task, given by a formal normative rule, and the decision-maker's actual answer" (Montibeller & von Winterfeldt, 2015, pp. 1231). This definition underscores the central role of normative decision analysis models in identifying and evaluating bias.

A prevailing view within decision analysis is that the structured and systematic nature of decision analysis methods should, in principle, mitigate biases by promoting more deliberate thinking (Keeney, 2004). However, empirical evidence indicates that many cognitive biases persist even when decision-makers follow formal elicitation procedures (Montibeller & von Winterfeldt, 2015). For instance, cognitive biases such as anchoring bias, equalizing bias, splitting bias can significantly impact the results of decision analysis methods (Borcherding & von Winterfeldt, 1988; Weber & Borcherding, 1993; Fischer, 1995; Fischer et al., 1987; Jacobi & Hobbs, 2007; Rezaei, 2021; Rezaei et al., 2022, 2024; Hammond et al., 1998; Bond et al., 2008). These findings suggest that the behavioral performance of decision analysis methods is more fragile than often assumed, and that the interplay between human judgment and structured elicitation procedures warrants systematic investigation.

The BDR literature has identified a broad set of biases relevant to decision analysis and has proposed taxonomies to understand their origins and potential for correction (Arkes, 1991;

Montibeller & von Winterfeldt, 2015). One widely used classification distinguishes among three types.

• Strategy-Based biases, arising from heuristic simplifications or effort-reduction strategies and are often viewed as more amenable to correction through structured elicitation. Examples include the status quo bias, the endowment effect, and the base rate fallacy.

• Association-Based biases, arising from the automatic activation of mental associations stored in memory, which becomes a liability when the associations triggered are judgmentally irrelevant or misleading. Examples include the overconfidence bias, the availability bias, and the omission bias.

• Psychophysically-Based biases, rooted in perceptual limitations and nonlinear mapping of physical stimuli to psychological responses. Examples include the anchoring bias, the gain-loss bias, and the proxy bias.

This classification provides insight into why "decision analysis mitigates bias" may hold for some cases but not others. In particular, the structured elicitation procedures may help reduce strategy-based biases by discouraging the use of simple strategies and superficial shortcuts. In contrast, association-based and psychophysically-based biases, being more deeply rooted in automatic associations and perceptual processes, may persist despite formal procedures. This highlights the need for a deeper and systematic investigation into how different biases interact with decision analysis methods, and how to devise methods such that the effect of these biases can be reduced.

1.2 Research Gaps

Although a growing body of research in behavioral decision research (BDR) has documented that cognitive biases can influence human judgments, several important gaps remain in understanding how these biases operate within decision analysis methods, particularly in MADM methods.

First, while cognitive biases have been extensively identified and studied in psychology, relatively little work has examined their impact within decision analysis. Psychological studies typically investigate simple or binary choices in unaided decision-making environments, where individuals rely on intuitive or holistic judgments (Gilovich et al., 2002; Tversky & Kahneman, 1974). In contrast, decision analysis addresses more complex MADM problems involving multiple attributes, multiple alternatives, and explicit tradeoffs, and is implemented within an aided decision-making environment where structured elicitation procedures and analyst guidance encourage deeper reflection (Keeney, 1982). Because of these fundamental differences in decision problem, decision environment, and process transparency, findings from psychological studies cannot be directly generalized to decision analysis. MADM methods decompose the decision process, for instance, by eliciting value functions, assessing weights, and aggregating preferences, thereby allowing researchers to observe where and how a bias enters the procedure and how it ultimately influences outcomes. This level of insight is not achievable in studies of unaided, holistic decision-making. Consequently, we still lack a systematic understanding of how cognitive biases affect each stage of MADM elicitation and how they shape the final decision results.

Second, most existing work on cognitive biases in decision analysis has focused on descriptive BDR, documenting the presence of biases and their influence on judgments within specific methods. While this descriptive evidence is essential, it does not address the prescriptive purpose of decision analysis, which is to support and improve decision-making. Therefore, it is critical to examine not only whether a bias exists but also how to reduce its influence and how to narrow the gap between the normatively optimal solution a method can produce and the judgmental deviations that arise during elicitation. Bias mitigation strategies in decision making can be generally divided into three categories (Fasolo et al., 2024): (i) debiasing, which aims to improve bias awareness and provide cognitive tools such as thinking strategies to reduce bias in judgments and decisions; (ii) choice architecture, which changes the structure of the decision environment or the presentation of information to support less biased judgments and decisions; and (iii) a dual approach that integrates both simultaneously. Since MADM methods have explicit, stepwise elicitation procedures, they provide unique opportunities for designing and embedding bias mitigation strategies into the method itself. The bias mitigation streams map onto the MADM context as follows: (i) debiasing primarily targets the judgmental inputs to the method, (ii) choice architecture operates through the design of the elicitation and aggregation procedures, and (iii) the dual approach utilizes the debiasing thinking strategies to redesign the elicitation procedures such that both the inputs and outputs of MADM methods are less biased. Yet, current research offers limited insight into how such prescriptive interventions should be developed or evaluated.

Third, despite increasing awareness that cognitive biases can affect decision analysis, rigorous empirical evidence remains limited, especially in the context of complex MADM settings. More systematic empirical research is needed to track how biases influence intermediate judgments, how they propagate to final decisions, and to compare different elicitation procedures in terms of their susceptibility to bias.

Together, these gaps highlight the need for a more comprehensive and empirically grounded understanding of how cognitive biases interact with the structure of MADM methods. Such understanding is essential for evaluating the behavioral robustness of decision analysis and for developing prescriptive strategies that improve judgment quality in complex decision contexts.

1.3 Research Focus

This dissertation examines how cognitive biases manifest within the Multi-Attribute Value Theory (MAVT) (Keeney & Raiffa, 1993), a well-known and widely used MADM method. This theory is grounded in the principles of utility theory, where the objective is to represent DM's preferences through a mathematical function that aggregates the values or utilities of different attributes into a single score. This dissertation focuses on four biases that are both theoretically central in behavioral decision research and highly relevant to the performance of decision analysis methods: three psychophysically-based biases (anchoring bias, loss aversion, and framing effect), and one strategy-based bias (status quo bias). These biases were selected because they are pervasive across laboratory and field studies, have significant impact on decision-making, have well-established psychological foundations, and are identified as highly relevant in decision analysis (Montibeller & von Winterfeldt, 2015). Examining these biases within the structured framework of MAVT allows us to identify the mechanisms through which biases enter the

elicitation process and to develop mitigation strategies. Through this analysis, the dissertation provides broader insights into the behavioral robustness of decision analysis methods.

Several considerations guided the selection of the cognitive biases examined in this dissertation. First, applying a decision analysis method involves four key elements: the decision-maker, the analyst, the method, and the decision problem. While the analyst is expected to anticipate potential biases and design the elicitation accordingly, the remaining three elements can still trigger systematic distortions in the decision-maker's judgments. These triggers differ in their origins even though the bias is ultimately expressed by the decision-maker. Anchoring bias and loss aversion may be activated by the elicitation procedures embedded in the method, such as numerical anchors introduced during the value function elicitation procedures. The framing effect can arise from the structure or presentation of the decision problem, such as how attributes or consequences are described. Status quo bias, in contrast, is not induced by the method or the problem framing but stems from the decision-maker's existing state, prior choices, or habitual reference point. Together, these four biases therefore capture distinct channels through which cognitive biases can enter a decision analysis process, enabling a comprehensive examination of how different bias types interact with a structured decision analysis method.

Second, the four biases are theoretically unified through the foundational principles of prospect theory and, in particular, its core assumption of reference-dependent evaluation. Prospect theory posits that individuals assess outcomes relative to a reference point, and such reference dependence gives rise to systematic deviations from normative rationality (Kahneman et al., 1979; Tversky & Kahneman, 1991). Loss aversion and the framing effect emerge directly from this mechanism: individuals code outcomes as gains or losses relative to a reference point, overweight losses compared to commensurate gains, and systematically shift their preferences when equivalent options are framed as gains or losses (Tversky & Kahneman, 1981). Anchoring bias can also be understood within a reference-dependent framework (Vassilopoulos et al., 2024; Berg & Moss, 2022; Godlonton et al., 2018; Sagi, 2006). Anchoring bias refers to the tendency to make insufficient adjustments towards the anchor (Tversky & Kahneman, 1974; Furnham & Boo, 2011), which could be a reference point. Finally, status quo bias reflects the tendency to use the current state as the reference point, such that deviations from the status quo are perceived as potential losses and weighed higher than the potential gains of a new option (Samuelson & Zeckhauser, 1988; Kahneman et al., 1991). Together, these four biases represent distinct but closely related manifestations of reference-dependent judgment, making them theoretically coherent for investigation.

Third, the framing effect and loss aversion are both rooted in people's sensitivity to equivalent gains and losses, and are generally referred to as gain-loss bias within BDR (Montibeller & von Winterfeldt, 2015). This allows us to study them jointly, enabling a deeper understanding of how multiple biases may interact within a decision analysis method.

Finally, the four biases allow us to engage directly with theoretical claims about the relative difficulty of debiasing (Montibeller & von Winterfeldt, 2015): psychophysically-based biases, such as anchoring, framing, and loss aversion, are thought to be harder to correct, whereas strategy-based biases, such as status quo bias, may be more amenable to mitigation through structured elicitation.

A wide range of multi-attribute decision-making (MADM) methods has been developed to support complex decision problems, including value-based methods (e.g., MAVT/MAUT

(Keeney & Raiffa, 1993)), pairwise comparison methods (e.g., AHP (Saaty, 1977), ANP (Saaty, 1996)), distance-based methods (e.g., TOPSIS (Hwang & Yoon, 1981), VIKOR (Opricovic & Tzeng, 2004)), and outranking methods (e.g., ELECTRE (Roy, 1968), PROMETHEE (Brans & Vincke, 1985)). These methods differ substantially in how preferences are elicited, represented, and aggregated, reflecting different assumptions about decision-makers' judgments and the nature of compensation across attributes. All MADM methods warrant systematic examination with respect to how cognitive bias can affect their inputs (judgments), and how that affects the decision-making process and decision outcomes.

Despite these differences, most commonly used MADM methods rely on elicitation procedures that integrate value judgments and tradeoff assessments in a largely holistic manner. Whether through pairwise comparisons, scoring rules, distance measures, or outranking relations, preferences are typically expressed in a way that does not explicitly separate evaluations of attribute levels from judgments about their relative importance. This intertwined structure limits the ability to disentangle where distortions arise during the decision-making process, as potential effects on value judgments and tradeoffs cannot be examined independently. In contrast, MAVT explicitly decomposes preference elicitation into two distinct stages: the elicitation of attribute-specific value functions and the elicitation of attribute weights. This decomposition provides a transparent and diagnostically precise framework for examining how different components of the decision-making process contribute to the final outcome. For this reason, this dissertation focuses on MAVT, while recognizing that extending similar analyses to other MADM methods remains an important direction for future research.

1.4 Research Questions

The main research question is therefore:

"*How do anchoring bias, loss aversion, framing effect, and status quo bias affect the decision-making process and outcome within multi-attribute value theory, and how can these influences be mitigated?*"

To address this main research question, the dissertation investigates three sub-research questions, each focusing on a specific bias or combination of biases and on the mechanisms through which they interact with MAVT.

RQ1: *How can anchoring bias affect the attribute-specific value function and weight elicitation within multi-attribute value theory, and how can such effects be mitigated?*

This research question examines how anchoring bias, one of the most pervasive cognitive biases arising from insufficient adjustment from an initial reference point (Tversky & Kahneman, 1974; Montibeller & von Winterfeldt, 2015; Epley & Gilovich, 2006, 2005), manifests within the structured elicitation processes of MAVT. Anchoring is theoretically important in this context because MAVT requires decision-makers to provide numerical judgments during value function elicitation and weight elicitation, both of which involve implicit or explicit starting points that may serve as anchors. If decision-makers rely disproportionately on these starting points, their judgments may deviate systematically from normative preferences, potentially distorting both intermediate outputs and final decision results.

By addressing this research question, the dissertation investigates the susceptibility of two core MAVT components, attribute-specific value functions and attribute weights, to anchoring bias. These elicitation steps involve subjective numerical assessments that are cognitively demanding and therefore particularly vulnerable to anchoring bias. Studying anchoring within MAVT provides novel theoretical and empirical insights because the structured, stepwise nature of MAVT enables detailed observation of how the bias enters specific stages of the elicitation process, propagates through the model, and influences final decisions. Such mechanistic understanding cannot be obtained from psychological studies that rely on unaided or holistic judgments.

To address this research question, a systematic literature review on anchoring bias is conducted, from which hypotheses are developed regarding its effects on MAVT elicitation procedures, namely value function elicitation and weight elicitation. This motivates two empirical studies, each targeting a distinct elicitation step within MAVT.

The first study investigates how anchors affect the attribute-specific value functions elicited through the midvalue splitting procedure (Keeney & Raiffa, 1993), and tests two debiasing strategies embedded directly into this elicitation protocol. The second examines anchoring in the tradeoff procedure and evaluates whether the Best–Worst Tradeoff method (Liang et al., 2022) reduces anchoring by relying on both best and worst attributes during elicitation. Each study is implemented through controlled experiments designed within the MAVT framework. After statistically analyzing the collected data, the results provide systematic evidence on the presence of anchoring within MAVT and on the effectiveness of the proposed mitigation strategies.

RQ2: *How do the loss aversion and the framing effect operate across multiple stages of multi-attribute value theory and collectively influence the decision-making process and outcomes?*

This research question examines the joint operation of multiple cognitive biases and addresses a gap that is largely overlooked in BDR. Most existing studies investigate biases in isolation, even though real decision-making, especially within MADM, almost always involves the simultaneous influence of several biases arising from multiple sources and acting across multiple elicitation procedures. For MAVT in particular, decisions are constructed through successive stages in which decision-makers evaluate attribute levels, form value functions, assess tradeoffs, and integrate preferences. At each of these stages, distinct cognitive processes are engaged, and thus multiple biases may collectively shape the final decision outcome. Understanding and mitigating the combined influence of multiple biases is therefore essential for assessing the behavioral robustness of MADM methods.

Within this context, this research question focuses on two theoretically closely related reference-dependent biases, loss aversion, and the framing effect, and investigates how they jointly influence the elicitation procedures of MAVT. Both biases originate from how individuals evaluate outcomes relative to a reference point, yet they operate through different mechanisms: framing determines whether an outcome is encoded as a gain or a loss, while loss aversion determines the relative psychological weight of those gains and losses. Although both have been extensively studied in simple, unaided decision tasks (Kahneman et al., 1979; Tversky & Kahneman, 1989, 1981, 1991), little is known about how they emerge, combine, or potentially amplify each other in multi-attribute settings where value construction and tradeoff assessments unfold step

by step. Because MAVT decomposes the decision-making process into sequential stages, these biases may enter at different points and interact in ways that cannot be observed in holistic or binary-choice environments. Investigating such interactions is therefore critical for identifying how multiple biases collectively distort the MAVT process and where targeted corrective interventions may be most effective.

To address this research question, an empirical study is conducted. Hypotheses are developed within MAVT regarding (i) how framing shapes the construction of attribute-specific value functions, (ii) how loss aversion influences attribute tradeoffs, and (iii) how the two biases interact across stages to alter the decision outcomes. This study provides structured evidence on how framing and loss aversion jointly shape decision-making within a formal MADM method, thereby contributing new insights into the behavioral performance of MAVT and the broader challenge of understanding and mitigating interacting biases in decision analysis.

RQ3: *How effective is multi-attribute value theory in reducing the status quo bias?*

Whereas the first two research questions focus on examining how cognitive biases can enter and distort specific elicitation procedures within MAVT, they also suggest that the structured nature of decision analysis methods may create opportunities for mitigating such biases when behavioral considerations are incorporated into the elicitation process. Building on this insight, the third research question shifts the focus from vulnerability to mitigation and examines whether MAVT itself can reduce cognitive biases.

This research question investigates whether MAVT, a structured decision analysis method, can reduce the influence of status quo bias, a strategy-based cognitive bias. Status quo bias is characterized by a disproportionate preference for the current or default option even when superior alternatives are available (Samuelson & Zeckhauser, 1988). Decision analysis methods, in principle, should mitigate strategy-based biases by encouraging decision-makers to articulate their objectives, reflect on attribute tradeoffs, and evaluate alternatives systematically. Yet, empirical evidence on whether, and through which mechanisms, decision analysis methods actually reduce status quo bias and generally strategy-based bias in MADM remains limited. Understanding this is essential for assessing the behavioral robustness of MAVT and, more broadly, for determining the extent to which structured decision procedures can mitigate intuitive errors that arise in unaided decision-making, especially strategy-based biases.

To address this question, an empirical study is conducted. I first conduct a systematic literature review of status quo bias and develop hypotheses that distinguish between different sources of the status quo, such as a participant's real status quo and an experimentally provided status quo. These hypotheses clarify how each source may enter the MAVT elicitation steps and how the multi-stage structure of MAVT may mitigate the bias. I then examine these hypotheses through an empirical study grounded in the MAVT procedure. The findings provide systematic evidence on the debiasing effectiveness of MAVT and contribute to broader discussions on the ability of structured decision analysis methods to reduce strategy-based biases.

1.5 Dissertation Outline

This dissertation is organized around four empirical studies designed to examine how four cognitive biases, anchoring bias, loss aversion, the framing effect, and status quo bias, enter the

elicitation procedures of multi-attribute value theory (MAVT), how they affect the decision-making process and outcome, and how such influences can be mitigated.

Chapter 1 introduces the research background, identifies research gaps, defines the research focus, formulates three research questions, and outlines the structure of the dissertation. **Chapter 2** investigates the influence of anchoring bias on the attribute-specific value function elicitation in MAVT. Through a controlled experiment, the chapter examines how numerical anchors introduced during the elicitation procedure distort decision-makers' judgments and lead to biased attribute-specific value functions, ultimately affecting the decision outcome, and how debiasing strategies can be designed to mitigate such effects. **Chapter 3** examines anchoring bias in the weight elicitation step, focusing on the tradeoff procedure commonly used in the MAVT. The study examines whether the most and least important attributes used to construct the indifference pairs can lead to biased weights and decision outcomes, and whether the Best-Worst Tradeoff method can mitigate anchoring bias by utilizing both attributes to construct the indifference pairs. **Chapter 4** explores two biases, loss aversion and framing effect, and their interactions within MAVT. This chapter explored not only how each bias may affect specific elicitation steps, but also the interaction between two biases and two elicitation procedures. **Chapter 5** evaluates the effectiveness of MAVT in mitigating a strategy-based bias: status quo bias. Besides exploring the effectiveness of mitigation, it also identifies how status quo bias enters the decision analysis process by examining its influence on the value function and weight elicitation procedures. **Chapter 6** synthesizes insights across all empirical studies. It provides an integrated conclusion on how cognitive biases enter different stages of MAVT, how distortions at one stage interact with subsequent stages, and how these biases can be mitigated. The chapter discusses the theoretical and practical implications of the findings, acknowledges limitations, and outlines several directions for future research on prescriptive behavioral decision analysis.

The outline of this dissertation is illustrated in Figure 1.1.

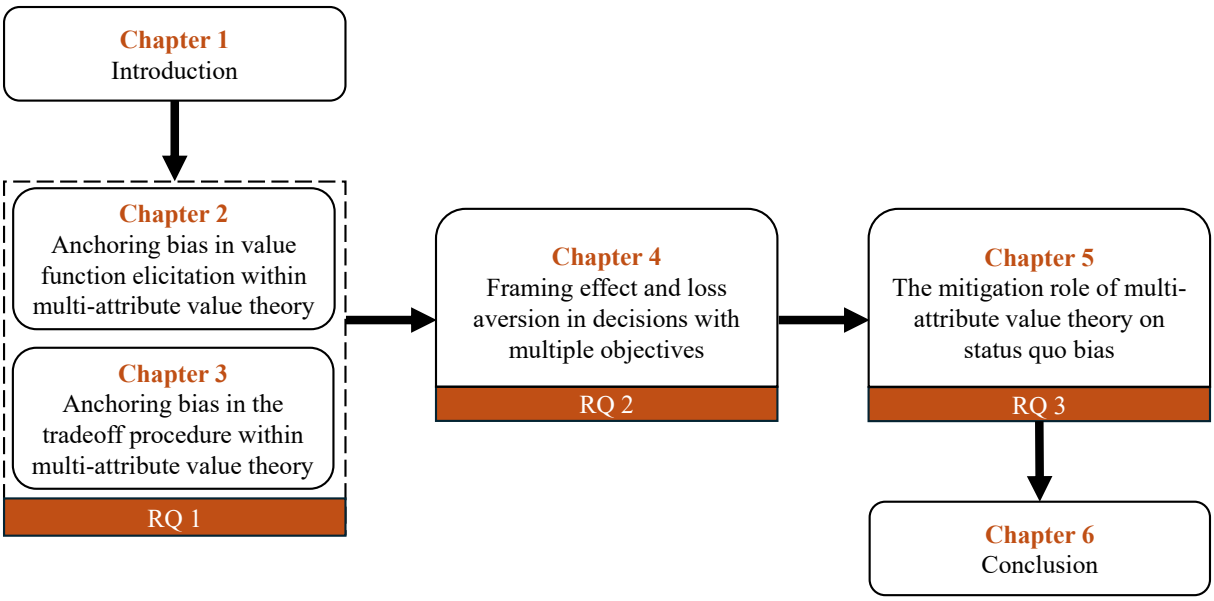


Figure 1.1: The outline of the dissertation

Bibliography

Arkes, H. R. (1991) Costs and benefits of judgment errors: Implications for debiasing., *Psychological Bulletin*, 110(3), pp. 486–498.

Belton, V., T. Stewart (2012) *Multiple criteria decision analysis: an integrated approach*, Springer Science & Business Media, New York.

Berg, S. A., J. H. Moss (2022) Anchoring and judgment bias: disregarding under uncertainty, *Psychological Reports*, 125(5), pp. 2688–2708.

Bond, S. D., K. A. Carlson, R. L. Keeney (2008) Generating objectives: can decision makers articulate what they want?, *Management Science*, 54(1), pp. 56–70.

Borcherding, K., D. von Winterfeldt (1988) The effect of varying value trees on multiattribute evaluations, *Acta Psychologica*, 68(1-3), pp. 153–170.

Brans, J.-P., P. Vincke (1985) Note—a preference ranking organisation method: (the promethee method for multiple criteria decision-making), *Management Science*, 31(6), pp. 647–656.

Cohen, L. J. (1981) Can human irrationality be experimentally demonstrated?, *Behavioral and Brain Sciences*, 4(3), pp. 317–331.

Dunning, D., D. W. Griffin, J. D. Milojkovic, L. Ross (1990) The overconfidence effect in social prediction, *Journal of Personality and Social Psychology*, 58(4), pp. 568–581.

Edwards, W. (1954) The theory of decision making, *Psychological Bulletin*, 51(4), pp. 380–417.

Einhorn, H. J., R. M. Hogarth (1981) Behavioral decision theory: Processes of judgement and choice, *Annual Review of Psychology*, 32(1981), pp. 53–88.

Epley, N., T. Gilovich (2005) When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors, *Journal of Behavioral Decision Making*, 18(3), pp. 199–212.

Epley, N., T. Gilovich (2006) The anchoring-and-adjustment heuristic: Why the adjustments are insufficient, *Psychological Science*, 17(4), pp. 311–318.

Fasolo, B., C. Heard, I. Scopelliti (2024) Mitigating cognitive bias to improve organizational decisions: An integrative review, framework, and research agenda, *Journal of Management*, 51(6), pp. 2182–2211.

Federspiel, F. M., G. Montibeller, M. Seifert (2024) Behavioral decision analysis: past, present and future, in: *Behavioral Decision Analysis*, Springer, pp. 1–14.

Figueira, J., S. Greco, M. Ehrgott (2005) *Multiple criteria decision analysis: state of the art surveys*, Springer Science & Business Media, New York.

Fischer, G. W. (1995) Range sensitivity of attribute weights in multiattribute value models, *Organizational Behavior and Human Decision Processes*, 62(3), pp. 252–266.

Fischer, G. W., N. Damodaran, K. B. Laskey, D. Lincoln (1987) Preferences for proxy attributes, *Management Science*, 33(2), pp. 198–214.

Fischhoff, B., S. B. Broomell (2020) Judgment and decision making, *Annual Review of Psychology*, 71(1), pp. 331–355.

Fishburn, P. C. (1970) *Utility theory for decision making*, John Wiley & Sons, New York.

Fishburn, P. C. (1989) Foundations of decision analysis: along the way, *Management Science*, 35(4), pp. 387–405.

French, S. (1989) *Readings in decision analysis*, CRC Press, Florida.

Furnham, A., H. C. Boo (2011) A literature review of the anchoring effect, *The Journal of Socio-Economics*, 40(1), pp. 35–42.

Gilovich, T. (1981) Seeing the past in the present: The effect of associations to familiar events on judgments and decisions, *Journal of Personality and Social Psychology*, 40(5), pp. 797–808.

Gilovich, T., D. Griffin, D. Kahneman (2002) *Heuristics and biases: The psychology of intuitive judgment*, Cambridge University Press, Cambridge.

Godlonton, S., M. A. Hernandez, M. Murphy (2018) Anchoring bias in recall data: Evidence from central america, *American Journal of Agricultural Economics*, 100(2), pp. 479–501.

Griffin, D., R. Gonzalez, C. Varey (2001) The heuristics and biases approach to judgment under uncertainty, in: *Blackwell handbook of social psychology: Intraindividual processes*, Blackwell Publishers Ltd, pp. 207–235.

Hammond, J. S., R. L. Keeney, H. Raiffa (1998) The hidden traps in decision making, *Harvard Business Review*, 76(5), pp. 47–58.

Hogarth, R. M. (1981) Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics, *Psychological Bulletin*, 90(2), pp. 197–217.

Howard, R. A. (2007) The foundations of decision analysis, *IEEE Transactions on Systems Science and Cybernetics*, 4(3), pp. 211–219.

Hwang, C.-L., K. Yoon (1981) Methods for multiple attribute decision making, in: *Multiple attribute decision making: methods and applications a state-of-the-art survey*, Springer, pp. 58–191.

Jacobi, S. K., B. F. Hobbs (2007) Quantifying and mitigating the splitting bias and other value tree-induced weighting biases, *Decision Analysis*, 4(4), pp. 194–210.

Jones, B. D. (1999) Bounded rationality, *Annual Review of Political Science*, 2(1), pp. 297–321.

Kahneman, D. (2011) *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Kahneman, D., J. L. Knetsch, R. H. Thaler (1991) The endowment effect, loss aversion, and status quo bias, *The Journal of Economic Perspectives*, 5(1), pp. 193–206.

Kahneman, D., A. Tversky (1972) Subjective probability: A judgment of representativeness, *Cognitive Psychology*, 3(3), pp. 430–454.

Kahneman, D., A. Tversky (1973) On the psychology of prediction, *Psychological Review*, 80(4), pp. 237–251.

Kahneman, D., A. Tversky, et al. (1979) Prospect theory: An analysis of decision under risk, *Econometrica*, 47(2), pp. 363–391.

Keeney, R. L. (1977) The art of assessing multiattribute utility functions, *Organizational Behavior and Human Performance*, 19(2), pp. 267–310.

Keeney, R. L. (1982) Decision analysis: an overview, *Operations Research*, 30(5), pp. 803–838.

Keeney, R. L. (2004) Making better decision makers, *Decision Analysis*, 1(4), pp. 193–204.

Keeney, R. L., H. Raiffa (1993) *Decisions with multiple objectives: preferences and value trade-offs*, Cambridge University Press, Cambridge.

Liang, F., M. Brunelli, J. Rezaei (2022) Best-worst tradeoff method, *Information Sciences*, 610, pp. 957–976.

Milkman, K. L., D. Chugh, M. H. Bazerman (2009) How can decision making be improved?, *Perspectives on Psychological Science*, 4(4), pp. 379–383.

Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63(2), pp. 81–97.

Montibeller, G., D. von Winterfeldt (2015) Cognitive and motivational biases in decision and risk analysis, *Risk Analysis*, 35(7), pp. 1230–1251.

Montibeller, G., D. von Winterfeldt (2024) Behavioral decision research: Descriptive and prescriptive perspectives, in: *Behavioral Decision Analysis*, Springer, pp. 15–40.

Opricovic, S., G.-H. Tzeng (2004) Compromise solution by mcdm methods: A comparative analysis of vikor and topsis, *European Journal of Operational Research*, 156(2), pp. 445–455.

Payne, J. W., J. R. Bettman, E. J. Johnson (1992) Behavioral decision research: A constructive processing perspective, *Annual Review of Psychology*, 43(1), pp. 87–131.

Peer, E., E. Gamliel (2013) Heuristics and biases in judicial decisions, *Court Review: The Journal of the American Judges Association*, 49, pp. 114–118.

Rezaei, J. (2021) Anchoring bias in eliciting attribute weights and values in multi-attribute decision-making, *Journal of Decision Systems*, 30(1), pp. 72–96.

Rezaei, J., A. Arab, M. Mehregan (2022) Equalizing bias in eliciting attribute weights in multiattribute decision-making: experimental research, *Journal of Behavioral Decision Making*, 35(2), e2262.

Rezaei, J., A. Arab, M. Mehregan (2024) Analyzing anchoring bias in attribute weight elicitation of smart, swing, and best-worst method, *International Transactions in Operational Research*, 31(2), pp. 918–948.

Roy, B. (1968) Classement et choix en présence de points de vue multiples, *Revue Française D'informatique Et De Recherche Opérationnelle*, 2(8), pp. 57–75.

Roy, B., D. Vanderpooten (1996) The european school of mcda: Emergence, basic features and current works, *Journal of Multi-Criteria Decision Analysis*, 5(1), pp. 22–38.

Saaty, T. L. (1977) A scaling method for priorities in hierarchical structures, *Journal of Mathematical Psychology*, 15(3), pp. 234–281.

Saaty, T. L. (1996) *Decision making with dependence and feedback: The analytic network process*, RWS Publications, Pittsburgh.

Sagi, J. S. (2006) Anchored preference relations, *Journal of Economic Theory*, 130(1), pp. 283–295.

Samuelson, W., R. Zeckhauser (1988) Status quo bias in decision making, *Journal of Risk and Uncertainty*, 1(1), pp. 7–59.

Shafir, E., R. A. LeBoeuf (2002) Rationality, *Annual Review of Psychology*, 53(1), pp. 491–517.

Simon, H. A. (1955) A behavioral model of rational choice, *The Quarterly Journal of Economics*, 69(1), pp. 99–118.

Simon, H. A. (1978) Rationality as process and as product of thought, *The American Economic Review*, 68(2), pp. 1–16.

Slovic, P., B. Fischhoff, S. Lichtenstein (1977) Behavioral decision theory., *Annual Review of Psychology*, 28, pp. 1–39.

Tversky, A., D. Kahneman (1973) Availability: A heuristic for judging frequency and probability, *Cognitive Psychology*, 5(2), pp. 207–232.

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

Tversky, A., D. Kahneman (1981) The framing of decisions and the psychology of choice, *Science*, 211(4481), pp. 453–458.

Tversky, A., D. Kahneman (1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment, *Psychological Review*, 90(4), pp. 293–315.

Tversky, A., D. Kahneman (1989) Rational choice and the framing of decisions, in: *Multiple criteria decision making and risk analysis using microcomputers*, Springer, pp. 81–126.

Tversky, A., D. Kahneman (1991) Loss aversion in risaess choice: A reference-dependent model, *The Quarterly Journal of Economics*, 106(4), pp. 1039–1061.

Tzeng, G.-H., J.-J. Huang (2011) *Multiple attribute decision making: methods and applications*, CRC press, Florida.

Vassilopoulos, A., A. C. Drichoutis, R. M. Nayga (2024) Reference dependence, expectations and anchoring in the becker-degroot-marschak mechanism, *Theory and Decision*, 97(4), pp. 637–683.

von Winterfeldt, D. (1999) On the relevance of behavioral decision research for decision analysis, in: *Decision science and technology: Reflections on the contributions of Ward Edwards*, Springer, pp. 133–154.

Weber, M., K. Borcherding (1993) Behavioral influences on weight judgments in multiattribute decision making, *European Journal of Operational Research*, 67(1), pp. 1–12.

Zanakis, S. H., A. Solomon, N. Wishart, S. Dublish (1998) Multi-attribute decision making: A simulation comparison of select methods, *European Journal of Operational Research*, 107(3), pp. 507–529.

Chapter 2

Anchoring Bias in Value Function Elicitation within Multi-Attribute Value Theory

---

Abstract: Anchoring bias refers to the human tendency to rely heavily on an initial piece of information when making judgments. This bias has significant implications for decision analysis methods that rely on human judgments. This study examines the influence of anchoring bias in the value function elicitation step of the multi-attribute value theory, specifically within the midvalue splitting procedure. We hypothesize that the starting point provided by the analyst during elicitation creates a bias in decision-maker's judgments, leading to distorted value functions and ultimately affecting decision outcomes. We also hypothesize that counter-anchoring and avoiding the use of anchors mitigate the effect of anchoring bias. To test the hypotheses, we designed an experiment and collected data from 320 subjects. The findings show that the starting point in the midvalue splitting procedure could change the attribute-specific value functions and, consequently, the overall value of the alternatives. Additionally, two debiasing strategies, counter-anchoring and avoiding the use of anchors, were found to be effective in reducing the effect of anchoring bias. The implications of this study can extend to other structured value function elicitation methods.

Keywords: Anchoring bias; value function; midvalue splitting procedure; multi-attribute value theory; debiasing

This chapter is based on the following journal paper:

Sun, G., Kroesen, M., Rezaei, J (2025) Anchoring Bias in Value Function Elicitation Within Multiattribute Value Theory, *Decision Analysis*, 22(4), pp.284-304.

2.1 Introduction

Cognitive bias refers to the systematic error in judgment that occurs when individuals process and interpret information (Tversky & Kahneman, 1974). It is a well-documented phenomenon in psychology and emerges from the human tendency to use heuristics, or mental shortcuts, to simplify decision-making in complex environments. While such shortcuts can be efficient in reducing cognitive load, they could lead to flawed judgments (Kahneman, 2011). Over the past decades, a growing body of research has revealed the pervasive influence of cognitive bias across various fields (Ariely, 2008; Bazerman & Moore, 2012; Thaler & Sunstein, 2008; Bushong & Gagnon-Bartsch, 2024). In financial contexts, for instance, overconfidence bias can lead investors to overestimate their knowledge or abilities to predict outcomes, causing them to make ill-informed investment choices and suffer significant financial losses (Barber & Odean, 2001). In managerial decision-making, the framing effect, where people respond differently to the same situation depending on how they are framed, can lead to inconsistent results. For instance, managers may opt for riskier choices when the same scenario is framed as avoiding losses rather than achieving gains (Tversky & Kahneman, 1981).

Given these widespread implications, cognitive bias has become a central concern in behavioral decision research (BDR) that is divided into two major branches: descriptive BDR and prescriptive BDR (Montibeller & von Winterfeldt, 2024). Descriptive BDR aims at developing theories or models to describe the biases in judgments, while prescriptive BDR focuses on developing debiasing strategies to correct those biases, improving the quality of decisions. Notably, different from the descriptive view above, cognitive bias in prescriptive research is defined as "a systematic discrepancy between the 'correct' answer in a judgmental task, given by a formal normative rule, and the decisionmaker's or expert's actual answer to such a task" (Montibeller & von Winterfeldt, 2015, pp. 1231). This definition highlights the role of the normative rule and, by extension, the role of methodological design, in prescriptive BDR.

In descriptive BDR, cognitive bias is generally assessed by measuring the deviation of human judgment from either the true value or the coherence of preferences in an unsupported decision-making environment (Montibeller & von Winterfeldt, 2024). Conversely, the prescriptive view introduces a medium, the decision analysis method, an aspect considered peripheral or non-central in the descriptive view. Within the decision analysis field, a variety of methods have been developed to support decision-makers (DMs) in structuring and quantifying their preferences, in order to evaluate and choose among alternatives characterized by multiple, often conflicting, attributes (Keeney & Raiffa, 1976; Belton & Stewart, 2012). The procedures of these methods typically involve a series of structured questions posed by an analyst to elicit the DM's preferences. These preferences are inferred based on the value judgments of DMs. The underlying assumption of these methods is that DMs hold coherent and stable preferences unaffected by irrelevant information during these judgmental tasks (Fischhoff, 1991), and that these methods can reduce human biases in decision-making (Keeney, 2004). However, this assumption has been extensively challenged by evidence of cognitive bias in decision analysis methods in prescriptive BDR.

The methodological design of the elicitation procedures plays an important role in prescriptive BDR. This is primarily because preferences are not merely revealed but constructed during the elicitation process (Slovic, 1995). Empirical research has shown that preferences can be shaped by contextual factors and the elicitation task itself (Payne et al., 1992). This construc-

tive view assigns critical importance to the elicitation method, when thoughtfully and carefully devised, the method can reduce potential biases during this process; however, if the methodological design neglects cognitive phenomena, it may inadvertently exacerbate existing biases. So, the goal of prescriptive BDR is not only to identify these deviations but also to develop procedures to mitigate these biases and ensure that decision analysis methods remain effective in real-world contexts.

This study investigates how anchoring bias, one of the most important cognitive biases (Tversky & Kahneman, 1974; Kahneman, 2011; Montibeller & von Winterfeldt, 2015), influences the multi-attribute value theory (MAVT), particularly the value function elicitation step. MAVT is a well-known and widely applied decision analysis method developed by Keeney & Raiffa (1976). The value function elicitation step involves determining attribute-specific value functions, which represent the DM's value judgments over different attribute levels. This step not only allows us to directly observe how cognitive bias may distort the representation of preference, but it also provides a quantifiable structure through which we can assess the impact of bias on the value function.

There are different ways to elicit a value function, including the standard difference procedure, the lock-step procedure, direct rating, successive comparison and curve fitting, among others (Beinat, 1997; Fishburn, 1967; Watson & Buede, 1987; Keeney & Raiffa, 1976). One of the most commonly used and the focus of this study is the midvalue splitting procedure. In this procedure, multiple midvalue points will be identified by the DM, which are then used to form the value function. To begin the process, the analyst provides a starting point for the DM to identify whether it is the first midvalue point. However, due to the effect of anchoring bias, the DM might adjust insufficiently away from the starting point, which results in a biased midvalue point. This will in turn affect the value function, and ultimately, the final outcome of the decision-making problem. Through this investigation, we aim to understand how anchoring bias impacts the value function shape and the final decision outcomes, as well as to identify effective strategies to mitigate its effects. To achieve this, we develop a few hypotheses and design an experiment to test them. Data were collected from 320 participants, who completed the MAVT procedure under varying anchoring conditions during the value function elicitation step. To assess the effects of anchoring bias on value function elicitation and the overall value of the alternatives, we conducted both parametric and non-parametric statistical analyses.

This study makes two main contributions to the field of decision analysis. First, it provides a detailed analysis of how anchoring bias affects the value function and, consequently, the decision results of MAVT. It enhances our understanding of anchoring bias in value assessment, offering insights that can extend to other value-based decision methods. Second, it develops effective debiasing strategies to mitigate the anchoring effect. These strategies aim to improve the consistency and reliability of value function elicitation procedure, thereby enhancing the credibility of MAVT and related decision analysis methods in real-world applications. While our focus is on the midvalue splitting procedure in the value function elicitation step, the implications can be extended to other value function elicitation methods, which will be discussed in Section 2.6.4.

The remainder of this paper is organized as follows: Section 2.2 introduces anchoring bias and its implications in decision analysis. Section 2.3 describes the MAVT, with a particular emphasis on the midvalue splitting procedure. Section 2.4 presents the research hypotheses. Section 2.5 outlines the experimental design. Section 2.6 discusses the results of the study.

Finally, Section 2.7 concludes the paper and suggests directions for future research.

2.2 Anchoring Bias

Anchoring bias refers to the human tendency to rely heavily on an initial piece of information and to insufficiently adjust away from it when making judgments; see Furnham & Boo (2011) for a review. This initial information, known as the "anchor", can be provided externally or generated internally (Epley & Gilovich, 2005). External anchors are reference points originating outside the DM's judgment, such as information provided by other people, text, pictures or other contextual stimuli. Internal anchors, on the other hand, are based on a DM's prior beliefs, experiences, or memories. For instance, if someone is asked to estimate a product price, an external anchor could be a price tag, while an internal anchor might be her/his memory of the price of a similar product.

Empirical studies of anchoring effects typically use one of two paradigms: estimation tasks, where anchoring distorts accuracy relative to true value; and valuation tasks, where it affects the internal coherence of preferences in the absence of a known correct answer. The foundational study by Tversky & Kahneman (1974) illustrates anchoring in an estimation task. Two groups of participants are asked to estimate (in five seconds without a calculator) the product of a sequence of numbers presented in reverse order ($8\times7\times6\times5\times4\times3\times2\times1$ versus $1\times2\times3\times4\times5\times6\times7\times8$). Despite both sequences containing the same numbers, the group starting with a higher initial partial product (i.e., the product of the first few numbers) gave a significantly higher median estimate than the group starting with a lower initial product (2250 versus 512, respectively). Even though both groups produced estimates much lower than the true value (40320), the difference in their estimates (due to the initial anchor) demonstrates how anchoring bias distorts individual assessments.

Beyond this seminal work, anchoring effects have been documented across a wide range of applied domains. In estimation tasks, Prava et al. (2016) explores how different anchors, such as an ignorance prior, can influence the probability assessments in surveys. Anchoring bias occurs when participants do not adjust sufficiently from such anchors, leading to partition dependence and carryover biases, distorting the probability assessments. In valuation tasks, anchoring bias is equally pervasive. In negotiations, an initial offer can serve as a powerful anchor, influencing the final agreement terms even if the offer is arbitrary or exaggerated (Galinsky & Mussweiler, 2001). In marketing, the first price that consumers encounter, such as an initial product price or a reference price, can anchor their perceptions of value. This initial price can influence how consumers perceive the value of any subsequent prices, such as a discounted price, ultimately affecting their willingness to pay (Wansink et al., 1998). In healthcare, anchoring bias can lead to admission control bias and path-dependent feedback, where initial diagnostic uncertainty disproportionately influences subsequent decisions and resource allocation, potentially impacting patient outcomes (Kim & Tong, 2024). Anchoring also plays a critical role in legal settings, where initial sentencing recommendations, whether reasonable or not, can influence the final sentencing decisions of judges (Englich et al., 2006).

Within decision analysis, studies of anchoring bias have primarily focused on valuation tasks. Research has shown that anchoring bias significantly affects weight elicitation—the process by which the relative importance of the attributes (or scaling constants) is elicited.

Buchanan & Corner (1997) explored the impact of anchoring bias in two interactive decision analysis methods. In the Zoints and Wallenius method, the DM is usually provided with a fixed starting point as part of the method's initialization, whereas in the free search interactive method, there is no fixed starting point. Instead, it allows the DM to explore the feasible region freely. The experiment results show that the starting solution in the Zoints and Wallenius method had a significant impact on the decision outcome, demonstrating the anchoring effect. In contrast, the anchoring effect was not significant in the free search method. An insight gain from this study is that designing decision analysis methods with flexible starting points and iterative feedback can help reduce the impact of anchoring.

Jacobi & Hobbs (2007) focused on value tree-induced biases in weight elicitation methods, particularly the splitting bias. Due to splitting bias, decomposing an attribute into multiple sub-attributes leads to a higher global weight of that attribute compared to directly assessing its relative importance. They suggest that the use of anchor-and-adjustment heuristics is the main cause of such bias. DMs initially allocate equal weights across attributes within each tree partition. The equal allocation serves as an anchor for their judgments. DMs then insufficiently adjust these weights to align with their preferences, resulting in weights being more uniform than they should be.

Collectively, these findings demonstrate the pervasive influence of anchoring bias and highlight the necessity of prescriptive interventions to mitigate its impact. In the context of decision analysis, this points to a critical yet often overlooked aspect: while these methods are mathematically sound and intended to guide rational decision-making, their structural frameworks can unintentionally induce cognitive bias, such as anchoring bias, by default. Therefore, we should reevaluate how these methods are designed and implemented and, when necessary, improve them to address cognitive aspects, ensuring their effectiveness.

Generic approaches for reducing cognitive bias, such as consulting with experts, providing regular feedback, and increasing awareness of biases, have shown to be somewhat effective in reducing anchoring bias (Adame, 2016; Lilienfeld et al., 2009; Gavirneni & Xia, 2009; Fasolo et al., 2024). For prescriptive BDR, we need more targeted strategies that focus on the heuristics underlying anchoring bias. This study incorporates two such targeted approaches:

Avoiding Anchors: A straightforward yet effective strategy is to design decision-making processes that avoid reliance on predefined starting points (Montibeller & von Winterfeldt, 2015). Practical implementations include randomizing the presentation order of attributes, structuring decision tasks in a way to prevent the introduction of starting points, or encouraging DMs to independently generate their assessments without external prompts. For example, the free search interactive method, which removes fixed starting points, enables DMs to explore alternatives freely, without being influenced by anchoring bias. However, this strategy has been primarily designed for the external anchors. Internal anchors, which are inherently subjective and often viewed as more credible by DMs (Mussweiler & Strack, 1999; Wilson et al., 1996; Chapman & Johnson, 2002), require different interventions. For instance, approaches such as monetary rewards for being correct and forewarning individuals about possible biases in judgment have been shown to be effective (Epley & Gilovich, 2005).

Consider-the-Opposite Strategy: This strategy encourages DMs to critically evaluate their initial judgments by considering contradictory information or alternative scenarios (Lord et al., 1984; Mussweiler et al., 2000). In the decision analysis context, counter-anchors are a practical

way to operationalize this strategy. For instance, while traditional methods such as SMART and Swing rely on unidirectional anchors, the Best-Worst Method (BWM) (Rezaei, 2015) incorporates a bidirectional evaluation process. BWM requires pairwise comparisons of the best attribute against others and the others against the worst attribute, which helps to balance the evaluation by reducing the influence of a single anchor (Rezaei et al., 2024). However, the use of counter-anchors in decision analysis in practice is not without its challenges. First, if counter-anchor values are too implausible, DMs may recognize the potential for bias and overcorrect their judgments (Brewer et al., 2007). Second, the repeated use of counter-anchors can complicate the procedure and increase cognitive load. This may lead to mental fatigue, reduced engagement, and potentially less reliable responses from the DM (Kahneman, 2011).

Building on the two strategies, avoiding anchors and employing counter-anchors were designed specifically for mitigating anchoring bias in the midvalue splitting procedure. Detailed implementations will be introduced in Section 2.5.

2.3 Multi-Attribute Value Theory

Multi-attribute value theory (MAVT) is a well-established decision analysis method developed by Keeney & Raiffa (1976). This theory is grounded in the principles of utility theory, where the objective is to represent DM's preferences through a mathematical function that aggregates the values or utilities of different attributes into a single score. MAVT not only aids in the systematic evaluation of the alternatives but also enhances transparency in the decision-making process by clearly articulating the rationale behind the choices made (Höfer & Madlener, 2020; Anthes et al., 2019). MAVT has been widely applied and shown great success in various decision fields, such as environmental management (Hostmann et al., 2005), healthcare (Labijak-Kowalska et al., 2023), and policy-making (Ferretti, 2016). MAVT consists of the following five steps (Keeney, 2009; Keeney & Raiffa, 1976; Keeney, 2002):

Identification of the Objectives, Attributes and Alternatives: First, the problem must be structured, which includes (i) the objectives; (ii) attributes, the evaluators used to evaluate how well alternatives meet the objectives; and (iii) alternatives, the possible solutions available for the decision problem. The structuring process often uses tools like value trees to organize these elements hierarchically (Belton & Stewart, 2012).

Value Function Elicitation: Attribute-specific value functions should be developed. An attribute-specific value function translates the performance of an alternative on a specific attribute into a value score, typically on a scale from 0 to 1, where 0 is assigned to the least preferred level of the attribute and 1 is assigned to the most preferred level.

Weight Elicitation: This step involves eliciting scaling constants for each attribute that determine their influence in the overall value function. The Tradeoff procedure is commonly used in this step, where the DM is asked to compare two attributes at a time and determine how much of one attribute they would sacrifice to gain more of another.

Aggregation: Each alternative's aggregated score is calculated using an additive value function:

$$v(a_i) = \sum_{j=1}^{N} w_j v_j(a_{ij}) \tag{2.1}$$

where $v(a_i)$ is the overall value of alternative i, scaled from 0 to 1. $v_j(a_{ij})$ is the attribute-specific value representing the performance of alternative i with respect to attribute j, and w_j is the scaling constant (or weight) of the attribute j. To use an additive value function, the two primary conditions must be satisfied (mutual preference independence and difference independence) (Dyer & Sarin, 1979; Keeney & Raiffa, 1976).

Definition 2.1
The attributes $X_1, \ldots, X_N$ are *mutually preferentially independent* if any subset of attributes is preferentially independent of the remaining attributes.

Definition 2.2
The attribute X_j is *difference independent* of the remaining attributes if the preference difference between two levels of X_j is not affected by the fixed levels on the other attributes.

If these conditions are not satisfied, other forms such as a multiplicative aggregation model can be used (Keeney, 1974).

Ranking and Selection: The alternatives are ranked based on their aggregated scores. Specifically, for any two alternatives a_k and a_l, the preference relation is expressed as follows:

$$\begin{cases} v(a_k) > v(a_l) \Leftrightarrow a_k \succ a_l, & a_k \text{ is strongly preferred to } a_l \\ v(a_k) = v(a_l) \Leftrightarrow a_k \sim a_l, & a_k \text{ is indifferent to } a_l \\ v(a_k) \geq v(a_l) \Leftrightarrow a_k \succsim a_l, & a_k \text{ is weakly preferred to } a_l \end{cases} \tag{2.2}$$

Among the five steps, this study focuses on the value function elicitation step and examines how anchoring bias can distort value function elicitation and, in turn, distort the evaluation of alternatives. Value functions transform the performance of alternatives into a standardized scale, allowing for consistent comparison across different attributes. In MAVT, both the value function and attribute weights influence the final decision. However, to isolate the impact of value function elicitation, this study fixes the attribute weights, allowing for a more focused examination of how the elicitation process affects decision results. A detailed explanation of the experiment design used is provided in Section 2.5.

The midvalue splitting procedure, commonly used in the value function elicitation step, relies on defining a monotonic value function that represents the DM's preferences. When a higher attribute value is preferred, an increasing function is applied. First, a formal definition (Kirkwood & Sarin, 1980):

Definition 2.3
x_1 is said to be the midvalue of the interval $[x_0, x_2]$ if the decision-maker will give up the same amount of some other attribute to go from x_0 to x_1 as from x_1 to x_2.

The procedure involves the following steps: (i) *Assigning Initial Values*: The analyst assigns a value score of 0 to the lowest attribute level x_{lowest} and a value score of 1 to the highest attribute level x_{highest}. (ii) *Determining the First Midvalue Point*: The analyst provides an initial point x_1 between x_{lowest} and x_{highest}. The DM is then asked whether they perceive the change from x_{lowest} to x_1 to be equally desirable as the change from x_1 to x_{highest} [1]. If the DM indicates indifference between these changes, x_1 is considered the first midvalue point and is assigned a value score of 0.5. If the DM is not indifferent between the changes, the analyst will propose an alternative point x_1, and this process will be repeated until the point of indifference is found. (iii) *Determining Subsequent Midvalue Points*: Using the same procedure, the DM defines the indifference point x_2 between x_{lowest} and x_1. Once x_2 is identified, it is assigned a value score of 0.25. Similarly, the indifference point x_3 is defined between x_1 and x_{highest}, which is then assigned a value score of 0.75. (iv) *Consistency Check*: We check if the DM is indifferent between the two changes from x_2 to x_1 and from x_1 to x_3. If not, the DM will be asked to revise x_1. (v) *Plotting the Value Function*: After obtaining the midvalue points (x_1, x_2, x_3), the value function can be plotted, providing a clear graphical representation of the DM's preferences across the attribute range. In cases when a lower attribute value is preferred, a decreasing monotonic function is used. The value score of 1 is assigned to the lowest attribute level x_{lowest}, and 0 is assigned to the highest attribute level x_{highest}, ensuring the function accurately represents inverse preferences. The remaining steps are similar to the case of an increasing function. Notably, more midvalue points can be identified if we want to ensure a more detailed presentation of preference using the value function. It is also important to note that during this procedure, the levels of the other attributes are at a specified fixed level, and mutual preferential independence is given.

2.4 Hypotheses Development

The midvalue splitting procedure, as outlined in Section 2.3, begins with the analyst asking the DM whether she/he agrees with the midvalue point provided by the analyst. If this starting point is at the lower end of the attribute range, regardless of a monotonic increasing or decreasing value function, it is called a low anchor, whereas a starting point at the upper end is called a high anchor. When the DM is provided with a low anchor by the analyst, the insufficient adjustment from the anchor by the DM could lead to the first midvalue point being smaller than that of when a high anchor is given ($m_L < m_H$). The analyst's selection of the starting point can be explained by his/her own judgment of the shape of the value function, a random choice, or other reasons. It is clear that using a low versus a high starting point can affect determining the midvalue point. Similarly, when the analyst uses the mid-performance point ($(x_{lowest} + x_{highest})/2$) as the starting anchor, anchoring bias may still lead to distortions. For a linear value function, the midvalue point is exactly the mid-performance point, but people might adjust away from the anchor (the mid-performance point), and thus distort the elicited midvalue point. For non-linear value functions, the mid-performance point is acting as the low or high anchor depending on the shape of the value function. Therefore, the same arguments for the low and high anchors apply here.

The initial deviation might not only affect the first midvalue point, but also is expected to

[1] Given that the additive assumption is validated, it is easy and natural to ignore the changes in values of the other attributes when answering these questions (Smith & Dyer, 2021).

spread to subsequent midvalue points, ultimately changing the shape of the attribute-specific value function. While the consistency check in the midvalue splitting procedure is intended to ensure consistent preferences, the systematic deviation across all midvalue points undermines its effectiveness in identifying such errors. We define such consistency as intra-consistency, meaning that the DM maintains internal consistency across all midvalue points within a single value function. However, the DM may produce different intra-consistent value functions when presented with different anchors. We call such differences a violation of inter-consistency. While there is no objective "true" value function against which to compare outcomes since preferences are constructed rather than preexisting (Slovic, 1995), the pattern of violating inter-consistency in value function across different anchors offers meaningful insights into the effects of anchoring bias.

For monotonically decreasing attribute-specific value functions, as illustrated in Figure 2.1, anchoring bias can alter their shapes in various ways. Specifically, when a high anchor is provided, the resulting value function could be different compared to a low anchor as follows: (a) a concave shape to a convex shape; (b) a concave shape to a linear shape; (c) an extreme concave shape to a concave shape; (d) a linear shape to a convex shape; and (e) a convex shape to an extreme convex shape. In all cases, the area under curve (AUC) for the value function resulted from a low anchor is *smaller* than that of a high anchor. For the monotonically increasing attribute-specific value functions, we could think of five similar situations (convex to concave, convex to linear, extreme convex to convex, linear to concave, and concave to extreme concave), for all of which the AUC for the value function resulted from a low anchor is *larger* than that of a high anchor. These shifts indicate the degree to which judgmental inter-inconsistencies arise due to anchoring. Therefore, we propose the following hypothesis:

H1: A low anchor, compared to a high anchor, for the first midvalue point, results in smaller midvalue points and smaller (larger) area under curve for a decreasing (increasing) attribute-specific value function.

To mitigate the effect of anchoring bias, two widely used approaches are (i) avoiding anchors, and (ii) employing counter-anchors. In the context of the midvalue splitting procedure, where the starting point often serves as an anchor, these strategies can be applied in different ways. One approach is to avoid the starting point entirely, allowing the DM to express midvalue points without any pre-specified/suggested midvalue by the analyst. Alternatively, counter-anchoring can be implemented by introducing both a low and a high starting point sequentially, in either a low-high or high-low order, with the intention of balancing the influence of each anchor.

When no-anchor or counter-anchoring is employed, the expectation is that the midvalue points will fall between the low and high anchor midvalue points. As a result, the shape of the attribute-specific value function is anticipated to lie between the low and high anchor attribute-specific value functions. Based on this rationale, we hypothesize the following:

H2: No-anchor and counter-anchoring reduce the impact of anchoring bias, leading to an attribute-specific value function with an AUC in between the AUC of low and high anchor attribute-specific value functions.

In order to investigate the impact of anchoring bias on the overall value of alternatives,

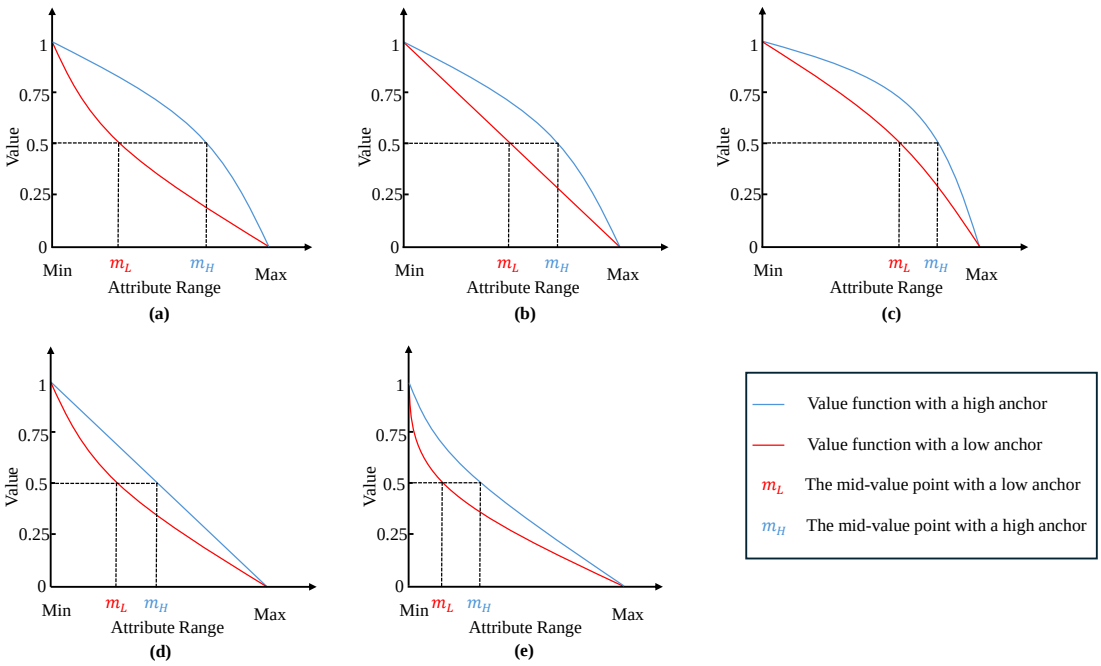


Figure 2.1: The possible effect of low and high anchors on value function shape

we develop a third hypothesis. The overall value of an alternative is calculated through the aggregation of weights and attribute-specific value functions (see equation 5.1). To isolate the anchoring effect of attribute-specific value function on the overall value, we control the weights through a carefully designed experimental setup (see Section 2.5).

For a monotonically decreasing value function, the value (v_L) for an attribute level derived from the low-anchored value function will be smaller than the value, v_H, derived from a high-anchored value function. Conversely, for a monotonically increasing value function, the value obtained from the low-anchored value function will be larger than the value obtained from a high-anchored value function. This relationship is illustrated in Figure 2.2.

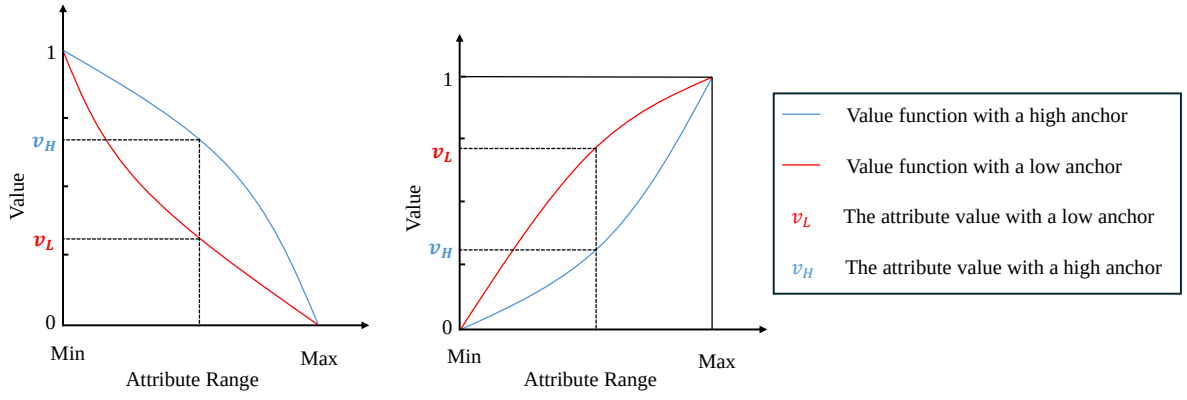


Figure 2.2: An example of possible changes in value under high vs. low anchoring

Considering a situation with two attributes and fixed weights, if both attributes are monotonically decreasing and have low-anchored value functions, the alternative may receive a lower

overall value compared to when both are high-anchored. Conversely, if both attributes are monotonically increasing and low-anchored, the overall value of an alternative may increase compared to the high-anchored scenario. For cases where one attribute is increasing and the other is decreasing, the opposing effects of anchoring may partially cancel each other out when both attributes are anchored in the same direction, making it challenging to draw a straightforward conclusion.

This reasoning can be extended to scenarios with more than two attributes, though the complexity significantly increases and experimental analysis becomes more difficult. In the next section, a carefully designed experiment (both attributes are monotonically decreasing, with fixed weights) is used to test the effect of changes in the attribute-specific value functions on the overall value of the alternatives. So, while it is clear how each attribute-specific value function contributes to the overall value, when it comes to experimental analysis, it is more convenient to design situations where the contributions are in the same direction. Specifically, we hypothesize that:

H3: A low anchor, compared to a high anchor, for the first midvalue point of a decreasing (increasing) attribute-specific value function will contribute to a decrease (increase) in the overall value of an alternative.

2.5 Experiment Design

2.5.1 The Experiment Overview

The experiment was designed following the MAVT steps to test the hypotheses. Before distributing this questionnaire, two pilot studies were conducted to refine the questions and instructions in the questionnaire. Participants were presented with a structured questionnaire including five parts.

Providing Informed Consent

Participants were first informed about the study's purpose, procedures, potential risks, and benefits. They were then asked to voluntarily provide their consent to participate by signing an informed consent form, acknowledging their understanding of the study and their rights as participants.

Presenting a Hypothetical Decision Problem

We designed an apartment renting decision problem with two attributes: monthly rent and commute distance, making it relatable for most participants. This case also provides measurable attributes. Additionally, choosing these two monotonically decreasing attributes helps to test the third hypothesis (as explained in Section 2.4). The decision problem is presented to the participants as follows:

"Suppose you find yourself at the end of your current rental contract and are actively searching for a new apartment. The rental agency has provided you with a list of options, all of which share identical features: each apartment is a 40 square meter studio with a standard 2-year lease renewal term, the layout, amenities and other features are also the same. However, the

monthly rent and commuting distance differ for each option.

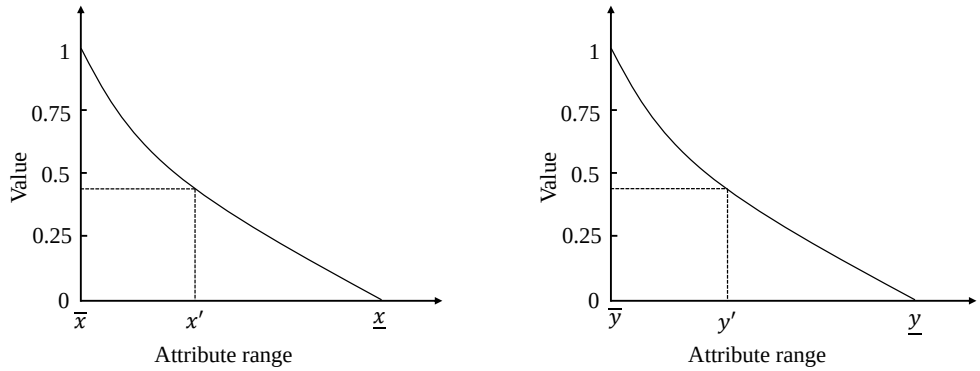
Monthly rent (euro): It is the amount of money one has to pay each month to rent the apartment. Commute distance (kilometer): This is the proximity of the apartment to your workplace."

The attribute range for 'Rent' has been set as [700, 1500], reflecting current rental market conditions in the countries of the experiment (Eurostat, 2023; Statistics Netherlands, 2024). This range ensures that the values used in the study are realistic and relevant to participants' experiences in the housing market. Similarly, the attribute range for 'Commute Distance' is defined as [5, 20] kilometers. This range is chosen to minimize the possibility of a non-monotonic value function, which could occur if DMs have personal preferences for certain commute distances. For example, some individuals may prefer a moderate commute distance rather than living too close to or too far from their workplace (Redmond & Mokhtarian, 2001). By setting these ranges, the study aims to capture realistic decision-making behaviors while ensuring a structured and logical approach to value function elicitation.

Eliciting and Fixing the Weights

This study examines how the effects observed during the value function elicitation step ultimately influence the decision results. In the additive aggregation model, the overall value of an alternative is determined by both the weights and the attribute-specific value functions (see equation 5.1). To isolate the impact in the value function elicitation step and ensure the results are unaffected by variations in the weight elicitation step, we first elicit and fix the weights. We achieve this by altering the attribute ranges used to elicit the attribute-specific value functions. Specifically, the procedure works as follows.

Let X and Y be two decreasing attributes[2] with predefined ranges: $X \in [\bar{x}, \underline{x}]$, where $\bar{x}$ is the minimum (best) performance score and $\underline{x}$ is the maximum (worst) performance score of attribute X. $Y \in [\bar{y}, \underline{y}]$, where $\bar{y}$ is the minimum (best) performance score and $\underline{y}$ is the maximum (worst) performance score of attribute Y. Figure 2.3 presents the hypothetical value functions of attribute X and Y.



(a) Value Function for Attribute X **(b) Value Function for Attribute Y**

Figure 2.3: Value functions for attributes X and Y

Step 1: We provide two alternatives $a_1 \equiv (\underline{x}, \bar{y})$ and $a_2 \equiv (\bar{x}, \underline{y})$, and ask if the DM

[2]In this paper, we use $\underline{x}$ and $\bar{x}$ to show the worst and best performance scores of attribute X. For a decreasing attribute, then the range is shown as $X \in [\bar{x}, \underline{x}]$, while for an increasing attribute the range is shown as $X \in [\underline{x}, \bar{x}]$.

equally prefers them. Note that $X \in [\overline{x}, \underline{x}], Y \in [\overline{y}, \underline{y}]$.

If the DM is indifferent between the two alternatives, following equation 2.2, it implies $w_X v(\underline{x}) + w_Y v(\overline{y}) = w_X v(\overline{x}) + w_Y v(\underline{y})$. Since $v(\underline{x}) = v(\underline{y}) = 0, v(\overline{x}) = v(\overline{y}) = 1$, we get $w_Y = w_X$. Given $w_X + w_Y = 1$, we derive $w_X = w_Y = 0.5$ and proceed to step 2.

If the DM prefers a_2 to a_1, it indicates $w_X > w_Y$ when $X \in [\overline{x}, \underline{x}], Y \in [\overline{y}, \underline{y}]$. We then ask the DM to replace $\bar{x}$ by x', where $x' \in (\bar{x}, \underline{x})$, such that he/she equally prefers the two alternatives. Later, we change the range of attribute X to $[x', \underline{x}]$ when eliciting the value function. Since the full range of attribute X is now $[x', \underline{x}]$, it means $v(x') = v(\bar{y}) = 1$. Therefore, the weights of attributes X and Y are still 0.5 when $X \in [x', \underline{x}], Y \in [\overline{y}, \underline{y}]$.

Conversely, if the DM prefers a_1 to a_2, it implies $w_X < w_Y$ when $X \in [\overline{x}, \underline{x}], Y \in [\overline{y}, \underline{y}]$. We ask the DM to replace $\bar{y}$ by y', where $y' \in (\bar{y}, \underline{y})$, so that he/she equally prefers the two alternatives. We then adjust the range of attribute Y to $[y', \underline{y}]$, when eliciting the value function, maintaining $w_X = w_Y = 0.5$ when $X \in [\overline{x}, \underline{x}], Y \in [y', \underline{y}]$.

Table 2.1: Scenarios of different attribute ranges for fixing weights

Scenario	Attribute X	Attribute Y
$w_X = w_Y$	$[\overline{x}, \underline{x}]$	$[\overline{y}, \underline{y}]$
$w_X > w_Y$	$[x', \underline{x}]$	$[\overline{y}, \underline{y}]$
$w_X < w_Y$	$[\overline{x}, \underline{x}]$	$[y', \underline{y}]$

For now, we elicited and fixed the weights of the two attributes to 0.5 by changing the attribute ranges (see Table 2.1). Consequently, in doing it, the overall values of a_1 and a_2 are also fixed to 0.5. To test Hypothesis 3 regarding the effect on the overall value of an alternative, we need another alternative with attribute levels not at the extreme levels of the attribute range.[3]

Step 2: For participants in the first and last scenarios (see Table 2.1), we provide a_3 with a random value within the attribute value range of X, like $\dot{x}$ and ask the DM to choose a value for Y like $\dot{y}$ such that he/she equally prefers the three alternatives. For participants in the second scenario, we provide a_3 with a random value within the attribute value range of Y, like $\dot{y}$ and ask the DM to choose a value for X like $\dot{x}$ such that he/she equally prefers the three alternatives.

For the first scenario $w_X = w_Y$, when $X \in [\overline{x}, \underline{x}], Y \in [\overline{y}, \underline{y}]$:

$$
\begin{aligned}
v(a_1) &= w_X v(\underline{x}) + w_Y v(\overline{y}) = w_Y \\
v(a_2) &= w_X v(\overline{x}) + w_Y v(\underline{y}) = w_X \\
v(a_3) &= w_X v(\dot{x}) + w_Y v(\dot{y})
\end{aligned}
$$

For the second scenario $w_X > w_Y$, when $X \in [x', \underline{x}], Y \in [\overline{y}, \underline{y}]$:

$$
\begin{aligned}
v(a_1) &= w_X v(\underline{x}) + w_Y v(\overline{y}) = w_Y \\
v(a_2) &= w_X v(x') + w_Y v(\underline{y}) = w_X v(x') \\
v(a_3) &= w_X v(\dot{x}) + w_Y v(\dot{y})
\end{aligned}
$$

[3]The additive assumptions for the aggregation model were verified between attributes through relevant questions in the questionnaire.

For the third scenario $w_X < w_Y$, when $X \in [\overline{x}, \underline{x}], Y \in [y', \underline{y}]$:

$$v(a_1) = w_X v(\underline{x}) + w_Y v(y') = w_Y v(y')$$
$$v(a_2) = w_X v(\overline{x}) + w_Y v(\underline{y}) = w_X$$
$$v(a_3) = w_X v(\dot{x}) + w_Y v(\dot{y})$$

In all scenarios, a_1 and a_2 are constructed using extreme attribute levels from the corresponding attribute ranges. The values of these extreme attribute levels are always 0 (for the worst) or 1 (for the best), regardless of any changes in the shape of the attribute-specific value functions. Therefore, the overall values of a_1 and a_2 remain fixed at 0.5 based on the calculation above. We design a_3 in a way that its attribute levels $\dot{x}$ and $\dot{y}$ are within the attribute ranges, thus any changes in the shape of the attribute-specific value functions due to anchoring bias can affect the values of $\dot{x}$ and $\dot{y}$. After eliciting the value functions for attributes X and Y in different experiment groups, we can calculate a new overall value of a_3, and **H3** can be tested by comparing $v(a_3)$ in different anchor groups. The equal preference among three alternatives serves as a benchmark, enabling us to attribute any deviations in $v(a_3)$ to anchoring bias. The experimental design establishes that $v(a_3)$ should theoretically be lower than 0.5 for the low anchor group, higher than 0.5 for the high anchor group, and equal to 0.5 for the three mitigation groups. This setup is intentionally structured to test the anchoring effect on decision-making outcomes, as well as the effectiveness of the mitigation strategies in neutralizing this bias.

Determining the Value Functions Using the Midvalue Splitting Procedure

In this part, participants are presented with a toy example to help them understand the midvalue splitting procedure. Then they are randomly assigned to one of five groups: low anchor group, high anchor group, low-high counter-anchor group, high-low counter-anchor group and no-anchor group. The first two groups are to test the anchoring bias and the last three groups are to test the debiasing strategies. To ensure consistency, all anchored groups receive two consecutive anchors spaced at 20% of the total attribute range. See Table 2.2 for the detailed information on anchors given in each group. A between-subject design is essential for this study, which ensures that the impact of each anchor group can be clearly isolated and measured without interference from prior conditions (Shadish et al., 2002). This design eliminates the risk that participants may adjust their responses based on prior exposure to a different anchor (e.g., shifting from a low to high anchor or vice versa), which could distort the observed effects of the anchors.

Table 2.2: Anchor values in different groups

Groups	**Starting Points Rent (euro)**	**Starting Points Commute Distance (km)**
Low Anchor	860 and 1020	8 and 11
No Anchor	none	none
Low-High Counter-Anchor	860 and 1340	8 and 17
High-Low Counter-Anchor	1340 and 860	17 and 8
High Anchor	1340 and 1180	17 and 14

Take the rent attribute value function elicitation in the high anchor group as an example; to identify the first midvalue point, the participants are presented with two possible rent drops:

"*Suppose you can get a lower rent for the apartment by increasing the commute distance. Suppose the drop in monthly rent would be either from 1500 euros to 1340 euros or from 1340 euros to 700 euros. For which drop in price would you accept a larger increase in commute distance?*" Here, 1340 euros serves as the high anchor. Regardless of the participants' answer, in the second question, we will replace 1340 euros with another high value, 1180 euros, to the DMs and ask if they are indifferent between the two latest rent drops. After that, participants will directly give a value for x_1 such that they are indifferent between the rent drops from 1500 euros to x_1 euros, and from x_1 euros to 700 euros. This value of x_1 becomes the first midvalue point, and is assigned a value score of 0.5. We understand that in a real-world scenario, an analyst will adjust the values in their following questions based on the DM's response and will continue asking questions until a midvalue point is reached. But for this experiment, we focus on the effect of anchors, and we stop after two questions for the sake of convenience and consistency (in terms of the number of adjustments) across all participants.[4]

This procedure is consistent across all groups, with only the starting points differing. Specifically, in the no-anchor group, no starting point is given to the participants, they directly give the first midvalue point. For the subsequent midvalue points, for all five groups, each subject is asked directly to give the points. After obtaining all three midvalue points, a consistency check is done by asking if the participants are indifferent between the change from $x_{0.25}$ to $x_{0.5}$ and from $x_{0.5}$ to $x_{0.75}$. If they are not indifferent between the two changes, then they will give a new score to $x_{0.5}$, and this new score becomes the first midvalue point.

Collecting Demographic Information

The final section of the questionnaire gathers demographic data such as age, gender, and education level. Additionally, information about the participant's current living space size, monthly rent or mortgage and commuting distance were also collected as control variables. This information provides a more comprehensive profile of the participants (Hammer, 2011; Creusen, 2010).

To implement the experiment in an online survey, we used the advanced and user-friendly Qualtrics platform, which allows for dynamic question flows, flexible layouts, and a variety of question types (Couper, 2000; Dillman et al., 2014). This platform enhances participant interaction by providing a seamless and intuitive survey experience.

2.5.2 Participants

Participants were recruited from six European countries: the Netherlands, Germany, France, Belgium, Denmark, and Luxembourg. These countries were chosen because they share similar cultures and apartment renting situations (Eurostat, 2023). Data collection was conducted using the online platform Prolific, which offers a large, diverse, and reliable participant pool (Palan & Schitter, 2018). Prolific's features for pre-screening and response verification enhance the quality of the data. The pre-screening function allowed us to limit participants based on their nationality and level of English proficiency. Additionally, the response verification process enabled us to reject incomplete answers or those that failed the midvalue splitting procedure.

[4]See for an example, Appendix A, the part of the questionnaire related to midvalue splitting procedure for participants in the high anchor condition of the first scenario.

Participants received a small monetary incentive for their participation, which has been shown to improve response rates, enhance the quality of responses, and accelerate data collection in research studies (Singer & Ye, 2013). This study was not incentivized in the traditional sense of providing performance-based rewards. While economists often use monetary incentives, arguing that they can elicit more realistic and effortful responses from participants, this is not universally accepted as the best practice, especially in psychology and behavioral decision-making fields (Hertwig & Ortmann, 2001). Hascher et al. (2021) suggested that the use of incentives in low-stakes valuation tasks may not be necessary and could potentially be counterproductive. The primary aim of our study is to investigate anchoring effects and value function elicitation within a controlled experimental setup. Introducing performance-based incentives could have influenced participants' responses, potentially leading to strategic behavior rather than authentic expressions of their preferences (Camerer & Hogarth, 1999).

In our study, participants spent an average of 16 minutes and 38 seconds, with a standard deviation of 8 minutes and 59 seconds, completing the experiment. This time includes participation in two experiments in one questionnaire, one of them being the present study. A total of 440 participants were recruited. After data cleaning, 36 responses were excluded because they did not complete the questionnaire. Additionally, in order to maintain logical consistency, 84 participants were excluded from the analysis due to failing the midvalue splitting procedure, as they provided values outside the specified ranges or identical values for all three midvalue points. This indicated either inattention to the questionnaire or a lack of understanding of the questions. This highlights a potential drawback of online questionnaires, where the lack of direct interaction may lead to misunderstandings or errors in following the procedures, resulting in inaccurate responses. Additionally, there were 19 participants who failed to give qualified answers when defining the initial indifference ranking, but all other responses adhered to the task logic and value ranges. This suggests they understood the instructions but may have momentary lapses on this specific question. Therefore, they were given a second chance to answer the initial ranking question and they provided qualified answers.

The final sample included 320 participants, whose data were used for the statistical analysis. Table 2.3 presents demographic information for the participants. Ethical approval for the study was obtained from the Institutional Review Board (IRB) of Delft University of Technology.

Table 2.3: Demographic characteristics of participants ($n = 320$)

Characteristics	Levels	Percent
Gender	Male	64%
	Female	35%
	Other	1%
Age	[18,24]	25%
	[25,34]	50%
	[35,44]	14%
	> 44	11%
Education	High School	14%
	Bachelor's degree	32%
	Master's degree	36%
	Other	18%

2.6 Results and Discussion

This section presents the findings from the experimental analysis designed to test the influence of anchoring bias on value function elicitation within the context of MAVT. The results are structured around the hypotheses, both parametric and non-parametric statistical methods were used.

The non-parametric tests identify consistent directional shifts caused by anchoring bias, irrespective of the scale or distribution of the data. Parametric tests quantify the magnitude of the difference between groups, offering insight into the size and practical significance of the anchoring effect. Together, the two distinct but complementary tests provide converging evidence of anchoring bias, demonstrating both its systematic influence on DM judgments and the extent of the effect. Notably, we conducted ANCOVA tests to examine whether subjects' current living space, housing costs, commuting distance, and demographics influenced the main results of anchoring bias in this study. The results indicated that none of these variables had a statistically significant effect on the primary dependent variables (all $p > 0.05$).

2.6.1 Hypothesis 1: Impact of Low vs. High Anchors on Midvalue Points and Value Function Shape

Hypothesis 1 posited that low starting points would result in smaller midvalue points and smaller (larger) area under the curve (AUC) compared to high starting points for decreasing (increasing) attribute-specific value functions. To test this, two key indicators were analyzed: the first midvalue point and the AUC.

The first midvalue point x_1 was selected because it represents the initial response most directly influenced by the starting point. A significant difference in this point would suggest that DM's initial adjustments were affected by anchoring bias. Notably, the attribute ranges during the value function elicitation step may vary across participants due to the experimental setup. To enable meaningful comparisons and analysis, we normalized the attribute ranges, which also adjusted the scores of the midvalue points accordingly.

The AUC, a measure of the total area under the curve relative to a reference axis, was used to quantify the overall shape of the value function. AUC was chosen over alternative indicators such as the sign of the second derivative for a value function, because AUC captures the nuanced changes in value function shape between groups. Specifically, while for a decreasing value function, a high anchor (compared to a low anchor) may shift a value function from extreme convex to convex, such changes cannot be fully captured by merely categorizing functions as convex or concave. AUC thus provides a more meaningful representation of these shifts and was recognized as the most appropriate measure for this study.

Rent Attribute

For the rent attribute, descriptive statistics showed that participants in the low anchor group produced lower midvalue points and smaller AUC values compared to those in the high anchor group, as detailed in Table 2.4.

An independent-samples t-test was conducted to compare the first midvalue point for rent

Table 2.4: Midvalue point and AUC for low and high anchor groups (Rent)

Anchor Group	Normalized First Midvalue Point Mean (SD)	AUC Value Mean (SD)
Low	0.42(0.11)	0.45(0.07)
High	0.49(0.12)	0.49(0.08)

between the low and high anchor groups. Levene's test for equality of variances was not significant, $F(1, 139) = 1.721$, $p = 0.192$, indicating that the assumption of equal variances was met. The results showed a statistically significant difference in the first midvalue point between the two groups, $t(139) = -3.496$, $p < 0.001$, confirming that participants in the low anchor group consistently provided lower midvalue points compared to the high anchor group. The effect size (Cohen's d) was 0.589, representing a medium effect magnitude. To further explore these results, a non-parametric Mann-Whitney U test was performed. This test, which compares the ranks rather than means of the midvalue points, also revealed a significant difference between the two groups, $Z = -3.586$, $p < 0.001$. The Mann-Whitney U test further confirmed that participants in the low anchor group consistently identify their midvalue points lower than those in the high anchor group, providing additional evidence of a systematic directional shift caused by anchoring bias.

The anchoring effect on the shape of the value function was examined using the AUC values. The independent-samples t-test for AUC values revealed that the low anchor group produced significantly lower AUC values, $t(139) = -3.276$, $p < 0.001$, with a corresponding medium effect size (Cohen's $d = 0.552$). The Mann-Whitney U test also indicated a significant difference in the ranks of AUC values between the two groups, $Z = -3.147$, $p = 0.002$, with the low anchor group having generally lower ranks compared to the high anchor group. These results imply that the influence of anchoring bias extends beyond the first midvalue point, affecting the overall shape of the value function for the Rent attribute.

Commute Distance Attribute

The descriptive statistics for the attribute commute distance showed a similar pattern: the low anchor group reveals lower first midvalue points and smaller AUC values than the high anchor group (see Table 2.5).

Table 2.5: Midvalue points and AUC for low and high starting point Groups (Commute Distance)

Anchor Group	Normalized First midvalue Point Mean (SD)	AUC Value Mean (SD)
Low	0.40(0.13)	0.43(0.08)
High	0.45(0.11)	0.45(0.07)

An independent-samples t-test was conducted to compare the first midvalue point for commute distance between the low and high anchor groups. Levene's test for equality of variances was not significant, $F(1, 139) = 1.574$, $p = 0.212$, indicating that the assumption of equal variances was satisfied. The t-test results showed a statistically significant difference in midvalue points between the low and high anchor groups, $t(139) = -2.504$, $p = 0.007$. The effect size (Cohen's d) was 0.422, representing a medium effect magnitude. Similarly, a Mann-Whitney U test further validated the significant difference between the distributions of midvalue points across the two groups, $Z = -2.605$, $p = 0.009$.

The t-test for AUC values showed significant difference ($t(139) = -2.271$, $p = 0.012$), with a medium effect size (Cohen's $d = 0.383$). The Mann-Whitney U test further corroborated the results, with $Z = -2.191$, $p = 0.028$. These results indicate that anchoring bias also affects the value function elicitation process for commute distance, though its effect is smaller than that of rent. This suggests that commute distance, as a non-monetary attribute, might introduce greater individual variability than rent, thus weakening the presence of bias. This aligns with evidence that non-monetary attributes reduce preference reversals compared to monetary attributes (Slovic et al., 1990).

2.6.2 Hypothesis 2: Effectiveness of Debiasing Strategies

Hypothesis 2 examines whether the use of no-anchor or counter-anchoring could mitigate the effects of anchoring bias. This hypothesis was tested by comparing the first midvalue points and AUC values generated by the low-high, high-low, and no-anchor groups with those from the low and high anchor groups, to see if the debiasing groups produced less extreme values compared to the low and high anchor groups.

Rent Attribute

According to Table 2.6, the first midvalue points and AUC values for the mitigation strategies (no-anchor, low-high, and high-low groups) fell between the extremes of the low and high anchor groups, suggesting that these strategies helped to mitigate the anchoring effect by reducing the reliance on one anchor.

Table 2.6: Midvalue points and AUC for five anchor groups (Rent)

	Low Anchor Mean (SD)	**No-Anchor Mean (SD)**	**Low-High Anchor Mean (SD)**	**High-Low Anchor Mean (SD)**	**High Anchor Mean (SD)**
Normalized First midvalue Point	0.42 (0.11)	0.43 (0.12)	0.45 (0.10)	0.46 (0.11)	0.49 (0.12)
AUC Value	0.45 (0.07)	0.45 (0.08)	0.46 (0.06)	0.47 (0.07)	0.49 (0.08)

In order to examine the anchoring effect across different conditions and to compare specific group differences, ANOVA test and post hoc analyses were used. ANOVA test for the first midvalue point showed significant differences between the five groups ($F(4, 315) = 3.705$, $p = 0.006$). Post-hoc comparisons identified two key findings regarding the effectiveness of debiasing strategies: the no-anchor group and low-high counter-anchor group produced significantly smaller midvalue points than the high anchor group ($p = 0.005$ and $p = 0.038$, respectively), and the high-low counter-anchor group resulted in significantly larger midvalue points than the low anchor group ($p = 0.047$). Importantly, there were no significant differences between the three debiasing strategy groups, indicating these strategies produce similar midvalue points.

For AUC values, the ANOVA test indicated significant differences between the five groups with ($F(4, 216) = 3.601$) at $p = 0.007$. Post-hoc analysis showed that the no-anchor group and low-high counter-anchor group had significantly smaller AUC values compared to the high anchor group ($p = 0.008$ and $p = 0.026$, respectively); the high-low counter-anchor group had significantly larger AUC values than the low anchor group ($p = 0.027$). Similar to the results for the first midvalue point, the AUC values for the three mitigation strategies were

non-significant, suggesting that they produce similar AUC values. This indicates that the three mitigation strategies can reduce anchoring bias to a similar extent.

Commute Distance Attribute

Table 2.7 presents the mean and standard deviation of the first midvalue points and the AUC values for the low and high anchor groups, alongside the three mitigation groups. Although the ANOVA for the first midvalue points ($F(4, 315) = 2.365$, $p = 0.053$) and for the AUC values ($F(4, 315) = 1.830$, $p = 0.123$) did not reach conventional levels of significance, these results do not imply that anchoring bias is absent in the low or high anchor groups. Rather, they suggest that the commute distance attribute is less susceptible to anchoring bias overall. The post hoc analysis revealed important findings for debiasing: the low-high counter-anchor group and high-low counter-anchor group produced significantly lower values than the high anchor group for both the first midvalue point and AUC values ($p = 0.008$, $p = 0.031$ for first midvalue point, respectively; $p = 0.036$ and $p = 0.037$ for AUC values, respectively). Besides, the three mitigation groups were not significantly different from each other, suggesting that they are equally effective in reducing anchoring bias.

Table 2.7: Midvalue points and AUC for five anchor groups (Commute Distance)

	Low Anchor Mean (SD)	**No Anchor Mean (SD)**	**Low-High Anchor Mean (SD)**	**High-Low Anchor Mean (SD)**	**High Anchor Mean (SD)**
Normalized First Midvalue Point	0.40 (0.13)	0.41 (0.13)	0.39 (0.11)	0.40 (0.11)	0.45 (0.11)
AUC Value	0.43 (0.08)	0.44 (0.08)	0.43 (0.07)	0.43 (0.07)	0.45 (0.07)

To better visualize these findings, we generated representative value functions (see Figure 2.4) for each experimental group by normalizing the attribute range and using the average midvalue points. For rent attribute, the representative value functions associated with the debiasing strategies generally fell between those observed in the low and high anchor conditions. The pattern for commute distance attribute was a bit different. That is, while the no-anchor condition produced value functions that lay intermediate to the low and high anchor groups, the counter-anchoring strategies did not consistently yield intermediate values. These results suggest that although anchoring bias is evident in the value function elicitation procedure for both attributes, and although the debiasing strategies are all effective for both attributes, the degree of effectiveness varies for the two attributes. Notably, while the normalization of the attribute ranges in Figure 2.4 visually compresses these differences, their practical impact remains non-trivial, especially considering the statistically significant differences and the medium effect sizes. The anchoring-induced shifts in midvalue points (7% for rent and 5% for commute distance between the low and high anchor groups) could significantly influence decision outcomes, especially in decisions involving critical factors such as financial commitments, safety, security, and health. In the next section, we will show how the decision results could change due to the changes in the value functions as a result of different anchors.

2.6.3 Hypothesis 3: Impact of Anchoring on the Overall Value of Alternatives

Hypothesis 3 suggests that the anchoring bias introduced in value function elicitation would also affect the overall value of alternatives and the proposed mitigation strategies would be

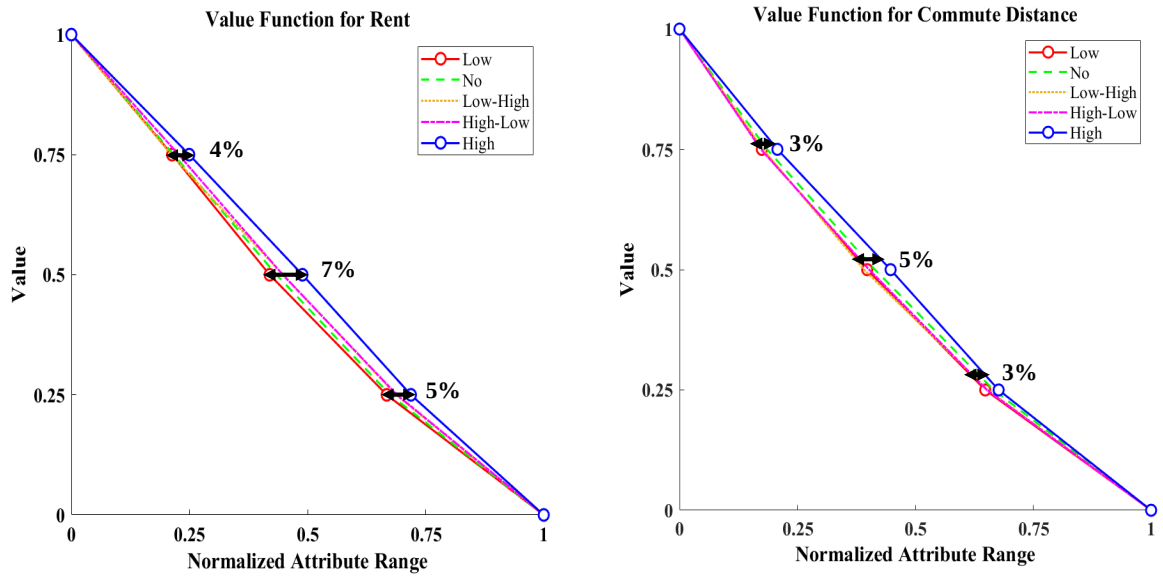


Figure 2.4: The representative value functions

effective in reducing this bias. Specifically, low anchors were expected to lead to lower overall values compared to high anchors for the same alternative in decreasing attribute-specific value functions. The overall value of alternative 3 was used to test this hypothesis.

As shown in Table 2.8, the group exposed to the low anchor produced the smallest overall value, while the high anchor group produced the largest overall value. The three mitigation strategies (no-anchor, low-high, high-low) resulted in overall values that fall between the extremes of the low and high anchor groups, aligning with the expectations of the hypothesis.

Table 2.8: Overall value for a_3 in different anchor groups

	Low Anchor Mean (SD)	**No-Anchor Mean (SD)**	**Low-High Anchor Mean (SD)**	**High-Low Anchor Mean (SD)**	**High Anchor Mean (SD)**
$v(a_3)$	0.52 (0.09)	0.54 (0.09)	0.54 (0.11)	0.56 (0.09)	0.59 (0.10)

ANOVA test results indicated significant differences among the five groups, $F(4, 315) = 4.805, p < 0.001$. Post-hoc comparisons revealed that the low anchor group provided significantly lower overall values compared to the high anchor group, with $p < 0.001$. Additionally, all three debiasing groups produced significantly lower overall values than the high anchor group ($p = 0.003$, $p = 0.005$ and $p = 0.039$, respectively). The high-low counter-anchor group produced a significantly higher value than the low anchor group ($p = 0.048$). Noticeably, no significant differences were observed among the three mitigation groups, suggesting that they have similar effects on the overall value.

Supporting these findings, the Kruskal-Wallis H test observed significant differences among the five groups ($\chi^2(4) = 22.217$, $p < 0.001$). Mann-Whitney U test for the overall value of a_3 revealed similar effects as with the ANOVA post hoc analysis results. All three debiasing groups produced significantly lower overall values than the high anchor group ($p < 0.001$, $p = 0.006$ and $p = 0.025$, respectively). The high-low counter-anchor group produced a significantly higher value than the low anchor group ($p = 0.024$). No significant differences were observed among the three mitigation strategies, which shows their mitigation effect is similar.

To examine how anchoring bias deviates participants from their initial equal preferences within each group, we tested if the overall value of a_3 is different from 0.5 across the five groups. One sample *t*-test results show that all groups are significantly above 0.5 ($p < 0.05$ for all), a result that initially seems counterintuitive. Ideally, one would expect the low anchor group to yield values below 0.5, the high anchor group to produce values above 0.5 and the debiasing groups to be closer to 0.5.

A closer inspection of individual-level data (see Table 2.9) provides important context beyond the group means. Although group means are all above 0.5, the distribution of responses shows systematic differences. Notably, for $v(a_3) < 0.5$, the low anchor group has the highest percentage of participants (33%), whereas the high anchor group has the lowest percentage (13%). The debiasing groups fall between the extremes. For $v(a_3) > 0.5$, the high anchor group shows the highest percentage (82%). In contrast, the low anchor group shows the lowest percentage (51%). The remaining groups fall in between. A Chi-square test indicated significant differences across the five groups, with $\chi^2(8, 320) = 19.392$ and $p = 0.013$. Further, Chi-square analysis among the group pairs showed that the low anchor group, no-anchor group and the low-high counter-anchor group are significantly different from the high anchor group ($p < 0.001$, $p = 0.015$ and $p = 0.038$, respectively).

Table 2.9: Distribution of individual-level overall values for a_3 across anchor groups

	$v(a_3)$			
	Below 0.5 (%)	**Equal to 0.5 (%)**	**Above 0.5 (%)**	**Total (%)**
Low Anchor	24 (33%)	12 (16%)	37 (51%)	73 (100%)
No-Anchor	12 (20%)	11 (18%)	37 (63%)	60 (100%)
Low-High Anchor	16 (28%)	6 (10%)	36 (62%)	58 (100%)
High-Low Anchor	13 (21%)	6 (10%)	42 (69%)	61 (100%)
High Anchor	9 (13%)	3 (4%)	56 (82%)	68 (100%)

In summary, although the overall means across groups are elevated (see Table 2.8), the individual-level analysis reveals clear, systematic differences that align with our theoretical expectations: low anchor leads to more values below 0.5, high anchor leads to more values above 0.5, and the debiasing groups produce values in between the extremes.

The experimental setting, where three alternatives were initially equally preferred, was designed to fix the weights and thereby isolate the effect of anchoring bias on the value function. In normal decision-making scenarios, where weights are not fixed, the effect of anchoring bias can be even more complex and pronounced due to its combined influence on both the weight elicitation (Rezaei et al., 2024) and value function elicitation processes. Furthermore, in real-world decision problems, even when the alternatives are not similar with respect to the attributes, the overall values of the alternatives are often close to each other, making the ranking highly sensitive to the inputs (i.e. value functions and weights). When the value functions or weights shift even slightly due to anchoring bias, the aggregated scores can cause significant and complex changes to the rankings.

2.6.4 Implications for Similar Procedures

While this study focuses on the midvalue splitting procedure, its implications can be extended to other value function elicitation methods as well. Below, we discuss three such methods and their susceptibility to anchoring bias. Future research could investigate the vulnerability of these methods to anchoring bias and examine the effectiveness of the proposed debiasing strategies.

Standard difference procedure (Beinat, 1997): This procedure divides the attribute range into equal value-spaced subintervals and plots the value function by solving a system of linear equations. When eliciting the value function for an increasing attribute X_j, the analyst will first define a value of $x_1 \in (\underline{x}, \bar{x})$ and then ask the DM to identify a value of $x_2 \in (x_1, \bar{x}]$ such that she/he is indifferent between the subintervals $(\underline{x}, x_1)$ and (x_1, x_2). The analyst continues to ask the DM to identify a value of x_N until $x_N = \bar{x}$. Consequently, the value function is defined using these indifference points, for example, $v_j(x_1) = \frac{v_j(x_N)}{N} = \frac{1}{N}$. Below, we discuss how anchoring bias could affect the value function elicited using this procedure.

During this process, the first indifference point x_1 set by the analyst can act as an anchor, distorting subsequent judgments in identifying the indifference points in two key ways: (i) Beginning this procedure at the lower versus upper end of the range can result in different value functions. When DMs identify subsequent indifference points $x_2, x_3, ..., x_N$, they might adjust insufficiently from x_1 due to anchoring bias. If the procedure starts at the lower end, the increments between indifference points may be smaller than they should be. Conversely, starting at the upper end can result in smaller decrements. Like the midvalue splitting procedure, this effect on the indifference points can lead to an overall effect on the shape of the value function due to the iterative structure. In case of an increasing value function, beginning this procedure at the lower end could then lead to a value function with a larger AUC compared to beginning at the upper end. (ii) Beginning the procedure with a smaller or a larger value of x_1 can result in different value functions. The identification of $x_2, x_3, ..., x_N$ can be affected by x_1 due to anchoring bias. Beginning this procedure with a smaller value of x_1 could lead to smaller increments between subsequent indifferent points compared to beginning this procedure at a larger value of x_1. In case of an increasing value function, this could also lead to a value function with a larger AUC when beginning this procedure with a smaller value of x_1. Building on the counter-anchoring strategy, debiasing could involve conducting the procedure first at the lower end and then repeating it at the upper end, or vice versa. By eliciting indifference points from both directions, the influence of any single starting point is reduced, and the final value function can be derived by averaging or synthesizing the results. Similarly, the procedure could be conducted using different starting values within the range to further mitigate anchoring effects.

Lock-step procedure (Keeney & Raiffa, 1976): This procedure constructs the attribute-specific value functions by iteratively defining indifference points between two attributes. Specifically, the worst outcomes for both attributes are defined at first, $v(\underline{x}, \underline{y}) = v_X(\underline{x}) = v_Y(\underline{y}) = 0$. Second, the analyst selects $x_1 > \underline{x}$ and sets $v_X(x_1) = 1$ to fix the "unit" of value for attribute X. Third, the analyst will ask the DM to identify y_1 such that she/he is indifferent between $(x_1, \underline{y})$ and $(\underline{x}, y_1)$, and define $v_Y(y_1) = 1$. This process continues until sufficient indifferent points (e.g., x_2, y_2, x_3, y_3,..., x_k, y_k) are defined to satisfy chains of indifference (e.g., $(x_2, \underline{y}) \sim (x_1, y_1) \sim (\underline{x}, y_2)$), with values incremented sequentially ($v_X(x_k) = v_Y(y_k) = k$). In this process, the analyst's selection of x_1 can also serve as an anchor, distorting subsequent judgments

in identifying the indifference points in a similar way as explained in the standard difference procedure. Both the starting point in the attribute range (lower vs. upper end) and the size of the initial value (x_1) affect the resulting value function. Counter-anchoring debiasing strategies, such as bidirectional elicitation and multiple starting values for x_1, can help mitigate these biases.

Direct rating (Fishburn, 1967): This procedure directly assigns 0 to the worst performance level and 100 to the best performance level and then asks the DM to assign scores to other performance levels. However, the order (rating from high-to-low performance level versus from low-to-high performance level) can also serve as anchors and affect the elicited values. For example, starting with high-performance levels may anchor DMs to overvalue subsequent mid-range levels. Building on the no-anchor strategy, debiasing could involve randomizing the presentation order. Building on the counter-anchoring strategy, debiasing could involve presenting performance levels in an alternating sequence, starting with the highest, then the lowest, followed by the second highest, then the second lowest, and so on.

In summary, these elicitation procedures share similar characteristics: the analyst's suggestions or presentation orders (anchors) that might affect the identification of indifference points or elicited values, and iterative procedures that can carry this effect to the value function. These common features allow the implications of this study to be extended to other value function elicitation methods.

2.7 Conclusion

In this study, we theorized and empirically tested the effect of anchoring bias on the midvalue splitting procedure, a common approach for value function elicitation, within the context of multi-attribute value theory (MAVT). We found that the starting point in the midvalue splitting procedure can serve as an anchor, influencing attribute-specific value functions. Consequently, this anchoring effect alters the overall value of the alternatives. The debiasing strategies tested in this study (no-anchor and counter-anchoring) effectively mitigate anchoring bias by producing less extreme results compared to the low and high anchor conditions. Moreover, the proposed mitigation strategies do not introduce additional bias when an attribute is less prone to bias, demonstrating their robustness.

The results of this study align with and extend the existing literature on anchoring bias and its impact on decision-making processes (Tversky & Kahneman, 1974; Englich et al., 2006). Our findings contribute to this field by showing its significant effects on the shape of elicited value functions and results in MAVT. In the context of MADM, while research has been primarily focused on the effect of anchoring bias in weight elicitation (Rezaei, 2021; Rezaei et al., 2024; Buchanan & Corner, 1997; Jacobi & Hobbs, 2007), this study provides empirical evidence in another essential part, the value function elicitation. The findings raise concerns about the reliability of commonly used elicitation procedures, as the characteristics of the value function directly affect the evaluation of the alternatives. Anchoring bias strongly influenced the overall value of alternatives even when only two attributes were considered. In real-world decision-making contexts, where the number of attributes and alternatives is typically much larger, the impact of anchoring bias can result in even stronger distortions in the evaluation process, potentially leading to systematically biased decisions.

The findings of this study carry important implications for analysts employing MAVT and similar methods to support decision-makers. First, analysts must recognize the susceptibility of these methods to anchoring bias in value function elicitation. To reduce such bias, analysts should remain neutral throughout all procedures. This includes avoiding the introduction of starting points that may unintentionally serve as anchors. When initial values or reference points are necessary, providing counter-anchors can help reduce their impact. Additionally, these insights extend beyond individual analysts to the design and implementation of decision-support systems that rely on value function elicitation. Such systems should be developed with mechanisms to counteract anchoring bias, ensuring that their structure and interfaces do not inadvertently introduce or reinforce biased reference points.

This study enhances our understanding of anchoring bias in value function elicitation and provides important implications for both theoretical and practical improvements of MAVT and other decision analysis methods. However, several limitations should be acknowledged. Despite refinements through two pilot experiments, some participants still failed the midvalue splitting procedure. This might suggest that the procedure imposes considerable cognitive demands on participants, potentially leading to fatigue and errors. The sample's cultural and geographical specificity may limit the generalizability of the findings (Norenzayan et al., 2007; Ma-Kellams, 2020). The simplified two-attribute decision problem may not fully capture real-world complexity. The degree of effectiveness of the debiasing strategies varies for the two attributes examined in this study, indicating different susceptibility to anchoring bias. These limitations highlight the need for replication studies to confirm the robustness of the observed effects.

Future research could build on these findings by examining the influence of anchoring bias in complex decision problems involving more than two attributes and different types of attribute value functions (e.g., a mix of increasing and decreasing value functions). Additionally, it would be valuable to investigate how anchoring bias affects other value function elicitation procedures and whether the proposed mitigation strategies remain effective in those contexts. We introduced some general ideas in Section 2.6.4, which future research can further develop and extend. Furthermore, conducting the experiment in a more controlled environment, such as a supervised laboratory setting, could help improve data quality by minimizing inattentive responses and enabling better identification of potentially invalid data, such as unusually fast responses on specific tasks. Finally, expanding the experiments to different cultural or geographical contexts could help generalize the study's findings.

Bibliography

Adame, B. J. (2016) Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy, *Learning and Motivation*, 53, pp. 36–48.

Anthes, R., M. Maier, S. Ackerman, R. Atlas, L. Callahan, G. Dittberner, J. Yoe, et al. (2019) Developing priority observational requirements from space using multi-attribute utility theory, *Bulletin of the American Meteorological Society*, 100(9), pp. 1753–1774.

Ariely, D. (2008) *Predictably irrational: the hidden forces that shape our decisions*, Harper Collins, New York.

Barber, B. M., T. Odean (2001) Boys will be boys: Gender, overconfidence, and common stock investment, *The Quarterly Journal of Economics*, 116(1), pp. 261–292.

Bazerman, M. H., D. A. Moore (2012) *Judgment in managerial decision making*, John Wiley & Sons, New Jersey.

Beinat, E. (1997) *Value functions for environmental management*, Springer, Dordrecht.

Belton, V., T. Stewart (2012) *Multiple criteria decision analysis: an integrated approach*, Springer, New York.

Brewer, N. T., G. B. Chapman, J. A. Schwartz, G. R. Bergus (2007) The influence of irrelevant anchors on the judgments and choices of doctors and patients, *Medical Decision Making*, 27(2), pp. 203–211.

Buchanan, J. T., J. Corner (1997) The effects of anchoring in interactive mcdm solution methods, *Computers & Operations Research*, 24(10), pp. 907–918.

Bushong, B., T. Gagnon-Bartsch (2024) Failures in forecasting: An experiment on interpersonal projection bias, *Management Science*, 70(12), pp. 8735–8752.

Camerer, C. F., R. M. Hogarth (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework, *Journal of Risk and Uncertainty*, 19(1), pp. 7–42.

Chapman, G. B., E. J. Johnson (2002) Incorporating the irrelevant: Anchors in judgments of belief and value, in: *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, New York, pp. 120–138.

Couper, M. P. (2000) Web surveys: A review of issues and approaches, *The Public Opinion Quarterly*, 64(4), pp. 464–494.

Creusen, M. E. (2010) The importance of product aspects in choice: the influence of demographic characteristics, *Journal of Consumer Marketing*, 27(1), pp. 26–34.

Dillman, D. A., J. D. Smyth, L. M. Christian (2014) *Internet, phone, mail, and mixed-mode surveys: The tailored design method*, John Wiley & Sons, New Jersey.

Dyer, J. S., R. K. Sarin (1979) Measurable multiattribute value functions, *Operations Research*, 27(4), pp. 810–822.

Englich, B., T. Mussweiler, F. Strack (2006) Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making, *Personality and Social Psychology Bulletin*, 32(2), pp. 188–200.

Epley, N., T. Gilovich (2005) When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors, *Journal of Behavioral Decision Making*, 18(3), pp. 199–212.

Eurostat (2023) Housing in europe, https://ec.europa.eu/eurostat/web/interactive-publications/housing-2023, accessed: 2023-10-12.

Fasolo, B., C. Heard, I. Scopelliti (2024) Mitigating cognitive bias to improve organizational decisions: An integrative review, framework, and research agenda, *Journal of Management*, 51(6), pp. 2182–2211.

Ferretti, V. (2016) From stakeholders analysis to cognitive mapping and multi-attribute value theory: An integrated approach for policy support, *European Journal of Operational Research*, 253(2), pp. 524–541.

Fischhoff, B. (1991) Value elicitation: is there anything in there?, *American Psychologist*, 46(8), pp. 835–847.

Fishburn, P. C. (1967) Methods of estimating additive utilities, *Management Science*, 13(7), pp. 435–453.

Furnham, A., H. C. Boo (2011) A literature review of the anchoring effect, *The Journal of Socio-Economics*, 40(1), pp. 35–42.

Galinsky, A. D., T. Mussweiler (2001) First offers as anchors: the role of perspective-taking and negotiator focus, *Journal of Personality and Social Psychology*, 81(4), pp. 657–669.

Gavirneni, S., Y. Xia (2009) Anchor selection and group dynamics in newsvendor decisions—a note, *Decision Analysis*, 6(2), pp. 87–97.

Hammer, C. S. (2011) The importance of participant demographics, *American Journal of Speech-Language Pathology*, 20(4), p. 261.

Hascher, J., N. Desai, I. Krajbich (2021) Incentivized and non-incentivized liking ratings outperformwillingness-to-pay in predicting choice, *Judgment and Decision Making*, 16(6), pp. 1464–1484.

Hertwig, R., A. Ortmann (2001) Experimental practices in economics: A methodological challenge for psychologists?, *Behavioral and Brain Sciences*, 24(3), pp. 383–403.

Höfer, T., R. Madlener (2020) A participatory stakeholder process for evaluating sustainable energy transition scenarios, *Energy Policy*, 139, 111277.

Hostmann, M., T. Bernauer, H. J. Mosler, P. Reichert, B. Truffer (2005) Multi-attribute value theory as a framework for conflict resolution in river rehabilitation, *Journal of Multi-Criteria Decision Analysis*, 13(2-3), pp. 91–102.

Jacobi, S. K., B. F. Hobbs (2007) Quantifying and mitigating the splitting bias and other value tree-induced weighting biases, *Decision Analysis*, 4(4), pp. 194–210.

Kahneman, D. (2011) *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Keeney, R. L. (1974) Multiplicative utility functions, *Operations Research*, 22(1), pp. 22–34.

Keeney, R. L. (2002) Common mistakes in making value trade-offs, *Operations Research*, 50(6), pp. 935–945.

Keeney, R. L. (2004) Making better decision makers, *Decision Analysis*, 1(4), pp. 193–204.

Keeney, R. L. (2009) *Value-focused thinking: A path to creative decisionmaking*, Harvard University Press, Cambridge.

Keeney, R. L., H. Raiffa (1976) *Decisions with multiple objectives: Preferences and value trade-offs*, Cambridge University Press, Cambridge.

Kim, S.-H., J. Tong (2024) Admission control bias and path-dependent feedback under diagnosis uncertainty, *Manufacturing & Service Operations Management*, 26(1), pp. 117–136.

Kirkwood, C. W., R. K. Sarin (1980) Preference conditions for multiattribute value functions, *Operations Research*, 28(1), pp. 225–232.

Labijak-Kowalska, A., M. Kadziński, I. Spychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, K. Farion (2023) Performance evaluation of emergency department physicians using robust value-based additive efficiency model, *International Transactions in Operational Research*, 30(1), pp. 503–544.

Lilienfeld, S. O., R. Ammirati, K. Landfield (2009) Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare?, *Perspectives on Psychological Science*, 4(4), pp. 390–398.

Lord, C. G., M. R. Lepper, E. Preston (1984) Considering the opposite: a corrective strategy for social judgment, *Journal of Personality and Social Psychology*, 47(6), pp. 1231–1243.

Ma-Kellams, C. (2020) Cultural variation and similarities in cognitive thinking styles versus judgment biases: A review of environmental factors and evolutionary forces, *Review of General Psychology*, 24(3), pp. 238–253.

Montibeller, G., D. von Winterfeldt (2015) Cognitive and motivational biases in decision and risk analysis, *Risk Analysis*, 35(7), pp. 1230–1251.

Montibeller, G., D. von Winterfeldt (2024) Behavioral decision research: Descriptive and prescriptive perspectives, in: *Behavioral Decision Analysis*, Springer, pp. 15–40.

Mussweiler, T., F. Strack (1999) Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model, *Journal of Experimental Social Psychology*, 35(2), pp. 136–164.

Mussweiler, T., F. Strack, T. Pfeiffer (2000) Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility, *Personality and Social Psychology Bulletin*, 26(9), pp. 1142–1150.

Norenzayan, A., I. Choi, K. Peng (2007) Perception and cognition, in: *Handbook of Cultural Psychology*, The Guilford Press, pp. 569–594.

Palan, S., C. Schitter (2018) Prolific.ac—a subject pool for online experiments, *Journal of Behavioral and Experimental Finance*, 17, pp. 22–27.

Payne, J. W., J. R. Bettman, E. J. Johnson (1992) Behavioral decision research: A constructive processing perspective, *Annual Review of Psychology*, 43(1), pp. 87–131.

Prava, V. R., R. T. Clemen, B. F. Hobbs, M. A. Kenney (2016) Partition dependence and carryover biases in subjective probability assessment surveys for continuous variables: model-based estimation and correction, *Decision Analysis*, 13(1), pp. 51–67.

Redmond, L. S., P. L. Mokhtarian (2001) The positive utility of the commute: Modeling ideal commute time and relative desired commute amount, *Transportation*, 28(2), pp. 179–205.

Rezaei, J. (2015) Best-worst multi-criteria decision-making method, *Omega*, 53, pp. 49–57.

Rezaei, J. (2021) Anchoring bias in eliciting attribute weights and values in multi-attribute decision-making, *Journal of Decision Systems*, 30(1), pp. 72–96.

Rezaei, J., A. Arab, M. Mehregan (2024) Analyzing anchoring bias in attribute weight elicitation of smart, swing, and best-worst method, *International Transactions in Operational Research*, 31(2), pp. 918–948.

Shadish, W., T. Cook, D. T. Campbell (2002) *Experimental and quasi-experimental designs for generalized causal inference*, Houghton Mifflin, Boston.

Singer, E., C. Ye (2013) The use and effects of incentives in surveys, *The Annals of the American Academy of Political and Social Science*, 645(1), pp. 112–141.

Slovic, P. (1995) The construction of preference, *American Psychologist*, 50(5), pp. 364–371.

Slovic, P., D. Griffin, A. Tversky (1990) Compatibility effects in judgment and choice, in: *Insights in decision making: A tribute to Hillel J. Einhorn*, The University of Chicago Press, Chicago, pp. 5–27.

Smith, J. E., J. S. Dyer (2021) On (measurable) multiattribute value functions: An expository argument, *Decision Analysis*, 18(4), pp. 247–256.

Statistics Netherlands (2024) Centraal bureau voor de statistiek, https://www.cbs.nl/, accessed: 2024-07-08.

Thaler, R. H., C. R. Sunstein (2008) *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale University Press, New Haven.

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

Tversky, A., D. Kahneman (1981) The framing of decisions and the psychology of choice, *Science*, 211(4481), pp. 453–458.

Wansink, B., R. J. Kent, S. J. Hoch (1998) An anchoring and adjustment model of purchase quantity decisions, *Journal of Marketing Research*, 35(1), pp. 71–81.

Watson, S. R., D. M. Buede (1987) *Decision synthesis: The principles and practice of decision analysis*, Cambridge University Press, Cambridge.

Wilson, T. D., C. E. Houston, K. M. Etling, N. Brekke (1996) A new look at anchoring effects: basic anchoring and its antecedents, *Journal of Experimental Psychology: General*, 125(4), pp. 387–402.

Chapter 3

Anchoring Bias in the Tradeoff Procedure within Multi-Attribute Value Theory

---

Abstract: Eliciting the weights of attributes is a key step in multi-attribute decision-making methods. The weights usually represent the relative importance of the attributes or the tradeoffs among them in forming a decision. Various weight elicitation methods exist, each based on different assumptions and procedures. Still, many of these methods do not explicitly account for the potential influence of cognitive biases in their design. This study examines the anchoring bias, a well-known cognitive bias, in the weight elicitation step (the Tradeoff procedure) of multi-attribute value theory (MAVT). We developed three hypotheses: (i) Using the most important (best) attribute to construct the indifference pairs in the Tradeoff procedure leads to higher weights for the best and worst attributes and lower weights for the other attributes; (ii) Using the least important (worst) attribute to construct the indifference pairs in the Tradeoff procedure leads to lower weights for the best and worst attributes and higher weights for the other attributes; (iii) Using both best and worst attributes to construct the indifference pairs (i.e., the Best-Worst Tradeoff: BWT), mitigates the anchoring bias. To test the hypotheses, we conducted an experiment by designing a questionnaire based on MAVT and collected data from 336 participants for a decision problem. The findings indicate that the anchoring bias has a significant impact on the Tradeoff procedure and that the BWT is effective in mitigating this bias.

Keywords: Anchoring bias; weight elicitation; tradeoff; best-worst tradeoff; debiasing

This chapter is based on the following manuscript:

Sun, G., Kroesen, M., Rezaei, J (2026) Anchoring Bias in the Tradeoff Procedure within Multi-Attribute Value Theory, *Journal of Behavioral Decision Making*, 39(2), e70069

3.1 Introduction

Many real-world decision problems are multi-attribute decision-making (MADM) problems, ranging from everyday life decisions to corporate or national governance problems, and in different fields such as healthcare, engineering, and finance. In MADM, the decision-maker (DM) evaluates multiple, often conflicting objectives. These objectives are represented by attributes, which provide measurable scales against which alternatives can be evaluated. Numerous MADM methods have been developed to support this process and to improve decision quality. One of the most important steps common to these methods is the elicitation of attribute weights. The weights usually represent the relative importance, contribution, or tradeoff among the attributes in forming a decision. These weights influence the final decision through various mechanisms, such as explicit aggregation, pairwise comparisons, or threshold-based rules. Consequently, if the weight elicitation process introduces biases or fails to capture the DM's preferences appropriately, the decision results may be misaligned with the DM's preferences. Therefore, ensuring that the elicited weights meaningfully reflect the DM's preferences is fundamental in supporting decision quality.

There are various methods for eliciting attribute weights, such as the Tradeoff procedure (Keeney & Raiffa, 1976), simple multi-attribute rating technique (SMART) (Edwards, 1977), Swing (von Winterfeldt & Edwards, 1993), analytic hierarchy process (AHP) (Saaty, 1980), and best-worst method (BWM) (Rezaei, 2015), among others. Each method has distinct assumptions and procedures to help the DM elicit their preferences and translate them into quantifiable weights. Consequently, different MADM problems can be solved by different methods depending on the context of the decision (Belton & Stewart, 2012). Importantly, the interpretation of attribute weights also varies across methods (Hämäläinen & Salo, 1997). As noted by Choo et al. (1999), in methods based on multi-attribute value theory (MAVT) (Keeney & Raiffa, 1976), such as the Tradeoff procedure, weights have a compensatory meaning: lower performance in one attribute can be offset by higher performance in another, depending on the weights. In contrast, methods such as AHP derive weights from pairwise comparisons and interpret weights as marginal contributions of each attribute to the overall value of an alternative, typically assuming a ratio scale. Understanding these differences is crucial, as it allows analysts to pose appropriate questions to the DM to elicit the relevant weights in different weight elicitation procedures.

These methods usually rely on the assumption that the DM is rational. In practice, however, humans are subject to cognitive biases that can systematically distort their judgments. As a result, even when these methods are implemented correctly, the elicited outputs may deviate from the normative expectations these methods are designed to reflect.

Cognitive bias is a systematic pattern of deviation from rationality when people process information. It arises from the use of mental shortcuts, or heuristics, in the decision-making process (Tversky & Kahneman, 1974; Gilovich et al., 2002). While heuristics can help the DM to arrive at a good enough decision quickly, they can also lead to systematic errors in judgments and decision-making. Research on cognitive biases has therefore been largely descriptive, seeking to identify new types of bias and to develop theories or models that explain these systematic deviations from rationality (Montibeller & von Winterfeldt, 2024). For instance, the anchoring bias is one of the most documented cognitive biases and has been identified in various areas (Tversky & Kahneman, 1974; Chapman & Johnson, 1999; Mussweiler & Strack, 2001). In the

medical context, Ly et al. (2023) demonstrated that pre-reviewing the document of a patient in the emergency department creates an anchoring bias in physicians' clinical decisions. In making a diagnosis, physicians can rely heavily on documentation and medical history (anchors), overlooking the possibility of other health conditions, leading to delays in proper examinations and diagnosis.

The insights from this descriptive line of research are highly relevant for MADM. They suggest that the fundamental assumptions (i.e., a rational DM) of MADM methods might not hold in practice, undermining the reliability of MADM methods. Consequently, research on cognitive bias in MADM has been primarily prescriptive, with an emphasis on developing debiasing strategies to account for these biases. The purpose of a prescriptive study, such as the present one, is not just to observe the biases but to improve the decision-making process by identifying and mitigating systematic deviations from normative decision rules. In doing so, prescriptive research provides actionable insights for MADM practitioners and contributes to improving the reliability of MADM methods.

Montibeller & von Winterfeldt (2015) and Morton & Fasolo (2009) have conducted reviews of the possible cognitive biases within the context of MADM and suggested mitigation strategies. Experimental studies have also been conducted to provide empirical evidence of cognitive bias or behavioral influences in MADM methods (von Nitzsch & Weber, 1993; Fischer et al., 1987; Fischer, 1995; Weber et al., 1988; Weber & Borcherding, 1993; Pöyhönen & Hämäläinen, 2000; Rezaei, 2021; Rezaei et al., 2022, 2024; Sun et al., 2025). Sun et al. (2025) investigated the anchoring bias in the value function elicitation step of MAVT, focusing on the midvalue splitting procedure. The present study complements and extends that work by shifting attention to the weight elicitation step, specifically the Tradeoff procedure. By examining another key stage of MAVT, this study deepens our understanding of how cognitive biases can arise at multiple stages of MADM methods, underscoring both the pervasiveness of such biases and the need for comprehensive debiasing strategies. Rezaei et al. (2024) examined the anchoring bias in the SMART and Swing methods. In those methods, the bias arises from explicit starting points, and their findings suggest that the BWM, which incorporates two directional anchors, can mitigate this effect. By contrast, the anchoring bias in the Tradeoff procedure does not stem from a starting point but from the selection of the reference attribute. This difference highlights that even within the same weight elicitation step of MAVT, distinct procedures can introduce bias through different mechanisms, requiring procedure-specific debiasing strategies. This study advances the literature by (i) theorizing and empirically demonstrating the anchoring bias in the Tradeoff procedure itself, and (ii) testing the Best–Worst Tradeoff (BWT) as a prescriptive debiasing strategy. While Liang et al. (2022) proposed that BWT might mitigate the anchoring bias in the Tradeoff procedure, empirical evidence has so far been lacking. By providing such evidence, the present study contributes to both the theoretical understanding of how cognitive bias influences different weight elicitation methods and the practical reliability of MADM applications.

This study investigates the effect of the anchoring bias in the weight elicitation step (the Tradeoff procedure) of MAVT. The Tradeoff procedure involves using attributes (best or worst) to construct indifference pairs and then deriving the weights (or scaling constants) from the indifference relations. We hypothesize that the selection of attributes (best or worst) might distort individual judgments due to the anchoring bias and lead to inconsistent weights across different conditions. We also hypothesize that the BWT method that utilizes both best and worst at-

tributes to form the indifference pairs can mitigate the anchoring bias in the Tradeoff procedure. To test these hypotheses, we designed a questionnaire based on the MAVT methodology. This questionnaire contained two decision problems: the first, with two attributes, was designed for the study reported in Sun et al. (2025) on value function elicitation; the second, with three attributes, was designed for the present study to investigate weight elicitation. Although both used the apartment selection context, the attribute ranges and experimental designs were different, ensuring independent datasets despite being combined into a single instrument for efficiency. For the current study, we employed a within-subject design to compare different Tradeoff procedures. We collected data from 336 participants and performed statistical analysis to test the hypotheses.

The remainder of this paper is structured as follows: Section 3.2 introduces the anchoring bias and its effects in MADM methods, especially in the weight elicitation step. Section 3.3 describes MAVT, the Tradeoff procedure, and the BWT. Section 3.4 develops three hypotheses for this study. Section 3.5 describes the experiment design. Section 3.6 presents the results and discussion of the study, and Section 3.7 concludes this study.

3.2 The Anchoring Bias and its Role in Multi-Attribute Decision Making

The anchoring bias was first introduced by Tversky & Kahneman (1974), explained by the anchoring-and-adjustment heuristic. According to this heuristic, individuals begin their judgment process with an anchor and then make adjustments from that anchor to arrive at a final judgment. However, the adjustments are insufficient, resulting in a final judgment that remains biased towards the anchor. This has been demonstrated in a range of experimental settings, either through estimation tasks (measured by deviation from the true value) or valuation tasks (measured by deviation from preference coherence and consistency). For example, in an estimation task, McElroy & Dowd (2007) asked participants to estimate the length of the Mississippi River after presenting them with either a low anchor (e.g., “more than 200 miles”) or a high anchor (e.g., “less than 20,000 miles”). Participants’ estimates were significantly biased in the direction of the anchor, with participants in the low anchor group providing significantly lower estimates (Mean=698.5) than those in the high anchor group (Mean=10,021.26).

Beyond the empirical evidence, the anchoring bias has been found to be impactful across a wide range of decision fields (Furnham & Boo, 2011). Accordingly, most field evidence concerns valuation tasks, in which anchoring is identified through systematic distortions in preference coherence and consistency rather than deviations from an objectively correct value. In legal settings, the anchoring bias can influence sentencing decisions made by legal professionals even when the anchor is the result of a dice roll (Englich et al., 2006), though factors such as legal expertise and the relevance of the anchor can moderate this effect (Bystranowski et al., 2021). In healthcare, the anchoring bias can influence diagnoses, where a physician’s initial hypothesis anchors subsequent evaluations, even when contradictory evidence emerges (Thirsk et al., 2022). In forecasting, Campbell & Sharpe (2009) found that professional economic forecasters rely heavily on recently realized values, leading to predictable forecast errors.

To reduce the effect of the anchoring bias, various mitigation strategies have been developed,

which can generally be grouped into three categories:

Encouraging critical thinking and reflection on the anchor's validity (Chapman & Johnson, 1999; Epley & Gilovich, 2005). This category includes strategies that encourage individuals to reflect on whether the anchor is appropriate. For instance, using incentives for accuracy can enhance individuals' motivation to make careful judgments, leading to more effortful cognitive processing and greater adjustments away from the anchor. For example, Epley & Gilovich (2005) demonstrated that when individuals generated their own anchors (i.e., self-generated anchors) and were provided with financial incentives for accuracy before the task, they then made greater adjustments away from the anchor and more accurate estimates compared to those who received no incentives. This suggests that incentives enhance motivation to engage in more effortful thinking, which reduces the influence of the anchoring bias.

Educating participants about the anchoring bias (Adame, 2016; Meub & Proeger, 2016; Morewedge et al., 2015). This approach involves increasing individuals' awareness of the anchoring bias, either implicitly through well-designed tasks or by explicitly providing information about the bias. The goal is to help individuals recognize how anchoring may influence their judgments. Meub & Proeger (2016) found that the anchoring bias is reduced through repeated forecasting tasks, suggesting that participants learned from experience and adapted their judgment strategies over time, becoming less influenced by the anchor in the later stages of the experiment.

Structuring decision-making processes to minimize reliance on arbitrary anchors (Mussweiler et al., 2000). This category focuses on altering the structure of the task to reduce the anchor's influence. A well-known strategy in this group is the "consider-the-opposite" strategy, which encourages individuals to critically evaluate their initial judgments by considering contradictory information or alternative scenarios, thereby reducing the influence of the anchor. Mussweiler et al. (2000) demonstrated the effectiveness of this strategy through two experiments, showing that asking participants to consider why an anchor might be inappropriate can reduce the anchoring bias.

The first two categories aim to shape how individuals think about or respond to anchors, and these strategies should be conducted before the actual judgment task. However, empirical studies have found that their effectiveness in reducing the anchoring bias is often limited (Chapman & Johnson, 2002; Wilson et al., 1996; Epley & Gilovich, 2005). The third category aims to directly alter the structure of the decision-making context and has shown more consistent success in mitigating the anchoring bias (Nagtegaal et al., 2020; Mussweiler et al., 2000). This structural approach resonates with the logic of MADM methods, which aim to help the DM make informed decisions by providing a structured way to evaluate the attributes and alternatives. While these methods are prescriptive in nature -intended to improve decision quality- they are typically developed under the assumption that the inputs, such as the weights, are derived from consistent human judgments. As a result, the design of many MADM methods does not explicitly account for cognitive biases, leaving them susceptible to a wide range of cognitive biases, such as the anchoring bias.

Several studies have examined the anchoring bias in MADM methods. In MADM, elicited judgments constitute valuation tasks, since they reflect subjective tradeoffs and preferences rather than estimates of objectively verifiable values. Therefore, most empirical studies define the anchoring bias in MADM as systematic deviations from preference coherence and consis-

tency in the elicited judgments or resulting decision outcomes. Sun et al. (2025) examined the anchoring bias in MAVT during the value function elicitation step. They found that the analyst's choice of starting point in the midvalue splitting procedure can systematically bias judgments, lead to biased value functions, and consequently biased decision results. Their study also tested "counter-anchor" and "no anchor" debiasing strategies, both of which effectively reduced this bias.

Buchanan & Corner (1997) investigated the effect of the anchoring bias in a multi-objective production scheduling decision problem with two interactive methods: the Zionts and Wallenius method (Zionts & Wallenius, 1983) and the Free Search method (Buchanan, 1997). The Zionts and Wallenius method starts from the current solution, and the DM iteratively chooses preferred directions of improvement until no further direction is preferred. In contrast, the Free Search method lets the DM explore the solution space directly, without a fixed starting point. The results showed that the decision outcome was significantly affected by the fixed starting point in the Zionts and Wallenius method, whereas no such effect was observed in the Free Search method. This highlights how structures designed to aid the decision-making process can themselves become a source of bias and unintentionally introduce or amplify the anchoring bias.

Different MADM methods have distinct procedures, and research shows that their structural design plays a crucial role in determining how susceptible the methods are to the anchoring bias. Methods that rely on a single directional anchor, such as SMART and Swing, have been shown to be particularly vulnerable. In SMART, scoring begins from the least important attribute, whereas in Swing, it begins from the most important. Despite this difference, both methods show the same pattern of the anchoring bias in an estimation task: the weights for the less important attributes are higher than their actual ones, while the weights for the more important attributes are lower than their actual ones (Rezaei, 2021). Building on this work, Rezaei et al. (2024) analyzed BWM, which incorporates two directional anchors through pairwise comparisons: one between the best attribute and the remaining attributes, and another between the worst attribute and the remaining attributes. The results showed that, compared to SMART and Swing, which rely on single-directional anchors, BWM can produce lower weights for the less important attributes and higher weights for the more important attributes. In another study, Rezaei (2022) demonstrated that methods relying on a single directional anchor are more prone to the anchoring bias, while the two-directional structure in BWM reduces its impact. These findings suggest that MADM methods with multiple or opposite anchors are less susceptible to the anchoring bias.

In summary, these studies highlight the significant role of method structure in shaping decision outcomes. While existing MADM methods like SMART and Swing are thoughtfully designed and account for many theoretical and practical considerations, they were not originally developed with the impact of cognitive biases in mind. Insights from behavioral decision research have since shown that preferences are not merely revealed but are often constructed during the elicitation process and are context-dependent (Slovic, 1995; Payne et al., 1992). This perspective reinforces the prescriptive function of MADM methods, which aim to guide the DM toward more informed decisions through structured procedures. On the one hand, if the design of these procedures does not account for cognitive biases from the outset, these procedures might become the source of the cognitive biases and fail the prescriptive aim of improving decision quality. On the other hand, recognizing that preferences are constructed during the elic-

itation process also opens the door to effective debiasing, as long as these behavioral aspects are considered in the method structure.

As one of the most widely used weight elicitation methods, the Tradeoff procedure has been included in prior investigations of behavioral effects. For instance, Weber & Borcherding (1993) examined several cognitive influences, such as range sensitivity, splitting bias, hierarchical structuring, and framing effects, across different elicitation methods, including the Tradeoff procedure. However, their work did not address the potential for the anchoring bias within the Tradeoff procedure's structure. This study focuses specifically on the anchoring bias, one of the most well-known cognitive biases (Tversky & Kahneman, 1974; Furnham & Boo, 2011), in the Tradeoff procedure, thereby extending this line of research by identifying and examining a previously overlooked behavioral vulnerability. The present contribution should be differentiated from recent studies by Rezaei and colleagues, who examined the anchoring bias in SMART and Swing. In those methods, anchoring arises from an explicit anchor, whereas in the Tradeoff procedure, the bias originates from the choice of reference attribute. This broadens our understanding of how biases can enter MADM methods, and practically, it calls for method-specific strategies rather than general ones.

3.3 An Overview of the Tradeoff Procedure

To introduce the Tradeoff procedure, it is essential to first position it within the context of multi-attribute value theory (MAVT). MAVT is a widely recognized MADM method developed by Keeney & Raiffa (1976). It assumes a DM's preferences can be represented by a value function consisting of attribute-specific value functions and corresponding weights (scaling constants). The Tradeoff procedure was originally developed within this theory to elicit the scaling constants by identifying indifference points between hypothetical alternatives. Since the calculation of weights using the Tradeoff procedure relies on the existence and structure of the underlying value function, MAVT provides the necessary conceptual foundation. This section introduces MAVT, the Tradeoff procedure, and the Best-Worst Tradeoff (BWT).

3.3.1 Multi-Attribute Value Theory

The first step of MAVT involves clearly defining the decision-making context. This includes identifying the objectives, attributes, and alternatives. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_M\}$ denote the set of alternatives being considered. Each alternative $a_i \in \mathcal{A}$ is evaluated based on a set of attributes $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$, where each attribute $X_j \in \mathcal{X}$ is used to assess how well an alternative meets the objectives. The goal is to determine which alternative best satisfies the objectives based on the evaluation of the attributes.

Once the decision-making context is established, the next step is to elicit attribute-specific value functions. An attribute-specific value function represents how a DM values different levels of performance for a specific attribute on a scale from 0 (the least preferred level) to 1 (the most preferred level). There are different value function elicitation procedures such as the mid-value splitting procedure, the lock-step procedure, the standard difference procedure, successive comparison and curve fitting, among others (Keeney & Raiffa, 1976; Beinat, 1997; Fishburn,

1967; Watson & Buede, 1987). The midvalue splitting procedure is used in the current study because it is one of the most commonly employed methods originally introduced by Keeney & Raiffa (1976). In the midvalue splitting procedure, the DM identifies several midvalue points within each attribute's range, which are used to plot the value function. When a higher attribute value is preferred, an increasing function is applied. A formal definition of the midvalue is (Kirkwood & Sarin, 1980):

Definition 3.1
$x_j^{0.5}$ is said to be the midvalue of the interval $[x_j^0, x_j^1]$ if the decision-maker will give up the same amount of some other attribute to go from x_j^0 to $x_j^{0.5}$ as from $x_j^{0.5}$ to x_j^1.

A similar definition can be used for the decreasing attribute-specific value function.

The third step involves eliciting the attribute weights (or scaling constants). Different weight elicitation methods can be used, such as SMART (Edwards, 1977) and Swing (von Winterfeldt & Edwards, 1993). The Tradeoff procedure is considered in this study because it is the one developed originally in MAVT, which will be introduced in detail in the next subsection.

Once the attribute-specific value functions and weights have been elicited, the next step is to aggregate them into an overall value for each alternative. Different aggregation models can be applied depending on the underlying preference assumptions (Keeney, 1974). The additive model is the simplest and most widely used. It is valid under the assumptions of mutual preferential independence and difference independence (Smith & Dyer, 2021; Keeney & Raiffa, 1976).

Definition 3.2
Attributes $X_1, \ldots, X_N$ are mutually preferentially independent if every subset of attributes is preferentially independent of its remaining attributes.

Definition 3.3
Attribute X_j is difference independent of the remaining attributes if the preference difference between any two levels of X_j is not affected by the fixed levels on the other attributes.

Under these assumptions, the additive model is expressed as:

$$v(a_i) = \sum_{j=1}^{N} w_j v_j(a_{ij}) \tag{3.1}$$

Here, $v(a_i)$ represents the overall value of alternative a_i, scaled from 0 to 1. $v_j(a_{ij})$ denotes the attribute-specific value of alternative a_i with respect to attribute X_j, and w_j is the scaling constant (or weight) associated with attribute X_j.

The final step of MAVT is to rank the alternatives based on their overall values or select the best-performing alternative. Specifically, for any two alternatives a_k and a_l,

$$v(a_k) \geq v(a_l) \Leftrightarrow a_k \succsim a_l$$

where the symbol $\succsim$ denotes "preferred or indifferent to" (Keeney & Raiffa, 1976) in the sense of the ordering implied by the overall value function (i.e., the additive model).

In MAVT, the overall value of an alternative is computed as a weighted sum of attribute-specific value functions if the preferential independence conditions hold, as shown in the aggregation model (3.1). This model implies that the outcome depends jointly on two elements: the attribute-specific value functions and the weights (scaling constants). If either component is biased, the overall evaluation will also be distorted. Thus, even if the attribute-specific value functions are properly elicited, the results can only be meaningful if the weights are also properly specified. It is, therefore, essential to ensure that the weights elicited reflect the DM's preference rather than being distorted by the anchoring bias during the Tradeoff procedure.

3.3.2 Tradeoff Procedure

The Tradeoff procedure typically involves the following steps:

Define attribute ranges. To define the full range of each attribute, it is first necessary to identify all possible performance levels that the attribute can take in the given decision context, within the compensatory range. Let $\underline{x}_j$ and $\overline{x}_j$ denote the worst and best performance levels for attribute X_j, respectively, where $j \in \{1, 2, \ldots, N\}$ for N attributes. Thus, $v_j(\underline{x}_j) = 0$ and $v_j(\overline{x}_j) = 1$, for all j.

Identify the importance order. The DM is presented with a set of hypothetical alternatives, each representing an extreme combination of attribute values. For a set of n attributes, each alternative corresponds to a combination where one attribute is at its best performance level (denoted $\overline{x}_j$) while the others are at their worst performance levels (denoted $\underline{x}_j$). The alternatives can be represented as:

$$a_1 := (\overline{x}_1, \underline{x}_2, \ldots, \underline{x}_N), \quad a_2 := (\underline{x}_1, \overline{x}_2, \ldots, \underline{x}_N), \quad \ldots, \quad a_N := (\underline{x}_1, \underline{x}_2, \ldots, \overline{x}_N)$$

The DM is asked to rank these alternatives based on their preferences. This ranking process reveals the importance order of the attributes. The attribute that is at its best performance level in the most preferred alternative is defined as the most important attribute for this specific problem. For instance, if the DM ranks a_1 as the most preferred alternative, then X_1 is considered as the most important attribute for the DM, and will be used to construct indifference pairs in the next step.

Construct indifference pairs. The DM is presented with a series of hypothetical comparisons, each involving two alternatives, in order to establish indifference relations. In each comparison, one attribute of the first alternative is set at its best possible level, while all other attributes are fixed at their worst levels. The second alternative in each comparison involves adjusting the level of the most important attribute until the DM is indifferent between the two alternatives, with all other attributes fixed at the worst levels. The adjusted level of the most important attribute at which the DM becomes indifferent reflects the tradeoff they are willing to make between the most important attribute and the other attribute. Attributes not involved in the comparison are held constant at their worst levels in order to isolate the tradeoff between the best attribute and each of the other attributes.

This process results in $N - 1$ indifference pairs for N attributes. For any indifference pair

between the most important attribute X_B, or in short B, and another attribute X_k, or in short k, it can be written as follows [1]:

$$(\underline{x}_1, \ldots, \underline{x}_B, \ldots, \overline{x}_k, \ldots \underline{x}_N) \sim (\underline{x}_1, \ldots, x_B^{B,k}, \ldots, \underline{x}_k, \ldots, \underline{x}_N) \tag{3.2}$$

where $\sim$ indicates the indifference relation, $x_B^{B,k}$ represents the adjusted levels of the most important attribute X_B that makes the DM indifferent between the two alternatives in each indifference pair.

Calculate weights. Each indifference pair generates an equation that quantifies the tradeoff between the most important attribute and another attribute. The first alternative has the most important attribute at its worst level $\underline{x}_B$ and attribute k at its best level $\overline{x}_k$. The second alternative has the most important attribute adjusted to a level $x_B^{B,k}$, while the other remaining attributes remain at their worst levels (with a value of zero). The indifference relation implies that the total value of the two alternatives is equal. Under the additive value function specified in (3.1), this can be expressed by the following equation:

$$w_k \underbrace{v_k(\overline{x}_k)}_{=1} + \sum_{\substack{j=1 \\ j \neq k}}^{N} w_j \underbrace{v_j(\underline{x}_j)}_{=0} = w_B v_B(x_B^{B,k}) + \sum_{\substack{j=1 \\ j \neq B}}^{N} w_j \underbrace{v_j(\underline{x}_j)}_{=0} \tag{3.3}$$

which collapses into:

$$w_k = w_B v_B(x_B^{B,k}) \tag{3.4}$$

This equation reflects the tradeoff the DM is willing to make between the most important attribute B and the attribute k. For N attributes, using the $N-1$ equations obtained from the indifference pairs and the constraint that the sum of the weights is equal to one, the weights for all attributes can be calculated as follows:

$$\begin{cases} w_1 = w_B v_B(x_B^{B,1}) \\ \ldots \\ w_k = w_B v_B(x_B^{B,k}) \\ \ldots \\ w_N = w_B v_B(x_B^{B,N}) \\ \sum_{j=1}^{N} w_j = 1 \\ w_j \geq 0, \quad j = 1, 2, ..., N \end{cases} \tag{3.5}$$

The weights derived using the Tradeoff procedure are based on the DM's indifference relations between constructed indifference pairs. In this process, the DM adjusts the level of the most important attribute starting from its worst level. This initial worst level may act as an anchor, leading to insufficient adjustments when determining the level of the most important attribute required for indifference. As a result, the anchoring bias might distort the assigned

[1] For simplicity, we use the index j to denote attribute X_j in the remainder of this paper.

scores for the most important attribute and, consequently, lead to biased weights (we discuss this in detail in Section 3.4).

3.3.3 Best-Worst Tradeoff Method

The Best-Worst Tradeoff (BWT) was developed by Liang et al. (2022). It builds on the traditional Tradeoff procedure by eliciting tradeoffs between the most important attribute and the remaining attributes, as well as between the least important attribute and the remaining attributes, and then using an optimization model to derive attribute scaling constants (weights) by minimizing the inconsistency in the judgments.

The BWT procedure involves six steps in total. The first two steps, which are (i) define the alternatives and attributes, and (ii) determine the attribute-specific value functions, are the same as in MAVT and will not be repeated here. The remaining steps are:

Identify best and worst attribute. The DM in BWT needs to identify both the best and the worst attributes. This identification process follows a similar logic to the "*Identify the Importance Order*" step in Section 3.3.2. The DM is presented with a set of hypothetical alternatives, each constructed such that one attribute is fixed at its best performance level while all other attributes are set to their worst performance levels. From the DM's ranking of these alternatives, the importance order of the attributes is derived. Throughout the paper, we use the terms *best* and *worst* to indicate the most and least important attributes, respectively. Accordingly, the most important attribute is denoted as the best attribute (B), and the least important attribute as the worst attribute (W). These two attributes are then used to construct the indifference pairs in the next steps.

Best-to-Others tradeoff. This step mirrors the traditional Tradeoff procedure, in which the best attribute is compared against each of the remaining attributes to establish tradeoff relations. For a decision problem involving N attributes, a total of $N-1$ indifference pairs are elicited between the best attribute B and every other attribute j, as described in (3.2).

If the DM reaches indifference when the level of the best attribute is adjusted to $x_B^{B,j}$, the value at that point, $v_B(x_B^{B,j})$, is denoted by b_{jB}, which represents the DM's elicitation of the ratio w_j/w_B. The reciprocal, $b_{Bj} = 1/b_{jB}$, then corresponds to the ratio w_B/w_j. To facilitate the mathematical representation, the BWT method defines the Best-to-Others (BTO) vector $\mathbf{b}^{BO} = \{b_{B1}, b_{B2}, \ldots, b_{BN}\}$, where each element b_{Bj} captures the tradeoff information between the best attribute B and another attribute j in the BTO comparisons. While the traditional Tradeoff procedure often constructs only the BTO indifference pairs, BWT also considers the tradeoff between the other attributes and the worst attribute.

Others-to-Worst tradeoff. This step involves comparing the other attributes to the worst attribute. For a decision problem with N attributes, this step also involves constructing $N-1$ indifference pairs, and each pair compares an attribute j, $j \neq W$, with the worst attribute W. These indifference pairs determine how much of an improvement in X_j is needed for the DM to consider it equivalent in changes of the worst attribute from its worst level to the best level. This process results in $N-1$ indifference pairs for N attributes. For any indifference pair between the least important attribute W and another attribute k, it can be written as follows:

$$(\underline{x}_1, \ldots, x_k^{k,W}, \ldots, \underline{x}_W, \ldots, \underline{x}_N) \sim (\underline{x}_1, \ldots, \underline{x}_k, \ldots, \overline{x}_W, \ldots, \underline{x}_N) \tag{3.6}$$

Similar to the traditional Tradeoff procedure, the indifference relations can also be expressed by equations using the additive value function (3.1), which is:

$$w_k v_k(x_k^{k,W}) + \sum_{\substack{j=1 \\ j\neq k}}^{N} w_j \underbrace{v_j(\underline{x}_j)}_{=0} = w_W \underbrace{v_W(\overline{x}_W)}_{=1} + \sum_{\substack{j=1 \\ j\neq W}}^{N} w_j \underbrace{v_j(\underline{x}_j)}_{=0} \tag{3.7}$$

which collapses into:

$$w_k v_k(x_k^{k,W}) = w_W \tag{3.8}$$

For N attributes, using the $N-1$ equations obtained from the OTW indifference pairs and the constraint that the sum of the weights is equal to one, the weights can be calculated as follows:

$$\begin{cases} w_1 v_1(x_1^{1,W}) = w_W \\ \dots \\ w_k v_k(x_k^{k,W}) = w_W \\ \dots \\ w_N v_N(x_N^{N,W}) = w_W \\ \sum_{j=1}^{N} w_j = 1 \\ w_j \geq 0, \quad j = 1, 2, ..., N \end{cases} \tag{3.9}$$

Suppose the DM reaches indifference when the level of X_j is adjusted to $x_j^{j,W}$, while the worst attribute changes from its worst level to the best level. The value $v_j(x_j^{j,W})$, is denoted as b_{Wj}, representing the ratio w_W/w_j. Its reciprocal, $b_{jW} = 1/b_{Wj}$, then corresponds to the ratio w_j/w_W. The Others-to-Worst (OTW) vector is defined as $\mathbf{b}^{OW} = \{b_{1W}, b_{2W}, \dots, b_{NW}\}$, where each element b_{jW} captures the OTW tradeoff between attribute j and the worst attribute W.

Find the optimal weights. After obtaining the BTO and OTW vectors, the following system of linear equations can be formed:

$$\begin{cases} b_{Bj} = \frac{w_B}{w_j}, \quad \forall j \neq B \\ b_{jW} = \frac{w_j}{w_W}, \quad \forall j \neq W \\ w_1 + w_2 + \dots + w_N = 1 \\ w_j \geq 0, \quad j = 1, 2, ..., N \end{cases} \tag{3.10}$$

However, the elicited judgments are often inconsistent, meaning that the equation system (3.10) typically does not admit an exact solution. To address this, a nonlinear optimization model can be employed to derive the optimal weights, formulated as follows:

$$
\begin{aligned}
\text{minimize} \quad & \xi \\
\text{subject to} \quad & \left| b_{Bj} - \frac{w_B}{w_j} \right| \leq \xi, \quad \forall j \neq B \\
& \left| b_{jW} - \frac{w_j}{w_W} \right| \leq \xi, \quad \forall j \neq W \\
& w_1 + w_2 + \cdots + w_N = 1 \\
& w_j \geq 0, \quad j = 1, 2, \ldots, N
\end{aligned}
\tag{3.11}
$$

In this model, w_j represents the weight of attribute j, B is the best attribute, and W is the worst attribute. The objective is to minimize the maximum absolute violation of the equations (denoted by ξ). This model ensures that the tradeoffs are as consistent as possible while adhering to the constraint that the sum of the weights is equal to one. BWT also includes a consistency check for indifference pairs and weights, providing a basis for the DMs to revise their judgments in the two tradeoff procedures. A linear optimization form is also presented in Liang et al. (2022).

3.4 Hypotheses Development

During the tradeoff procedure, the DM is asked to compare two hypothetical alternatives that differ in their attribute levels. One alternative features a specific attribute k at its best level, while all other attributes are fixed at their worst levels. In the other alternative, all the attributes, including attribute k, are set to their worst level, and the DM adjusts the level of attribute l until the DM is indifferent between the two alternatives. This adjustment starts from an initial reference point, such as the worst level of attribute l in the second alternative, which may serve as an anchor. Due to the anchoring bias, the DM might make insufficient adjustments from this anchor, leading to distorted indifference judgments and, ultimately, biased weight elicitation.

In the BWT method, two types of tradeoff tasks are conducted: the traditional Best-to-Others (BTO) tradeoffs and the Others-to-Worst (OTW) tradeoffs. Although the two procedures differ in direction, the potential anchoring mechanism operates similarly in both cases.

Best-to-Others tradeoffs

In the BTO tradeoffs, the DM adjusts the best attribute B from its worst level until she is indifferent between this adjustment and the full range change (from its best level to its worst level) of another attribute $k, k \neq B$. Due to the anchoring bias, the worst level of the best attribute $\underline{x}_B$ might act as an anchor for the DM and result in insufficient adjustments in defining the adjusted level $x_B^{B,k}$. That is to say, the DM might not make enough distance from the low anchor and could assign a number closer to the worst level $\underline{x}_B$ when defining $x_B^{B,k}$ due to the insufficient adjustment.

In general, for an increasing (decreasing) attribute-specific value function of attribute B, this leads to a lower (higher) adjusted level $x_B^{B,k}$, which in turn leads to a lower $v_B(x_B^{B,k})$. Since $v_B(x_B^{B,k})$ is associated to the ratio w_k/w_B (See equation (3.4)). If $v_B(x_B^{B,k})$ is lower due to the anchoring bias, the derived weights will be distorted accordingly. Specifically, the weight of the attributes $k, w_k, k \neq B$, may decrease while the weight of the best attribute B, w_B may

increase.

Others-to-Worst tradeoffs

In the OTW tradeoffs, the DM adjusts the attribute $k, k \neq W$ from its worst level until she is indifferent between this adjustment and the full range change (from its best level to its worst level) of the worst attribute W. Similar to the BTO procedure, the worst level of the attribute k, $\underline{x}_k, k \neq W$, might act as a low anchor to the DM, and result in insufficient adjustments in defining the adjusted level $x_k^{k,W}$. That is to say, the DM might assign a number closer to the worst level $\underline{x}_k$ when defining $x_k^{k,W}$ due to the insufficient adjustment.

For an increasing (decreasing) attribute-specific value function of $k, k \neq W$, this might lead to lower (higher) adjusted levels $x_k^{k,W}$ for all OTW tradeoffs, which in turn leads to a lower $v_k(x_k^{k,W})$. Since $v_k(x_k^{k,W})$ represents the ratio w_W/w_k (See equation (3.8)). From (3.8), it is easy to observe that if $v_k(x_k^{k,W})$ is lower due to the anchoring bias, the weight of the worst attribute w_W may decrease while the weight of the non-worst attribute k, $w_k, k \neq W$, may increase.

It is important to note that one tradeoff pair, the comparison between the best and worst attributes, is common to both the BTO and OTW procedures. Therefore, any differences in the resulting weights between the two procedures originate from the remaining $2(N-2)$ tradeoffs: the BTO comparisons between the best attribute and the remaining attributes (excluding the worst), and the OTW comparisons between the worst attribute and the remaining attributes (excluding the best).

Before presenting the propositions on how anchoring affects the weights in the BTO and OTW procedures, we first show how the weights are derived from each procedure separately, as follows.

Calculating attribute weights based on BTO and OTW

Assume a decision problem with N attributes $\{X_1, \ldots, X_N\}$. Let B denote the most important (best) attribute, W the least important (worst) attribute, and $\mathcal{K} := \{1, \ldots, N\} \setminus \{B, W\}$ the index set of the remaining attributes.

Best-to-Others (BTO) tradeoff. Eliciting the ratios $b_{Bk} = w_B/w_k$ for every $k \in \mathcal{K} \cup \{W\}$ and imposing the normalization $\sum_{j=1}^{N} w_j = 1$ gives

$$w_B = \frac{1}{1 + \sum_{j \neq B} \frac{1}{b_{Bj}}}, \qquad w_k = \frac{\frac{1}{b_{Bk}}}{1 + \sum_{j \neq B} \frac{1}{b_{Bj}}}, \quad k \neq B \tag{3.12}$$

Others-to-Worst (OTW) tradeoff. If instead the judgments use the worst attribute, one elicits $b_{kW} = w_k/w_W$ for every $k \in \mathcal{K} \cup \{B\}$. Solving the same normalization condition yields the following result.

$$w_W = \frac{1}{1 + \sum_{j \neq W} b_{jW}}, \qquad w_k = \frac{b_{kW}}{1 + \sum_{j \neq W} b_{jW}}, \quad k \neq W \tag{3.13}$$

Consistency without anchoring. When the tradeoff answers are internally consistent, the cross-ratio condition

$$b_{BW} \;=\; b_{Bk}\, b_{kW}, \quad \forall\, k \in \mathcal{K} \tag{3.14}$$

holds (see Liang et al., 2022). Under this condition, the weight vectors obtained from BTO, OTW, and BWT are identical.

Effect of the anchoring bias. Anchoring typically inflates both b_{Bk} and b_{kW}, violating the above equality. Because the shared ratio $b_{BW} = w_B/w_W$ is present in *both* procedures, we treat it as fixed and attribute any differences in the resulting weights to bias-induced changes in the remaining ratios. Propositions 1 and 2 formalise how these changes push the BTO and OTW weights in opposite directions.

Proposition 1. In the BTO procedure, the low-anchored $v_B(x_B^{B,k}), k \neq \{B, W\}$ results in an increase in w_B and w_W and a decrease in $w_k, k \neq \{B, W\}$.

Proof : Let $Q := v_B\big(x_B^{B,k}\big)$ and $S := \sum_{j \neq \{B,k\}} v_B\big(x_B^{B,j}\big)$. From the BTO identities

$$w_B = \frac{1}{1+S+Q}, \qquad w_k = \frac{Q}{1+S+Q}, \qquad w_W = w_B\, v_B\big(x_B^{B,W}\big)$$

where $v_B\big(x_B^{B,W}\big)$ is constant in this perturbation. Differentiating,

$$\frac{\partial w_B}{\partial Q} = -\frac{1}{(1+S+Q)^2} < 0, \qquad \frac{\partial w_k}{\partial Q} = \frac{1+S}{(1+S+Q)^2} > 0, \qquad \frac{\partial w_W}{\partial Q} = v_B\big(x_B^{B,W}\big)\,\frac{\partial w_B}{\partial Q} < 0$$

Thus, a downward shift in Q (a low anchor) increases w_B and w_W but decreases w_k. □

Proposition 2: In the OTW procedure, the low-anchored $v_k(x_k^{k,W}), k \neq \{B, W\}$ results in a decrease in w_B and w_W and an increase in $w_k, k \neq \{B, W\}$.

Proof : Let $P := v_k\big(x_k^{k,W}\big)$ and $R := \sum_{j \neq \{W,k\}} 1/v_j\big(x_j^{j,W}\big)$. Then

$$w_W = \frac{1}{1+R+1/P}, \qquad w_k = \frac{w_W}{P}, \qquad w_B = \frac{w_W}{v_B\big(x_B^{B,W}\big)}.$$

Differentiating,

$$\frac{\partial w_W}{\partial P} = \frac{1}{(1+R+1/P)^2 P^2} > 0, \qquad \frac{\partial w_B}{\partial P} = \frac{1}{v_B\big(x_B^{B,W}\big)}\,\frac{\partial w_W}{\partial P} > 0$$
$$\frac{\partial w_k}{\partial P} = -\frac{w_W}{P} + \frac{1}{P}\,\frac{\partial w_W}{\partial P} = \frac{1-(1+R+1/P)P^2}{(1+R+1/P)^2 P^3} < 0$$

Thus, a downward shift in P decreases w_W and w_B but increases w_k, completing the proof. □

Drawing from the above discussion, we propose the following hypotheses:

H1: Using the most important (best) attribute to construct the indifference pairs in the Tradeoff procedure leads to higher weights for the best and worst attributes and lower weights for the other attributes.

H2: Using the least important (worst) attribute to construct the indifference pairs in the Tradeoff procedure leads to lower weights for the best and worst attributes and higher weights for the other attributes.

Since the BWT method utilizes both BTO and OTW tradeoffs, it inherently incorporates the "consider-the-opposite" debiasing strategy in its structure. "Consider-the-opposite" (Mussweiler et al., 2000) is one of the most effective debiasing strategies developed for reducing the anchoring bias. It encourages individuals to critically evaluate their initial judgments by considering contradictory information or alternative scenarios. In the context of MADM methods, BWT achieves this by requiring the DM to evaluate tradeoffs from two different perspectives: using the best and worst attributes to construct the indifference pairs. This is not merely a matter of increasing the number of elicited comparisons; it systematically encourages the DM to conduct tradeoffs from opposing perspectives.

As illustrated in the previous discussions, the anchoring bias tends to distort the weights in opposite directions in the BTO and OTW tradeoff procedures. Thus, when the two sets of tradeoffs are combined in the BWT method, these opposite distortions may cancel each other out, reducing the anchoring bias introduced by any single anchor. The optimization model solves for the weight that best fits both tradeoff perspectives simultaneously and leads to weights that are less likely to be biased by a single direction anchor.

Therefore, we hypothesize:

H3: Using the Best-Worst Tradeoff method can reduce the anchoring bias in the Tradeoff procedure.

Remark. When the decision problem involves exactly *three* attributes, the weight of the worst attribute, w_W, shows a special behavior. Consider X_1 as the best attribute, X_3 as the worst attribute and X_2 as the remaining one. From (3.12) and (3.13), the weights for the three attributes can be obtained for BTO and OTW, respectively, as follows.

$$BTO\begin{cases} w_1 = \frac{1}{1+\frac{1}{b_{12}}+\frac{1}{b_{13}}} \\ w_2 = \frac{1}{b_{12}}\frac{1}{1+\frac{1}{b_{12}}+\frac{1}{b_{13}}} \\ w_3 = \frac{1}{b_{13}}\frac{1}{1+\frac{1}{b_{12}}+\frac{1}{b_{13}}} \end{cases} \qquad OTW\begin{cases} w_1 = b_{13}\frac{1}{1+b_{13}+b_{23}} \\ w_2 = b_{23}\frac{1}{1+b_{13}+b_{23}} \\ w_3 = \frac{1}{1+b_{13}+b_{23}} \end{cases} \tag{3.15}$$

If the anchoring bias does not affect the tradeoff process, and the preferences elicited are fully consistent (refer to (3.14)), then $b_{13} = b_{12}b_{23}$ holds. Under this condition, BTO and OTW yield identical weights. Wu et al. (2024) established a theorem[2] showing that there is a unique optimal solution for the problem under a not-fully consistent comparison system when three attributes are involved. Specifically, according to Wu et al. (2024) the optimal weights are:

[2]For proof of this theorem, please read Wu et al. (2024).

$$\begin{cases} w_1 = \frac{b_{13}-\xi^*}{b_{13}+b_{23}+1} \\ w_2 = \frac{b_{23}+\xi^*}{b_{13}+b_{23}+1} \\ w_3 = \frac{1}{b_{13}+b_{23}+1} \end{cases} \text{, if } b_{12} \times b_{23} < b_{13} \qquad \begin{cases} w_1 = \frac{b_{13}+\xi^*}{b_{13}+b_{23}+1} \\ w_2 = \frac{b_{23}-\xi^*}{b_{13}+b_{23}+1} \\ w_3 = \frac{1}{b_{13}+b_{23}+1} \end{cases} \text{, if } b_{12} \times b_{23} > b_{13} \tag{3.16}$$

From (3.15) (OTW weights) and (3.16), it is clear that w_3 has the same weight for OTW and BWT (regardless of the consistency level of BWT).

3.5 Experimental Design

To test the hypotheses developed in Section 3.4, we designed a hypothetical decision problem with three attributes and implemented it in an online questionnaire (Qualtrics) that followed the steps of MAVT. The questionnaire consisted of a larger experiment with two independent decision problems. The first, a two-attribute problem, has been reported elsewhere (Sun et al., 2025) and focused on the anchoring bias in value function elicitation. The current study focused on the second, a three-attribute problem specifically designed to examine the anchoring bias in weight elicitation. In this problem, participants evaluated three attributes, and the experimental variation focused on different versions of the Tradeoff procedure (BTO, OTW, and BWT).

Combining the two problems in a single questionnaire was both efficient and methodologically sound. First, the problems targeted different steps of MAVT (value functions vs. weights), so there was no conceptual overlap that could cause contamination. Second, any potential learning or familiarity effects from the first problem would apply equally across all conditions of the second problem. Since the analysis compares differences between different Tradeoff procedures (BTO, OTW, BWT), such general familiarity does not bias results in favor of one method.

The first part of the questionnaire was to inform the participants about the content of the study, including the objective, procedures, risks, and benefits. The participants were then asked to voluntarily provide their consent to participate in this study.

In the second part, participants were presented with a hypothetical decision problem. As illustrated in Table 3.1, an apartment selection problem was designed. The rent attribute range was set based on the current rental market conditions in the target countries of the experiment. The lower bound for commute distance was set at a sufficiently high level to ensure that the value function remains monotonic. This prevents situations where individuals might favor an intermediate commute distance over living too close to or too far from their workplace. The distance to the shopping center was deliberately given a narrow range to reduce its overall importance. This design was intended to encourage a similar attribute importance order across participants, reducing variance due to individual differences in the experiment. Importantly, while the design encouraged consistent importance rankings for the attributes, the subsequent analysis was not dependent on the specific attributes themselves, but rather on their relative importance rankings as reported by each participant. That is, all analyses were conducted based on whether an attribute was ranked as most, least, or intermediately important, regardless of which physical attribute (e.g., rent, commute, shopping distance) it represented.

The third part of the experiment consisted of three tasks: eliciting attribute-specific value

Table 3.1: Attributes used in the decision problem

Attribute	Unit	Description	Range
Monthly Rent	euro	It is the amount of money one has to pay each month to rent the apartment.	[600,1500]
Commute Distance	kilometer	This is the proximity of the apartment to your workplace.	[5,15]
Distance to the Shopping Center	meter	This is the proximity of the apartment to the nearest shopping center where you buy groceries.	[100,500]

functions, eliciting weights, and validating the additivity assumption. While the value functions were elicited separately, we integrated the weight elicitation with the additivity assumption validation because both procedures require constructing indifference pairs (See Appendix A for an example of the tradeoff questions). This integration reduced the total number of indifference judgments one has to make, allows for a more efficient experimental structure, and minimizes unnecessary cognitive effort. Reducing redundancy in elicitation tasks is particularly important in multi-criteria decision experiments where complex tradeoffs are involved, and quantitative judgments are required (Larichev, 1992; Larichev & Brown, 2000; Jaspersen & Montibeller, 2015).

To implement this integration, participants first determined the importance order of the three attributes following the procedure described in Section 3.3. The identified best and worst attributes are then used to construct the BTO and OTW tradeoffs. Assume X_1 was the best attribute, X_3 was the worst attribute, and X_2 was the remaining attribute. In total, each participant evaluated three indifference pairs across the two tradeoff procedures (see Table 3.2), noting that P_2 is identical in the two tradeoff procedures.

Table 3.2: Indifference pairs in the experiment

Indifference Pairs	Best-to-Others tradeoff	Others-to-Worst tradeoff
P_1	$(\underline{x}_1, \overline{x}_2, \underline{x}_3) \sim (x_1^{1,2}, \underline{x}_2, \underline{x}_3)$	$(\underline{x}_1, x_2^{2,3}, \underline{x}_3) \sim (\underline{x}_1, \underline{x}_2, \overline{x}_3)$
P_2	$(\underline{x}_1, \underline{x}_2, \overline{x}_3) \sim (x_1^{1,3}, \underline{x}_2, \underline{x}_3)$	$(x_1^{1,3}, \underline{x}_2, \underline{x}_3) \sim (\underline{x}_1, \underline{x}_2, \overline{x}_3)$

To verify the additivity assumption, we assessed an additional set of three indifference relations. Take P_1 from the BTO tradeoffs as an example (see Table 3.2): once a participant identified the indifference value $x_1^{1,2}$, we modified the original pair by changing the third attribute level from $\underline{x}_3$ to $\overline{x}_3$ and asked the participants whether they were still indifferent between the two options. If so, it indicated that the preferences between X_1 and X_2 were independent of X_3. We performed similar checks for the other two pairs, $\{X_1, X_3\}$ and $\{X_2, X_3\}$. Once all indifference pairs had been assessed, the additivity assumption was fully validated. In the formal BWT procedure, there will be necessary revisions after checking the ordinal and cardinal consistency ratios, but this step was not performed due to the experimental constraints.

Importantly, this experiment employed a within-subject design under three conditions: BTO tradeoff, OTW tradeoff, and BWT tradeoff. Since the BWT method is inherently constructed from the BTO and OTW tradeoffs, a within-subject design was particularly well-suited. By applying a within-subject design, the experiment ensured that any observed differences across the BTO, OTW, and BWT conditions can be directly attributed to the tradeoff procedures them-

selves, rather than to differences among participants (Greenwald, 1976). In contrast, a between-subject design would introduce inter-participant variability, such as the participant's prior experiences, knowledge, or decision-making styles, that could potentially obscure these effects (Charness et al., 2012).

Since the tradeoff procedure requires attribute-specific value functions to calculate weights, we used the midvalue splitting procedure to elicit these functions after completing the tradeoff tasks. Although this study focuses on the anchoring bias in the weight elicitation step, value functions are essential for converting indifference judgments into weights under the additive model. To ensure that any observed bias in the weights stemmed from the tradeoff elicitation itself, and not from inconsistencies in value function elicitation, all participants used the same value function elicitation procedure. Take eliciting the attribute-specific value function for the rent as an example. To obtain the first midvalue point, we asked, "*Suppose the price drop in monthly rent would be from 1500 to a certain value (r_1) or from that same value (r_1) to 600. Please assign a value to r_1 such that for the two price drops, you would accept the same increase in commute distance and shopping center.*" r_1 was identified as the first midvalue point. The subsequent midvalue points were also determined following similar questions, and the attribute-specific value function for rent could be plotted using these midvalue points (Keeney & Raiffa, 1976, Section 3.4.8).

The last step was collecting demographic information such as age, gender, and education level to have a comprehensive understanding of the participant's profile. Besides, the participant's current rent or mortgage, commute distance, and distance to the shopping center were also collected as control variables.

Data were collected via the online platform Prolific (Palan & Schitter, 2018), which offers various pre-screening features and a response verification process that enhances data quality. We recruited a total of 440 participants from six European countries: the Netherlands, Germany, France, Belgium, Denmark, and Luxembourg. These countries share similar rental market conditions and were thus selected. Since the questionnaire was in English, we also limited participation to individuals fluent in English using the pre-screening functions. Moreover, the response verification function enabled us to reject incomplete answers or those respondents that provided answers outside of the value ranges. Participants received a small monetary incentive for successful completion. Such incentives have been shown to increase response rates and improve the quality of answers (Singer & Ye, 2013). Of the 440 participants recruited, 36 were excluded for not completing the experiment. An additional 68 participants were removed as they provided values outside the specified ranges or identical values across all midvalue points. This indicated either inattention or insufficient understanding of the MAVT questions. The entire questionnaire, which comprised the two decision problems, was completed on average of 16 minutes and 38 seconds, with a standard deviation of 8 minutes and 59 seconds.

The final sample included 336 participants; the detailed demographics of the sample are presented in Table 5.2. After data collection, the statistical analyses were conducted in SPSS, Version 29. We employed both parametric and non-parametric tests. ANOVA with post hoc tests was used to capture the magnitude of the effects, while Wilcoxon Signed-Rank tests were applied to assess the direction of the changes. We considered both aspects important: although the magnitude indicates the strength of the anchoring bias, the direction provides a more direct test of whether the bias systematically shifts the results in the hypothesized way.

Table 3.3: Demographic characteristics of participants ($n = 336$)

Characteristics	Levels	Percent
Gender	Male	62%
	Female	35%
	Other	3%
Age	[18,24]	25%
	[25,34]	49%
	[35,44]	15%
	> 44	11%
Education	High School	13%
	Bachelor's degree	34%
	Master's degree	33%
	Other	20%

3.6 Results and Discussion

The experiment was designed to test the effect of the anchoring bias in the Tradeoff procedure within the context of MAVT. Both parametric and nonparametric tests were used to perform the analysis. Notably, ANCOVA tests were conducted to examine whether the participant's current living space, housing costs, commute distance, shopping distance, and demographics affected the main results of this study. The results indicated that none of the control variables had a statistically significant effect ($p > 0.05$ for all).

Since the hypotheses were based on the relative importance of attributes, and participants may prioritize the three attributes differently, it was necessary to classify attributes according to each participant's individual ranking rather than by fixed attribute labels. This approach ensured that the analysis focused on differences in weights for the *best*, *worst*, and *other* attributes as perceived by each participant, rather than being influenced by the specific attribute names. Specifically, the highest-ranked attribute for a participant was treated as their *best* attribute, the lowest-ranked as their *worst*, and the remaining one as *other*, regardless of whether the attribute was rent, commute distance, or distance to the shopping center. In general, participants tended to prioritize rent, with 85% ranking it as their most important attribute, followed by 12% for commute distance and 3% for distance to the shopping center.

The average weights based on this classification were presented in Figure 3.1. The results indicated that using the Best-to-Others (BTO) tradeoffs in the tradeoff procedure resulted in higher weights for the *best* and *worst* attributes and lower weights for the *other* attribute, compared to using the Others-to-Worst (OTW) tradeoffs in the tradeoff procedure. The Best-Worst Tradeoff (BWT) method produced weights in between the two extremes for the *best* and *other* attributes and similar weights for the *worst* attribute compared to the OTW tradeoff procedure.

ANOVA tests indicated statistically significant differences in attribute weights across the three conditions: *best* attribute, $F(2, 1005) = 5.106, p = 0.006$; *other* attribute, $F(2, 1005) = 15.144, p < 0.001$; and *worst* attribute, $F(2, 1005) = 4.819, p = 0.008$. As shown in Table 3.4, post hoc analyses revealed that, compared to the OTW group, the BTO group assigned significantly higher weights to both the *best* attribute (BTO: Mean $= 0.592$, OTW: Mean $= 0.553$, $p = 0.002$) and *worst* attribute (BTO: Mean $= 0.135$, OTW: Mean $= 0.116$, $p = 0.007$),

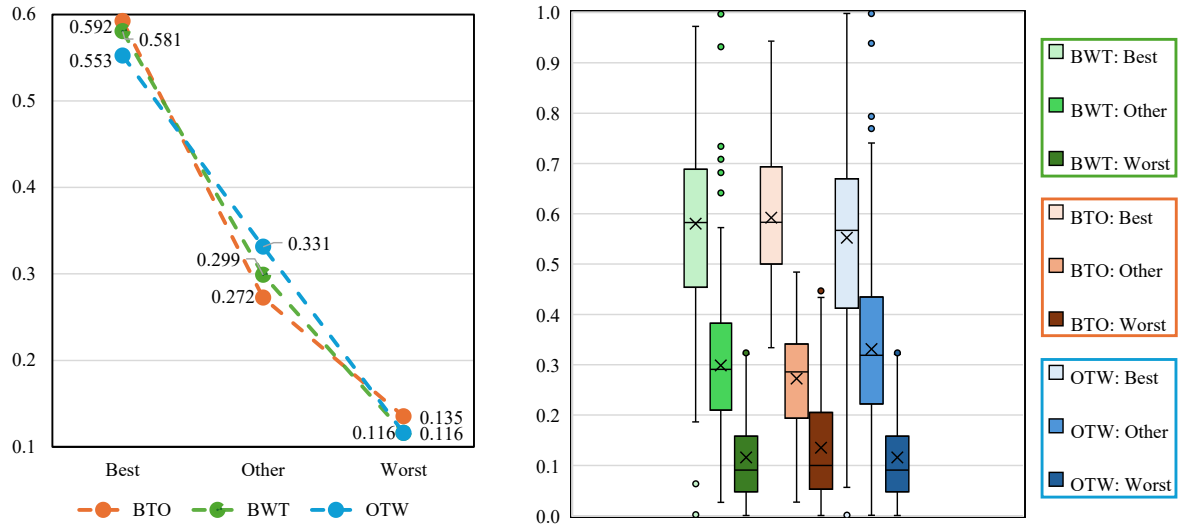


Figure 3.1: Average weight based on individual ranking

and significantly lower weights to the *other* attribute (BTO: Mean $= 0.272$, OTW: Mean $= 0.331$, $p < 0.001$). Moreover, maximum observed differences illustrated that the anchoring bias could have a substantial impact at the individual level. For instance, the largest difference for the *other* attribute between BTO and OTW was 0.6553. These findings supported hypotheses **H1** and **H2**, demonstrating the presence of the anchoring bias in the Tradeoff procedure.

Table 3.4: The mean difference of the best, other, and worst attributes of all Tradeoff method pairs

Attribute	**Tradeoff Method Pair**	**Mean Difference**	**Sig.**	95% Confidence Interval	Maximum Difference
Best	BWT-BTO	-0.012	0.359	[-0.037, 0.013]	-0.5219
Best	BWT-OTW	0.028	0.029	[-0.003, 0.053]	0.3795
Best	BTO-OTW	0.040	0.002	[0.015, 0.065]	0.5291
Other	BWT-BTO	0.027	0.014	[0.005, 0.048]	0.6543
Other	BWT-OTW	-0.033	0.002	[-0.054, -0.012]	-0.5856
Other	BTO-OTW	-0.059	< 0.001	[-0.080, -0.038]	-0.6553
Worst	BWT-BTO	-0.019	0.007	[-0.033, -0.005]	-0.3104
Worst	BWT-OTW	0.000	1.000	[-0.014, 0.014]	0.000
Worst	BTO-OTW	0.019	0.007	[0.005, 0.033]	0.3104

To test the effectiveness of BWT in producing more balanced weights and thus reducing the anchoring bias in the Tradeoff procedure, we compared the results of the BWT group with both the BTO and OTW groups. Compared to the BTO group, the BWT method resulted in significantly higher weights for the *other* attribute ($p = 0.014$) and significantly lower weights for the *worst* attribute ($p = 0.007$). Although the weights for the *best* attribute were also lower in the BWT group, this difference did not reach statistical significance ($p = 0.355$). In comparison to the OTW group, the BWT group produced significantly higher weights for the *best* attribute ($p = 0.032$) and significantly lower weights for the *other* attribute ($p = 0.002$). The BWT and OTW produced the same weights for the worst attribute ($p = 1$), consistent with expectations.

These results generally supported **H3**, indicating that the BWT method effectively reduced the anchoring bias. However, the non-significant difference in the best attribute between the

BWT and BTO groups ($p = 0.359$) was unexpected and warrants further investigation.

While the ANOVA highlighted differences in the magnitude of weights across groups, it did not capture the direction of change. To better understand how the anchoring bias systematically shifted weight elicitation under different conditions, we conducted Wilcoxon Signed-Rank tests. This non-parametric analysis focused on within-subject directional changes and provided complementary evidence that could reinforce the overall pattern of anchoring effects observed in the ANOVA and post hoc analyses. As shown in Table 3.5, the results not only aligned with but also extended the results of the ANOVA and post hoc analyses, with the previously non-significant difference in the *best* attribute between the BWT and BTO groups now reaching significance ($p = 0.028$), offering stronger support for the three hypotheses.

With respect to **H1** and **H2**, participants in the OTW group assigned significantly lower weights to the *best* and *worst* attributes ($p < 0.001$ for both) and significantly higher weights to the *other* attribute ($p < 0.001$) compared to those in the BTO group. These results confirmed the presence of the anchoring bias in the Tradeoff procedure, where it distorted the weights in opposite directions.

Regarding **H3**, the Wilcoxon Signed-Rank test showed that the BWT group produced significantly lower weights for the *best* attribute compared to the BTO group ($p = 0.028$), and significantly higher weights than the OTW group ($p < 0.001$). For the *other* attribute, weights in the BWT group were significantly higher than those in the BTO group ($p < 0.001$) and significantly lower than those in the OTW group ($p < 0.001$). For the *worst* attribute, the BWT group assigned significantly lower weights than the BTO group ($p < 0.001$). These statistically significant results indicated that the weights produced by the BWT method consistently fell between the two extremes of the BTO and OTW procedures, except for the *worst* attribute, where BWT and OTW produced the same weights ($p = 1$). These results supported **H3** and demonstrated that BWT was effective in mitigating the anchoring bias in the Tradeoff procedure.

Table 3.5: Wilcoxon signed-rank test

	BTOBest-BWTBest	**OTWBest-BWTBest**	**OTWBest-BTOBest**	**BTOOther-BWTOther**	**OTWOther-BWTOther**	**OTWOther-BTOOther**	**BTOWorst-BWTWorst**	**OTWWorst-BWTWorst**	**OTWWorst-BTOWorst**
Z	-2.200^b	-5.863^a	-5.071^a	-4.282^a	-6.267^b	-5.775^b	-7.753^b	0.000^c	-7.753^a
Sig.	0.028	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	1	< 0.001

[a] Based on positive ranks. [b] Based on negative ranks. [c] The sum of negative ranks equals the sum of positive ranks.

To better illustrate how the anchoring bias could distort the Tradeoff procedure, we presented the results from a single participant, demonstrating how the weights elicited for the same DM varied under different conditions due to this bias.

Participant 64 identified the importance order of the three attributes as follows: rent, commute distance, and distance to shopping center. Based on the attribute ranges (see Table 3.1) and the indifference pairs specified for this experiment (see Table 3.2), three indifference judgments were elicited, as summarized in Table 3.6.

After eliciting the attribute-specific value functions, the attribute weights were derived using BTO equation systems (3.12), OTW equation systems (3.13), and BWT optimization model (3.11). Table 3.7 presents the derived weights under the three tradeoff procedures.

As shown, noticeable differences emerged across the three conditions. For the most impor-

Table 3.6: All indifference pairs and responses for participant 64

Indifference Pairs	Attribute Comparison	Participant's Answer
BTO:P_1	$(1500, 5, 500) \sim (x_1^{1,2}, 15, 500)$	$x_1^{1,2} = 1000$
OTW:P_1	$(1500, x_2^{2,3}, 500) \sim (1500, 15, 100)$	$x_2^{2,3} = 10$
BTO&OTW:P_2	$(1500, 15, 100) \sim (x_1^{1,3}, 15, 500)$	$x_1^{1,3} = 1200$

Table 3.7: Weights for participant 64

Tradeoff Procedure	w_{Rent}	$w_{CommuteDistance}$	$w_{DistancetoShoppingCenter}$
BTO	0.600	0.250	0.150
BWT	0.593	0.264	0.143
OTW	0.571	0.286	0.143

tant attribute, rent, the BTO procedure yielded the highest weight (0.600), the OTW procedure yielded the lowest (0.571), and the BWT method produced an intermediate weight (0.593). For commute distance, the OTW procedure yielded the highest weight (0.286), the BTO procedure yielded the lowest (0.250), and the BWT method again produced an intermediate weight (0.264). For the least important attribute, distance to shopping center, the BTO procedure produced the highest weight (0.150), while the OTW and BWT procedures produced identical, lower weights (0.143).

Our experimental design was based on a relatively simple decision problem with three attributes. Real-world MADM applications often involve more attributes, sometimes ten or more (Belton & Stewart, 2012). When the number of attributes increases, each attribute's normalized weight tends to become smaller, which may reduce the influence of the anchoring bias in individual weights. However, the use of a simplified three-attribute setup allowed us to provide a clean and controlled test of the theoretically derived hypotheses regarding anchoring in weight elicitation, without the additional complexity that larger attribute sets might have introduced, such as cognitive overload or interaction effects (Payne et al., 1993; Keeney & Raiffa, 1976). At the same time, it was important to examine whether the observed effects had practical implications for decision outcomes.

To address this, we conducted a simulation study to test whether differences across the three weight elicitation methods (BTO, OTW, and BWT) led to differences in the final decision outcomes. We constructed and evaluated a set of simulated alternatives. Specifically, we generated five random numbers between zero and one that summed to one, and then mapped these numbers onto attribute levels within the experimental ranges. This procedure produced five non-dominated alternatives for each simulation. Using the additive aggregation model, we combined each participant's attribute-specific value functions and weights to obtain an overall value for each alternative and thus a complete ranking of the five alternatives under each weight elicitation method. As a result, each participant produced three distinct rankings of the same set of alternatives. To assess the extent to which these rankings agreed or disagreed with each other, we applied the Kendall τ_b coefficient (Kendall, 1948). This measure captured the correlation between two rankings, with a value of 1 indicating perfect agreement, 0 indicating no systematic association, and 1 indicating complete reversal of order.

To ensure that our findings were not dependent on the specific choice of alternatives, we

repeated this procedure 100 times with independently generated sets of five non-dominated alternatives. In each replication, we obtained τ_b values for the three method pairs (BWT–BTO, BWT–OTW, and BTO–OTW). We then calculated the average Kendall τ_b across the 100 replications, which provided a more stable estimate of the similarity in rankings produced by different weighting methods.

As shown in Table 3.8, the results of the one-sample *t*-tests showed that these average τ_b values were significantly below 1, indicating that the agreement between the rankings obtained from different weighting methods was far from perfect. This confirmed that the discrepancies in elicited weights across methods were not merely numerical differences but had practical consequences for the ranking of alternatives. Moreover, the τ_b for BTO-OTW comparison was the lowest among all, indicating that the anchoring bias pulled the two tradeoff procedures toward more extreme and divergent weights, whereas BWT helped to reduce these extremes.

Table 3.8: Results of one-sample t-tests of Kendall's τ_b (Test value = 1)

Tradeoff Method Comparison	**Mean τ_b**	**Std. Deviation**	t	**One-Sided p**
BWT-BTO	0.8077	0.0508	-37.852	< 0.001
BWT-OTW	0.8657	0.0379	-35.452	< 0.001
BTO-OTW	0.6915	0.0768	-40.171	< 0.001

These results aligned with our hypothesis, illustrating that the anchoring bias could distort the weights when only one type of tradeoff is used. Furthermore, they demonstrated that the BWT method can mitigate this bias by producing weights between the two extremes, thus reducing the influence of the anchoring bias.

3.7 Conclusion

In this study, we investigated the effect of the anchoring bias in the Tradeoff procedure within multi-attribute value theory. We found that the selection of the best or worst attribute to construct indifference pairs in the Tradeoff procedure can serve as an anchor and result in biased weights due to the anchoring bias. Specifically, using the best attribute leads to higher weights for the best and worst attributes and lower weights for the other attributes compared to using the worst attribute in the tradeoff procedures. Additionally, the BWT method was found to be effective in reducing the anchoring bias by incorporating both indifference pairs.

These findings align with previous research (Rezaei, 2021; Rezaei et al., 2024; Rezaei, 2022), indicating that weight elicitation methods using a single reference point (e.g., SMART, Swing, or one vector of BWM) are susceptible to the anchoring bias, while methods employing two counter-reference points, such as the BWM, can reduce its impact. It is worth noting that, although these studies all examined weight elicitation, they focused on different methods, each with distinct interpretations of weights and distinct procedures that introduce the anchoring bias. This study extends this line of research by examining the Tradeoff procedure. While Liang et al. (2022) proposed that BWT has an inherent debiasing mechanism for the anchoring bias, empirical evidence supporting this claim has been lacking. This study addresses this gap by providing empirical evidence for the debiasing effect of BWT.

This study enhances our understanding of the anchoring bias in weight elicitation methods, offering important implications for both the theoretical and practical aspects of the Tradeoff procedure. It also underscores the need to account for cognitive biases, particularly the anchoring bias, when developing and applying MADM methods. While the structure of a MADM method may unintentionally introduce such biases, thoughtful design can help mitigate their effects.

The limitations of this study should be acknowledged. First, the sample is skewed toward highly educated participants (about two-thirds with a Bachelor's or Master's degree) and contained a gender imbalance (one-third female, two-thirds male), from six European countries, which may restrict the generalizability of the findings (Henrich et al., 2010). Higher education is often associated with stronger numeracy and literacy, which facilitates understanding of structured elicitation tasks (Peters et al., 2006). Gender can also influence preference and susceptibility to heuristics (Croson & Gneezy, 2009). Although we found no demographic effects in our ANCOVA analysis, future research should explicitly test whether our findings can be generalized across populations with different educational backgrounds and more balanced gender distributions. Second, the decision problem is simplified, involving only three attributes, which may not fully capture the complexity of real-world decision-making. Moreover, under three attributes, the BWT and Other-to-Worst tradeoff procedures produce the same weight for the worst attribute, limiting the ability to fully observe the debiasing effects of the BWT method. Future research could address these limitations by investigating decision problems with more than three attributes, both to further explore the debiasing role of BWT and to test whether the anchoring bias persists in more complex and realistic settings. Such extensions would be particularly relevant in high-stakes application areas where biased decisions can lead to substantial negative consequences, such as healthcare, infrastructure planning, or supplier selection. Compared to the traditional Tradeoff procedure, which requires only $N-1$ indifference pairs, BWT requires $2N-3$ for N attributes (Liang et al., 2022). This additional effort, combined with more attributes, may increase cognitive load. It remains an open question whether BWT can still mitigate the anchoring bias effectively under such conditions. Still, BWT remains far less demanding than pairwise comparison approaches, which require $N(N1)/2$ (unidirectional) or $N(N1)$ (bidirectional) judgments (Liang et al., 2022). This suggests that BWT may provide a favorable balance between cognitive effort and bias reduction, but its performance in larger-scale decision problems should be systematically investigated. In addition, recruiting a more diverse and global participant pool would help assess the robustness of the findings across different populations. Finally, future studies could examine whether similar patterns of the anchoring bias occur across different weight elicitation methods.

Bibliography

Adame, B. J. (2016) Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy, *Learning and Motivation*, 53, pp. 36–48.

Beinat, E. (1997) *Value functions for environmental management*, Springer, Dordrecht.

Belton, V., T. Stewart (2012) *Multiple criteria decision analysis: an integrated approach*, Springer, New York.

Buchanan, J. T. (1997) A naive approach for solving mcdm problems: The guess method, *Journal of the Operational Research Society*, 48(2), pp. 202–206.

Buchanan, J. T., J. Corner (1997) The effects of anchoring in interactive mcdm solution methods, *Computers & Operations Research*, 24(10), pp. 907–918.

Bystranowski, P., B. Janik, M. Próchnicki, P. Skórska (2021) Anchoring effect in legal decision-making: A meta-analysis, *Law and Human Behavior*, 45(1), pp. 1–23.

Campbell, S. D., S. A. Sharpe (2009) Anchoring bias in consensus forecasts and its effect on market prices, *Journal of Financial and Quantitative Analysis*, 44(2), pp. 369–390.

Chapman, G. B., E. J. Johnson (1999) Anchoring, activation, and the construction of values, *Organizational Behavior and Human Decision Processes*, 79(2), pp. 115–153.

Chapman, G. B., E. J. Johnson (2002) Incorporating the irrelevant: Anchors in judgments of belief and value, in: *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, Cambridge, pp. 120–138.

Charness, G., U. Gneezy, M. A. Kuhn (2012) Experimental methods: Between-subject and within-subject design, *Journal of Economic Behavior & Organization*, 81(1), pp. 1–8.

Choo, E. U., B. Schoner, W. C. Wedley (1999) Interpretation of criteria weights in multicriteria decision making, *Computers & Industrial Engineering*, 37(3), pp. 527–541.

Croson, R., U. Gneezy (2009) Gender differences in preferences, *Journal of Economic Literature*, 47(2), pp. 448–474.

Edwards, W. (1977) How to use multiattribute utility measurement for social decisionmaking, *IEEE Transactions on Systems, Man, and Cybernetics*, 7(5), pp. 326–340.

Englich, B., T. Mussweiler, F. Strack (2006) Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making, *Personality and Social Psychology Bulletin*, 32(2), pp. 188–200.

Epley, N., T. Gilovich (2005) When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors, *Journal of Behavioral Decision Making*, 18(3), pp. 199–212.

Fischer, G. W. (1995) Range sensitivity of attribute weights in multiattribute value models, *Organizational Behavior and Human Decision Processes*, 62(3), pp. 252–266.

Fischer, G. W., N. Damodaran, K. B. Laskey, D. Lincoln (1987) Preferences for proxy attributes, *Management Science*, 33(2), pp. 198–214.

Fishburn, P. C. (1967) Methods of estimating additive utilities, *Management Science*, 13(7), pp. 435–453.

Furnham, A., H. C. Boo (2011) A literature review of the anchoring effect, *The Journal of Socio-Economics*, 40(1), pp. 35–42.

Gilovich, T., D. Griffin, D. Kahneman (2002) *Heuristics and biases: The psychology of intuitive judgment*, Cambridge University Press, Cambridge.

Greenwald, A. G. (1976) Within-subjects designs: To use or not to use?, *Psychological Bulletin*, 83(2), pp. 314–320.

Hämäläinen, R. P., A. A. Salo (1997) The issue is understanding the weights, *Journal of Multi-Criteria Decision Analysis*, 6, pp. 340–343.

Henrich, J., S. J. Heine, A. Norenzayan (2010) The weirdest people in the world?, *Behavioral and Brain Sciences*, 33(2-3), pp. 61–83.

Jaspersen, J. G., G. Montibeller (2015) Probability elicitation under severe time pressure: A rank-based method, *Risk Analysis*, 35(7), pp. 1317–1335.

Keeney, R. L. (1974) Multiplicative utility functions, *Operations Research*, 22(1), pp. 22–34.

Keeney, R. L., H. Raiffa (1976) *Decisions with multiple objectives: Preferences and value trade-offs*, Cambridge University Press, Cambridge.

Kendall, M. G. (1948) Rank correlation methods., in: *Public Program Analysis*, Springer, Boston, pp. 146–163.

Kirkwood, C. W., R. K. Sarin (1980) Preference conditions for multiattribute value functions, *Operations Research*, 28(1), pp. 225–232.

Larichev, O. I. (1992) Cognitive validity in design of decision-aiding techniques, *Journal of Multi-Criteria Decision Analysis*, 1(3), pp. 127–138.

Larichev, O. I., R. V. Brown (2000) Numerical and verbal decision analysis: comparison on practical cases, *Journal of Multi-Criteria Decision Analysis*, 9(6), pp. 263–273.

Liang, F., M. Brunelli, J. Rezaei (2022) Best-worst tradeoff method, *Information Sciences*, 610, pp. 957–976.

Ly, D. P., P. G. Shekelle, Z. Song (2023) Evidence for anchoring bias during physician decision-making, *JAMA Internal Medicine*, 183(8), pp. 818–823.

McElroy, T., K. Dowd (2007) Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues, *Judgment and Decision Making*, 2(1), pp. 48–53.

Meub, L., T. Proeger (2016) Can anchoring explain biased forecasts? experimental evidence, *Journal of Behavioral and Experimental Finance*, 12, pp. 1–13.

Montibeller, G., D. von Winterfeldt (2015) Cognitive and motivational biases in decision and risk analysis, *Risk Analysis*, 35(7), pp. 1230–1251.

Montibeller, G., D. von Winterfeldt (2024) Behavioral decision research: Descriptive and prescriptive perspectives, in: *Behavioral Decision Analysis*, Springer, Cham, pp. 15–40.

Morewedge, C. K., H. Yoon, I. Scopelliti, C. W. Symborski, J. H. Korris, K. S. Kassam (2015) Debiasing decisions: Improved decision making with a single training intervention, *Policy Insights from the Behavioral and Brain Sciences*, 2(1), pp. 129–140.

Morton, A., B. Fasolo (2009) Behavioural decision theory for multi-criteria decision analysis: a guided tour, *Journal of the Operational Research Society*, 60(2), pp. 268–275.

Mussweiler, T., F. Strack (2001) The semantics of anchoring, *Organizational Behavior and Human Decision Processes*, 86(2), pp. 234–255.

Mussweiler, T., F. Strack, T. Pfeiffer (2000) Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility, *Personality and Social Psychology Bulletin*, 26(9), pp. 1142–1150.

Nagtegaal, R., L. Tummers, M. Noordegraaf, V. Bekkers (2020) Designing to debias: Measuring and reducing public managers' anchoring bias, *Public Administration Review*, 80(4), pp. 565–576.

Palan, S., C. Schitter (2018) Prolific.ac—a subject pool for online experiments, *Journal of Behavioral and Experimental Finance*, 17, pp. 22–27.

Payne, J. W., J. R. Bettman, E. J. Johnson (1992) Behavioral decision research: A constructive processing perspective, *Annual Review of Psychology*, 43, pp. 87–131.

Payne, J. W., J. R. Bettman, E. J. Johnson (1993) *The adaptive decision maker*, Cambridge University Press, New York.

Peters, E., D. Västfjäll, P. Slovic, C. Mertz, K. Mazzocco, S. Dickert (2006) Numeracy and decision making, *Psychological Science*, 17(5), pp. 407–413.

Pöyhönen, M., R. P. Hämäläinen (2000) There is hope in attribute weighting, *INFOR: Information Systems and Operational Research*, 38(3), pp. 272–282.

Rezaei, J. (2015) Best-worst multi-criteria decision-making method, *Omega*, 53, pp. 49–57.

Rezaei, J. (2021) Anchoring bias in eliciting attribute weights and values in multi-attribute decision-making, *Journal of Decision Systems*, 30(1), pp. 72–96.

Rezaei, J. (2022) The balancing role of best and worst in best-worst method, in: *Advances in Best-Worst Method: Proceedings of the Second International Workshop on Best-Worst Method (BWM2021)*, Springer, pp. 1–15.

Rezaei, J., A. Arab, M. Mehregan (2022) Equalizing bias in eliciting attribute weights in multiattribute decision-making: experimental research, *Journal of Behavioral Decision Making*, 35(2), e2262.

Rezaei, J., A. Arab, M. Mehregan (2024) Analyzing anchoring bias in attribute weight elicitation of smart, swing, and best-worst method, *International Transactions in Operational Research*, 31(2), pp. 918–948.

Saaty, T. L. (1980) The analytic hierarchy process (ahp), *The Journal of the Operational Research Society*, 41(11), pp. 1073–1076.

Singer, E., C. Ye (2013) The use and effects of incentives in surveys, *The ANNALS of the American Academy of Political and Social Science*, 645(1), pp. 112–141.

Slovic, P. (1995) The construction of preference, *American Psychologist*, 50(5), pp. 364–371.

Smith, J. E., J. S. Dyer (2021) On (measurable) multiattribute value functions: An expository argument, *Decision Analysis*, 18(4), pp. 247–256.

Sun, G., M. Kroesen, J. Rezaei (2025) Anchoring bias in value function elicitation within multi-attribute value theory, *Decision Analysis*, 22(4), pp. 284–304.

Thirsk, L. M., J. T. Panchuk, S. Stahlke, R. Hagtvedt (2022) Cognitive and implicit biases in nurses' judgment and decision-making: a scoping review, *International Journal of Nursing Studies*, 133, 104284.

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

von Nitzsch, R., M. Weber (1993) The effect of attribute ranges on weights in multiattribute utility measurements, *Management Science*, 39(8), pp. 937–943.

von Winterfeldt, D., W. Edwards (1993) *Decision analysis and behavioral research*, Cambridge University Press, Cambridge.

Watson, S. R., D. M. Buede (1987) *Decision synthesis: The principles and practice of decision analysis*, Cambridge University Press, Cambridge.

Weber, M., K. Borcherding (1993) Behavioral influences on weight judgments in multiattribute decision making, *European Journal of Operational Research*, 67(1), pp. 1–12.

Weber, M., F. Eisenführ, D. von Winterfeldt (1988) The effects of splitting attributes on weights in multiattribute utility measurement, *Management Science*, 34(4), pp. 431–445.

Wilson, T. D., C. E. Houston, K. M. Etling, N. Brekke (1996) A new look at anchoring effects: basic anchoring and its antecedents., *Journal of Experimental Psychology: General*, 125(4), pp. 387–402.

Wu, Q., X. Liu, L. Zhou, J. Qin, J. Rezaei (2024) An analytical framework for the best–worst method, *Omega*, 123, 102974.

Zionts, S., J. Wallenius (1983) An interactive multiple objective linear programming method for a class of underlying nonlinear utility functions, *Management Science*, 29(5), pp. 519–529.

Chapter 4

Framing and Loss Aversion in Decisions with Multiple Objectives

---

Abstract: The asymmetric way individuals respond to equivalent gains versus losses (i.e., the gain-loss bias) has been extensively studied in decisions with a single objective. Yet, its role in decisions with multiple objectives remains largely underexplored, where tradeoffs among objectives must be made. In such contexts, outcomes depend not only on how the decision-maker values performance on each objective but also on how tradeoffs are constructed, making the influence of gain–loss bias more complex. This study examines how two sources of gain-loss bias, the framing effect and the loss aversion, influence the multi-attribute value theory, a theory of decision making with multiple objectives. We develop and test three hypotheses on (i) how loss aversion affects the tradeoff procedure, (ii) how framing effect shapes the attribute-specific value functions, and (iii) how these two biases interact across elicitation stages. To test these hypotheses, we design a water-filter decision problem and collect data from 283 subjects. The results indicate that the gain-loss bias has a significant effect on the attribute-specific value functions, the tradeoffs, and the decision outcome. Furthermore, we demonstrate that the interactions between biases can potentially cancel out their individual effects, providing valuable insights for bias mitigation. We also test other practical bias mitigation strategies, such as group decision-making, which are shown to be effective in reducing these biases.

Keyword: Gain-loss bias; bias mitigation; multiple objectives; tradeoff; multi-attribute value theory

This chapter is based on the following manuscript:

Sun, G., Kroesen, M., Rezaei, J. Framing and Loss Aversion in Decisions with Multiple Objectives (*manuscript under review*)

4.1 Introduction

In real-world decision-making, individuals often deviate from normative principles due to cognitive biases. One of the most extensively studied biases is the gain-loss bias, where people respond differently to equivalent gains and losses. This phenomenon is evident in two main ways: (i) as a framing effect, where different descriptions of the same outcome (e.g., framed as gains vs. losses) lead to different preferences (Tversky & Kahneman, 1981), and (ii) as loss aversion, where outcomes are evaluated relative to a reference point and losses loom larger than equivalent gains (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991).

Tversky & Kahneman (1981) introduced a famous framing effect example using the Asian Disease Problem. The participants were asked to select programs after a disease was expected to kill 600 people. When the outcomes were framed as gains (e.g., 200 lives saved vs. 1/3 probability of 600 lives saved and 2/3 probability of no lives saved), most participants preferred the sure option; when framed as losses (e.g., 400 lives lost vs. 1/3 probability of no lives lost and 2/3 probability of 600 lives lost), most participants preferred the risky option. This reversal in preference demonstrated how logically equivalent scenarios can lead to different preferences only based on framing.

Kahneman & Tversky (1979) laid the foundation of loss aversion with the development of prospect theory. They showed that individuals evaluate outcomes relative to a reference point and consistently value losses more heavily than equivalent gains. The famous coffee mug trade problem provides direct evidence of loss aversion (Knetsch, 1989; Kahneman et al., 1991). Participants randomly received coffee mugs, candies, or nothing and were later asked to trade them or state a selling price. They found that sellers demanded significantly higher prices than buyers were willing to pay. This indicated that giving up the mug was perceived as a loss and thus valued more than the gain of acquiring it.

These studies illustrate how gain-loss bias emerges consistently in single-objective and binary choice settings across diverse contexts, including medical decision-making (Gong et al., 2013; McNeil et al., 1982), goal-directed behavior (Heath et al., 1999; Allen et al., 2017), consumer behavior (Levin & Gaeth, 1988), retirement saving behavior (Benartzi & Thaler, 1995), and risky financial decisions (Tom et al., 2007). Yet, relatively little attention has been paid to their implications for decisions with multiple objectives.

In practice, many high-stakes decisions, such as those in engineering, environment, or healthcare fields, require evaluating multiple, often conflicting objectives. These objectives are represented by attributes, which provide measurable scales for evaluating alternatives. Since these objectives are often conflicting, improvements in one attribute often require sacrifices in another, tradeoffs between them must be made in the decision analysis process. In this context, gain-loss bias may not only distort how individual attribute levels are valued but also how attributes are weighted. For example, if one attribute is framed as a gain and another as a loss, the loss attribute might have a bigger impact on the tradeoff judgment, leading to biased weights compared to when both attributes are framed as gains. As the evaluation of an alternative is affected by both the value of the attribute level and the weights of the attributes, biases in the two components can interact in complex ways. Their combined effects may amplify, offset, or otherwise complicate the influence of the gain-loss bias on the final decision outcome. In high-stakes decision areas, such distortions may translate into major financial losses, misallocated

resources, and unintended societal consequences.

There are many studies that have focused on how the gain-loss bias influences unaided decisions within a multi-attribute environment (Bier & Connell, 1994; Kim et al., 2014), but they do not address the bias in a formal decision analysis method. As a result, important questions, such as where and how biases enter the decision analysis process, whether decision analysis methods are robust to gain-loss bias, and how bias mitigation strategies can be developed and tested, remain unsolved.

The aim of this study is to investigate the effect of the gain-loss bias in aided decisions with multiple objectives, to examine this issue within a multi-attribute decision-making method, and to develop practical bias mitigation strategies. We developed three hypotheses regarding the framing effect, loss aversion, and their interaction[1] within multiple stages of a multi-attribute decision-making method.

To test the hypotheses, we developed a water-filter purchase problem with two attributes and designed an experiment following the steps of a multi-attribute decision-making method. The experiment was carried out in a questionnaire developed using Qualtrics, and data were collected from 283 participants via Prolific. We also explored several practical bias mitigation strategies, which were found to be effective. These included (i) using a homogeneous frame for all attributes; (ii) exposing DM to both gain and loss conditions and averaging the results; (iii) employing group decision-making; and (iv) using the potential cancellation effect between the two biases.

This study makes several contributions to the literature on behavioral decision-making and multi-attribute decision-making. First, it extends the literature on gain-loss bias within aided decision-making to multiple-objectives decisions. Specifically, it is the first to empirically demonstrate how gain-loss bias influences the attribute-specific value functions and weights. Second, by clearly identifying different sources and forms of gain-loss bias and examining their interactions, this study contributes to a deeper understanding of how biases arise and interact across different stages of multi-attribute decision-making methods. Lastly, prior behavioral decision research primarily offers descriptive insights, this study provides prescriptive insights on how to reduce the gain-loss bias in multi-attribute decision-making methods.

The remainder of the paper is structured as follows. Section 4.2 reviews related work on gain-loss bias. Section 4.3 introduces the multi-attribute decision-making method. Section 4.4 presents the hypotheses. Section 4.5 describes the experimental design. Section 4.6 discusses empirical results and the proposed bias mitigation strategies. Section 4.7 concludes with theoretical and practical implications.

4.2 Gain-Loss Bias and its Role in Multi-Attribute Decision-Making Methods

The term gain-loss bias used in this study refers to the asymmetric way in which individuals respond to equivalent gains versus losses (Montibeller & von Winterfeldt, 2015). In the multi-

[1]The term *interaction* refers to the mutual influence between biases or between elicitation stages of the multi-attribute decision-making method in shaping the final decision outcome.

attribute decision-making context, this concept provides a broad lens for understanding deviations from rational choice that arise from different types of gains and losses. The gain-loss bias includes two related but distinct mechanisms: loss aversion and the framing effect. Although both are rooted in sensitivity to gains and losses, they operate through different psychological processes, and they can lead to different empirical findings.

Prospect theory (Kahneman & Tversky, 1979) provides the theoretical basis for understanding gain-loss bias. It posits that individuals evaluate outcomes relative to a reference point, giving rise to a value function that is concave for gains, convex for losses, and steeper in the loss domain (Tversky & Kahneman, 1991). This reference-dependent asymmetry, known as loss aversion, was initially studied in risky choice but later extended to riskless decision contexts as well (Thaler, 1980; Knetsch, 1989; Hardie et al., 1993; Carmon et al., 2003). Gächter et al. (2022) measured loss aversion by comparing participants' willingness to accept compensation for giving up an item versus their willingness to pay to acquire it. They found that loss aversion is stronger in riskless conditions than in risky ones, highlighting its relevance to multi-attribute decision-making contexts, where many decision problems involve tradeoffs among deterministic attributes and clearly defined alternatives.

The framing effect emerges directly from this reference-dependent evaluation. When equivalent outcomes are described differently, individuals can exhibit systematically different preferences (Tversky & Kahneman, 1981). Framing can take multiple forms, including risky-choice framing, attribute framing, and goal framing, each operating through different psychological mechanisms (Levin et al., 1998). Empirical studies (Levin et al., 2002; Barnes et al., 2025; Payne et al., 2013; Meyerowitz & Chaiken, 1987; Ganzach & Karsahi, 1995) demonstrate that these framing types are largely independent, and that different forms of framing can produce distinct patterns of behavior. This study focuses on attribute framing, which is about how the attributes are framed.

Research on loss aversion and the framing effect has led to the observation of a wide range of behavioral patterns that deviate from classical rationality assumptions (Levy, 1992). These biases have been extensively studied across domains such as finance, healthcare, environmental decision-making, and consumer behavior (Gong et al., 2013; Benartzi & Thaler, 1995; Klapper et al., 2005; Almashat et al., 2008; Barberis, 2013). For example, Genesove & Mayer (2001) provided evidence of loss aversion in the real estate market, showing that homeowners are reluctant to sell a property for less than its purchase price, even when doing so would be financially rational.

Given the widespread impact of loss aversion and the framing effect, researchers have developed various strategies to mitigate their influence. General debiasing techniques such as raising awareness of biases, using experienced DM, and increasing cognitive effort have been shown to reduce susceptibility to both biases (Cheng & Wu, 2010; Fu et al., 2018; Mrkva et al., 2020; Thomas & Millar, 2012; Beratšová et al., 2016). Additionally, individual differences such as gender, cultural background, and cognitive or emotional capacity can influence the degree of susceptibility to these biases (Bibby & Ferguson, 2011; Miu & Crişan, 2011; Cassotti et al., 2012; Sokol-Hessner et al., 2009). Besides these, more targeted strategies grounded on the choice architecture literature include using frame-neutral presentations, exposing individuals to both gain and loss frames, and group decision-making (Almashat et al., 2008; Garcia-Retamero & Dhami, 2013; Yaniv, 2011; Montibeller & von Winterfeldt, 2015; Hammond et al., 1998). Compare to generic debiasing strategies, these targeted strategies are particularly relevant when

integrating behavioral insights into decision analysis methods, as they act on the mechanisms of the biases rather than on DMs. By directly addressing the source of the bias, these strategies can be incorporated into the structure of decision analysis methods, thereby improving the reliability of the elicited judgments and preferences.

In the context of multi-attribute decision-making, Bier & Connell (1994) found that individuals are ambiguity-seeking in positively framed scenarios, which contradicts the prospect theory that posits individuals are risk-averse in the gain domain. This suggests that people may employ different cognitive strategies in multi-attribute settings compared to single-attribute settings, underscoring the need for further research on the framing effect in more complex decision-making environments. Kim et al. (2014) examined both attribute framing and goal framing in multi-attribute decisions, focusing on two components unique in this context: attribute-level evaluation and attribute weight elicitation, aspects that cannot be separated in single-attribute tasks. Their results showed that the evaluation of the positively framed attribute is higher than that of the negative condition, and the weight of the framed attribute is greater in positive goal framing than in the negative condition. These findings suggest that framing can systematically distort both the perceived importance and evaluation of attributes in multi-attribute decision-making. While these studies demonstrate that gain-loss bias can influence multi-attribute judgments in unaided decisions, they do not examine how such biases operate within formal decision analysis methods. Decision analysis methods provide a structured way of evaluating decision problems. Whether this very structure may create new entry points for bias, reduce bias, or amplify it, remains an open question.

An important line of research addressing this gap has explored loss aversion in the tradeoff procedure. Several violations of procedural invariance have been observed in this context. Delquié (1993) showed that equivalent tradeoffs can yield inconsistent results depending on which attribute is adjusted. In their two-stage design, participants first adjusted one attribute to reach indifference between two options and then adjusted the other attribute for the same indifference condition. Although the two stages were logically equivalent, participants gave systematically different responses depending on which attribute they adjusted. To mitigate the loss aversion in the traditional tradeoff procedure, Delquié (1997) developed a new procedure, "bi-matching", where participants simultaneously adjust both attributes to find an indifference point. Building on Delquie's work, Bleichrodt & Pinto (2002) designed a more elaborate version of the tradeoff task to disentangle the effects of loss aversion and scale compatibility (the tendency to give more weight to the attribute that is used as the response scale). Their results showed distinct patterns of deviation depending on the adjusted attributes, suggesting that both biases operate independently and can systematically distort the elicited tradeoff values. Notably, the study also highlighted that the interaction between the two biases can sometimes reduce the overall distortion, offering prescriptive insights.

Several research gaps remain unaddressed in the literature. First, gain-loss bias has received limited attention in aided multi-attribute decision-making settings. In particular, while framing effects have been widely studied in single-attribute, choice-based contexts, as well as in unaided multi-attribute decisions, their influence on decision analysis methods has not been systematically explored. Montibeller & von Winterfeldt (2015) identified gain-loss bias in several stages of decision analysis methods, but empirical evidence on its effect and potential mitigation strategies remains lacking.

Second, previous research on loss aversion in tradeoff procedures has primarily investigated

procedural consistency (Delquié, 1993; Bleichrodt & Pinto, 2002), focusing on how adjusting different attributes in a two-stage tradeoff task leads to inconsistent matching responses. In contrast, our study investigates how reversing the direction of adjustment for the same attribute (gaining from the lowest level vs. losing from the highest level) affects the elicited weights. This represents a distinct mechanism through which loss aversion can influence tradeoff judgments. Moreover, we focus on whether such inconsistencies translate into systematic distortions in the derived attribute weights. By emphasizing the tradeoff procedure as a weight elicitation method rather than only as a judgmental task, we uncover bias mechanisms that previous studies could not capture.

Third, as noted by Bleichrodt & Pinto (2002), different cognitive biases can interact in complex ways, which may amplify or mitigate their individual effects. This study investigates two such interactions: (i) the interaction across elicitation stages within the multi-attribute decision-making method, and (ii) the interaction between framing effects and loss aversion.

Finally, this study emphasizes not only identifying the effects of gain-loss bias but also exploring strategies to mitigate them. By focusing on how biases interact and propagate across different steps of the multi-attribute decision-making method, it offers insights into step-specific and potentially unified bias mitigation strategies throughout the entire process.

4.3 Multi-Attribute Value Theory

Multi-attribute value theory (MAVT), originally developed by Keeney & Raiffa (1976), is a widely used multi-attribute decision-making method for evaluating alternatives characterized by multiple attributes. Its central idea is to represent a DM's preferences as a single overall value for each alternative, enabling the alternatives to be compared and ranked on a common value scale. This overall value is determined by a value function constructed from attribute-specific value functions and scaling constants (weights) that capture the tradeoff among attributes.

The process begins with identifying objectives, alternatives, and attributes. The objectives are represented by the attributes, and the alternatives are described by their performance levels with respect to those attributes. This early step can already introduce attribute framing bias, as the same attribute can be formulated in different ways (e.g., lives saved vs. lives lost), which might affect the subsequent attribute evaluations.

The attribute-specific value function translates the attribute performance levels to a normalized value scale, usually ranging from 0 (least preferred) to 1 (most preferred). There are different elicitation methods, such as the midvalue splitting procedure, the lock-step procedure, the standard difference procedure, and curve fitting (Beinat, 1997; Fishburn, 1967; Watson & Buede, 1987; Keeney & Raiffa, 1976). This study uses the midvalue splitting procedure, as it is one of the original methods developed within MAVT. The procedure defines a monotonic value function that reflects the DM's preferences over the attribute range. When higher attribute levels are preferred, the procedure begins by assigning a value score of 0 to the least preferred level and 1 to the most preferred level. The DM is then asked to identify an indifference point between these two levels such that the improvement from the least preferred level to this midvalue point is considered equally desirable as the improvement from this midvalue point to the most preferred level. This indifference point is the first midvalue point and is assigned a value

score of 0.5. The same logic is then applied iteratively within the resulting intervals to identify subsequent midvalue points (i.e., 0.25, 0.75). Additional midvalue points can be identified to obtain a more detailed representation of preferences. These midvalue points can then be used to plot the attribute-specific value functions. If lower levels of the attribute are preferred, the procedure is mirrored by assigning a value of 1 to the lowest level and 0 to the highest level, while following the same elicitation steps.

The tradeoff procedure is developed within MAVT to derive the scaling constants (weights). First, the DM is presented with a set of hypothetical alternatives. In each alternative, one attribute is set to its best performance level, while all others are at their worst performance levels. The DM ranks these alternatives, and the attribute at its best performance level in the most preferred alternative is identified as the most important attribute x_B in the decision context. Then, a series of indifference pairs is constructed to quantify the tradeoffs between the most important attribute and each of the remaining attributes. In each pair, one alternative is defined with the most important attribute x_B at its worst performance level and another attribute x_k at its best performance level; in the second alternative, the attribute x_k is at its worst performance level, and the DM will adjust the level of the most important attribute until she is indifferent between the two alternatives. All other attributes not included will be fixed during this procedure. The level of the most important attribute at indifference, denoted $x_B^{B,k}$, reflects the tradeoff the DM is willing to make between x_B and x_k. Repeating this process for all other attributes yields $N-1$ indifference pairs for N attributes, together with the constraint that the weights sum to one, the weights of the attributes can be calculated.

Once both the attribute-specific value functions and weights are elicited, they are integrated into the overall value function. This overall value function can be additive or non-additive, based on the DM's preference structure over the attributes (Keeney, 1974; Keeney & Raiffa, 1976). The additive model is the most widely used one, its validity depends on two key preference conditions: mutual preference independence and difference independence (Smith & Dyer, 2021; Dyer & Sarin, 1979; Keeney & Raiffa, 1976). Mutual preferential independence requires that any subset of attributes is preferentially independent of the remaining attributes, while difference independence requires that the preference difference between two levels of a given attribute is not influenced by the levels of the other attributes.

The additive value function is defined as follows (Keeney & Raiffa, 1976; Keeney, 2009):

$$v(a_i) = \sum_{j=1}^{N} w_j v_j(a_{ij}) \tag{4.1}$$

where $v(a_i)$ is the overall value of alternative i, scaled from 0 to 1. $v_j(a_{ij})$ is the attribute-specific value representing the performance of alternative i with respect to attribute j, and w_j is the scaling constant (or weight) of the attribute j.

The final step of MAVT is to compare and rank the alternatives based on their overall values, or to select the alternative with the highest overall value. Formally, for any two alternatives a_k and a_l, $v(a_k) \geq v(a_l) \Leftrightarrow a_k \succsim a_l$, where the symbol $\succsim$ denotes "preferred or indifferent to" (Keeney & Raiffa, 1976), reflecting the ordering implied by the overall value function.

4.4 Hypotheses Development

This section develops three hypotheses on how gain–loss bias can affect multiple stages of MAVT.

4.4.1 Loss Aversion in the Tradeoff Procedure

In MAVT, attribute weights are commonly elicited through the tradeoff procedure, where the level of the most important attribute is adjusted to compensate for changes in another attribute being traded off, forming a set of indifference pairs. This procedure can be implemented in two directions, framed as a gain or a loss on the most important attribute. Although the two frames of tradeoff are logically the same, individuals may have different interpretations of the same amount of gain and loss (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991), which could lead to biased indifference judgments and biased weights being elicited.

As shown in Table 4.1, in the gain-framed tradeoff, there is a gain on the most important attribute x_B from its lowest level to a point $x_B^{B,k}$ at which the DM is indifferent between the two options. The attribute x_k is fixed at the highest performance level in the first alternative, and is fixed at the lowest performance level in the second alternative. All other attributes not involved in this tradeoff are fixed.

Table 4.1: Gain-framed tradeoff indifference pairs

Indifference Pair	Direction of Change	Tradeoff Frame
$(\underline{x}_1, \ldots, \underline{x}_B, \ldots, \overline{x}_k, \ldots \underline{x}_N) \sim (\underline{x}_1, \ldots, x_B^{B,k}, \ldots, \underline{x}_k, \ldots, \underline{x}_N)$	Gain on x_B	Gain-framed Tradeoff

As introduced in Section 3, the tradeoff procedure will result in $N-1$ indifference pairs. Solving the resulting system of equations together with the constraint that the weights sum to one leads to the following expressions for the weights under the gain-framed tradeoff procedure:

$$w_B^{gain} = \frac{1}{1 + \sum_{j \neq B} v(x_B^{B,j})} \tag{4.2}$$

$$w_k^{gain} = \frac{v(_B^{B,k})}{1 + \sum_{j \neq B} v(x_B^{B,j})}, \quad k \neq B \tag{4.3}$$

In contrast, there is a loss on the most important attribute x_B for the loss-framed tradeoff (see Table 4.2). Attribute x_k is fixed at the lowest performance level in the first alternative, and it is fixed at the highest performance level in the second alternative. The DM adjusts attribute x_B from its highest level to a lower level $x_{B'}^{B,k}$ such that she is indifferent between the two options.

Table 4.2: Loss-framed tradeoff indifference pairs

Indifference Pair	Direction of Change	Tradeoff Frame
$(\underline{x}_1, \ldots, \overline{x}_B, \ldots, \underline{x}_k, \ldots \underline{x}_N) \sim (\underline{x}_1, \ldots, x_{B'}^{B,k}, \ldots, \overline{x}_k, \ldots, \underline{x}_N)$	Loss on x_B	Loss-framed Tradeoff

This loss-framed tradeoff also results in $N-1$ indifference pairs, and with the constraint that

the weights sum to one, the attribute weights derived from the loss-framed tradeoff procedure are:

$$w_B^{loss} = \frac{1}{n - \sum_{j \neq B} v(x_{B'}^{B,j})} \tag{4.4}$$

$$w_k^{loss} = \frac{1 - v(\overline{x}_{B'}^{B,k})}{n - \sum_{j \neq B} v(x_{B'}^{B,j})}, \quad k \neq B \tag{4.5}$$

Normatively, the two tradeoff procedures should yield equivalent weights as they are derived from a logically equivalent set of indifference pairs. However, loss aversion in prospect theory suggests that people value the same amount of gains and losses differently (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991), and this might affect the value judgment during the tradeoff procedure and lead to biased weights. Figure 4.1 illustrates how framing may introduce asymmetries in how tradeoffs are evaluated.

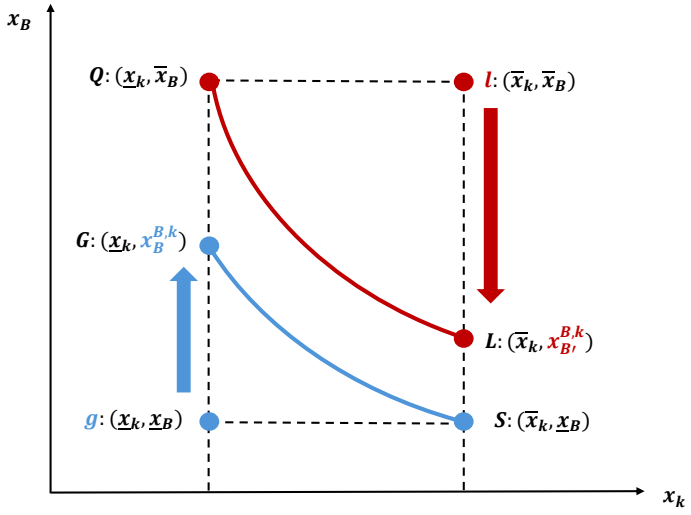


Figure 4.1: An illustration of loss aversion in the tradeoff procedure

In the gain-framed tradeoff, the DM adjusts $x_B^{B,k}$ to identify the indifference relation between options S and G. As shown in Table 4.1, this adjustment process of attribute x_B starts from its worst performance level, $\underline{x}_B$, until indifference is reached at level $x_B^{B,k}$ in option G. Since the attribute x_k in option G is fixed at its worst performance level $\underline{x}_k$, this setup implicitly positions $g : (\underline{x}_k, \underline{x}_B)$ as the reference point, and implies that the advantage of moving from $\underline{x}_B$ to $x_B^{B,k}$ is the same as the improvement from $\underline{x}_k$ to $\overline{x}_k$ on the other attribute.

In contrast, the loss-framed tradeoff requires identifying $x_{B'}^{B,k}$ from a reference point l such that the DM is indifferent between option Q and L. Here, the adjustment starts from the highest level $\overline{x}_B$ and moves downward, until indifference is reached at $x_{B'}^{B,k}$ in option L. Since the attribute x_k in option L is fixed at its highest performance level $\overline{x}_k$, this setup positions $l : (\overline{x}_k, \overline{x}_B)$ as the reference point. The DM considers the disadvantage of moving from $\overline{x}_B$ to $x_{B'}^{B,k}$ relative to the disadvantage from $\overline{x}_k$ to $\underline{x}_k$.

Although the two tradeoffs involve the same amount of changes on attribute x_k, they differ by sign, loss aversion implies that the disadvantage has a bigger effect than the advantage. As a

result, the advantage from $\underline{x}_B$ to $x_B^{B,k}$ in the gain-framed tradeoff is smaller than the disadvantage from $\overline{x}_B$ to $x_{B'}^{B,k}$ in the loss-framed tradeoff. In other words, the value difference produced by a gain (i.e., $v(x_B^{B,k}) - v(\underline{x}_B)$) could be smaller than the value difference from an equivalent loss (i.e., $v(\overline{x}_B) - v(x_{B'}^{B,k})$). Formally, this implies:

$$v(x_B^{B,k}) < 1 - v(x_{B'}^{B,k}) \tag{4.6}$$

This asymmetry leads to the following proposition.

Proposition 1. Let attribute weights w_B^{gain} and w_B^{loss} be derived from the gain- and loss-framed tradeoff procedures, respectively, using the most important attribute x_B. If

$$v(x_B^{B,k}) < 1 - v(x_{B'}^{B,k}) \quad \text{for all } k \neq B, \tag{4.7}$$

then the weight of x_B in the gain-framed tradeoff is strictly greater than in the loss-framed tradeoff:

$$w_B^{gain} > w_B^{loss}. \tag{4.8}$$

Proof: Given $v(x_B^{B,k}) < 1 - v(x_{B'}^{B,k})$, it follows that

$$\sum_{k \neq B} v(x_B^{B,k}) < \sum_{k \neq B} 1 - v(x_{B'}^{B,k}) \tag{4.9}$$

as

$$\sum_{k \neq B} 1 - v(x_{B'}^{B,k}) = (n-1) - \sum_{k \neq B} v(x_{B'}^{B,k}) \tag{4.10}$$

Therefore,

$$1 + \sum_{j \neq B} v(x_B^{B,j}) < n - \sum_{j \neq B} v(x_{B'}^{B,j}) \tag{4.11}$$

From equations (4.2) and (4.4), the weight of the most important attribute is given by $w_B^{gain} = \frac{1}{1+\sum_{j \neq B} v(x_B^{B,j})}$ and $w_B^{loss} = \frac{1}{n-\sum_{j \neq B} v(x_{B'}^{B,j})}$.

Inequality (4.11) shows that the denominator of w_B^{gain} is strictly smaller than the denominator of w_B^{loss}. Since both weights are defined as the reciprocal of these positive denominators, it follows directly that

$$w_B^{gain} > w_B^{loss}.$$

This completes the proof. □

Drawing from the above discussion, we propose the following hypothesis:

H1: The gain-framed tradeoff procedure will lead to a larger weight for the most important attribute compared to using the loss-framed tradeoff procedure.

4.4.2 Framing Effect on Attribute-Specific Value Functions

Framing an attribute as a gain or a loss can affect how DM perceives and evaluates it, despite the underlying information being logically equivalent (Tversky & Kahneman, 1981). While the term framing effect serves a descriptive role in behavioral decision research, it constitutes a bias within the normative framework of MAVT, as the preference elicitation is assumed to be invariant with equivalent representations. A gain frame emphasizes desirable outcomes and is associated with an increasing value function, whereas a loss frame emphasizes undesirable outcomes and is associated with a decreasing value function.

The value functions in prospect theory and MAVT differ in their formation. Prospect theory defines value over gains and losses relative to a psychological reference point, which can shift based on context or framing (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991). The value function is shown in Figure 4.2. The lowest level for gains and losses (point of zero) is interpreted as the reference point. Applying this to MAVT, the lowest attribute level can be regarded as the reference point. The attribute-specific value function in MAVT measures the DM's preference over attribute performance rather than the DM's preference over gains or losses. Despite this difference, prospect theory offers valuable insights: when an attribute is framed differently as a gain or a loss, the perceived reference point (the lowest attribute level) in MAVT may effectively shift. As a result, the shape of the elicited value function in MAVT can vary systematically across frames, reflecting the asymmetric valuation patterns predicted by prospect theory.

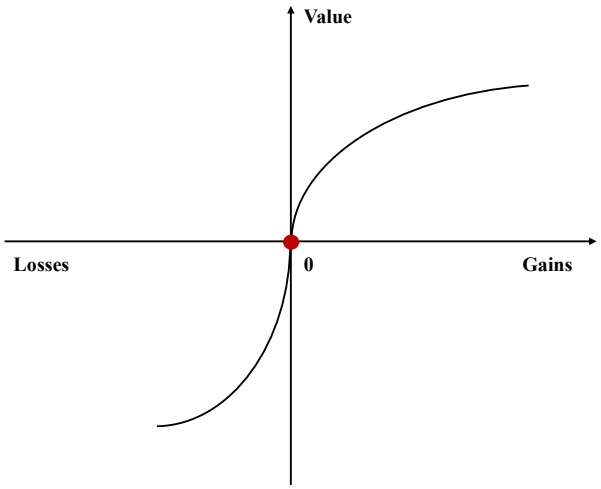


Figure 4.2: The value function in prospect theory

To enable a meaningful comparison between gain- and loss-framed value functions in MAVT, we transform the loss-framed attribute scale so that it increases in the same direction as the gain-framed scale (see Figure 4.3). Specifically, we reverse the x-axis of the loss-framed attribute so that both value functions represent increasing preference from the least to the most desirable outcome. This ensures that both value functions are defined over a common scale and direction, allowing for a direct comparison of their curvature and shape differences resulting from the framing manipulation.

When an attribute is framed in terms of gains, lower attribute levels (closer to the reference point) are perceived as small gains, and higher levels as larger gains. Due to diminishing sensi-

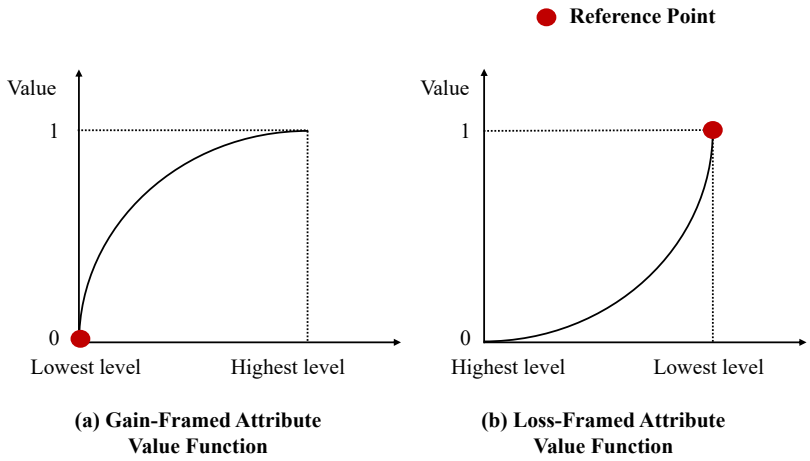


Figure 4.3: Elicited value functions for the same attribute in gain vs. loss framing

tivity in the gain domain from prospect theory, individuals assign relatively large marginal value to improvements near the lower end, but decreasing value to improvements near the upper end. This results in a concave increasing value function: steep initially and flattening as the attribute level increases.

Conversely, when the same attribute is framed as a loss and transformed to match the gain-framed direction. The reference point is still the lowest attribute level, but now it is the most preferred performance level. Due to diminishing sensitivity in the loss domain from prospect theory, individuals assign relatively large marginal value to changes near the reference point (the lowest attribute level), and smaller marginal value to changes near the end (the highest attribute level). This produces a convex increasing value function: initially flat and becoming steeper toward the end.

Some attributes may naturally elicit value functions that are linear, concave, or convex, depending on how preferences change across the attribute range. When such an attribute is framed in terms of gains or losses, the resulting value function may shift in degree or shape rather than category. For example, a concave function might become more steeply concave under a gain frame or less steep under a loss frame.

To detect and compare such differences, we use the area under the curve (AUC) as a measure that captures the overall shape of the value function. The AUC measures the total area under the curve relative to a reference axis. Unlike binary classifications (e.g., concave vs. convex), AUC can detect subtle but meaningful differences in curvature (Sun et al., 2025). For example, the AUC of an extremely concave function in the gain frame will be larger than the AUC of a moderately concave function in the loss frame, a difference that cannot be fully captured by categorical labels alone. Therefore, we hypothesize that:

H2: For the same attribute, the gain-framed value function will exhibit a larger area under the curve (AUC) than the loss-framed value function.

4.4.3 Framing Effect on Weight

As discussed earlier, this study not only examines individual biases in isolation, but also investigates how biases may interact with one another and across elicitation phases. Hypothesis 3 addresses such interactions and explores their impact on the derived weights, an aspect that, to our knowledge, has not been empirically examined in the gain-loss bias literature.

First, the framing effect can occur at multiple elicitation steps of MAVT and result in a complex influence on the derived weights. It influences both the elicitation of value functions and the tradeoff procedure. According to prospect theory, individuals are more sensitive to losses than to equivalent gains (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991). When an attribute is framed as a loss, this greater sensitivity leads to smaller adjustments in tradeoff tasks, reflecting reluctance to sacrifice on losses. At the same time, as shown in Hypothesis 2, framing also alters the curvature of the elicited value function. Since both the tradeoff judgments and the attribute-specific value functions jointly determine the attribute weights, the framing effect in both steps can interact and affect the derived weights.

Second, the two biases, the framing effect and the loss aversion, can also jointly affect the attribute weights. Hypothesis 1 established that loss aversion affects the weights differently depending on whether the tradeoff procedure is gain- or loss-framed. Hypothesis 2 showed that framing alters the shape of value functions. When combined, the asymmetry due to loss aversion interacts with the curvature differences induced by framing, producing an additional layer of divergence in the weights derived under the two tradeoff procedures.

The influences of the two types of interactions on the tradeoff procedure are illustrated in Figure 4.4. In the gain-framed tradeoff, the DM adjusts the most important attribute upward from its lowest level until indifference is reached. When the attribute itself is gain-framed, individuals tend to make larger adjustments from the lowest level compared to when the attribute is framed as losses (i.e., GA_$x_B^{B,k}$ > LA_$x_B^{B,k}$). As established in Hypothesis 2, gain-framed attributes are also associated with value functions with a larger AUC than loss-framed attributes. Together, these effects produce a higher value for $v(x_B^{B,k})$ under gain framing. According to equation (4.2), where $w_B^{gain} = \frac{1}{1+\sum_{j\neq B} v(x_B^{B,j})}$, this translates into a smaller weight for the most important attribute.

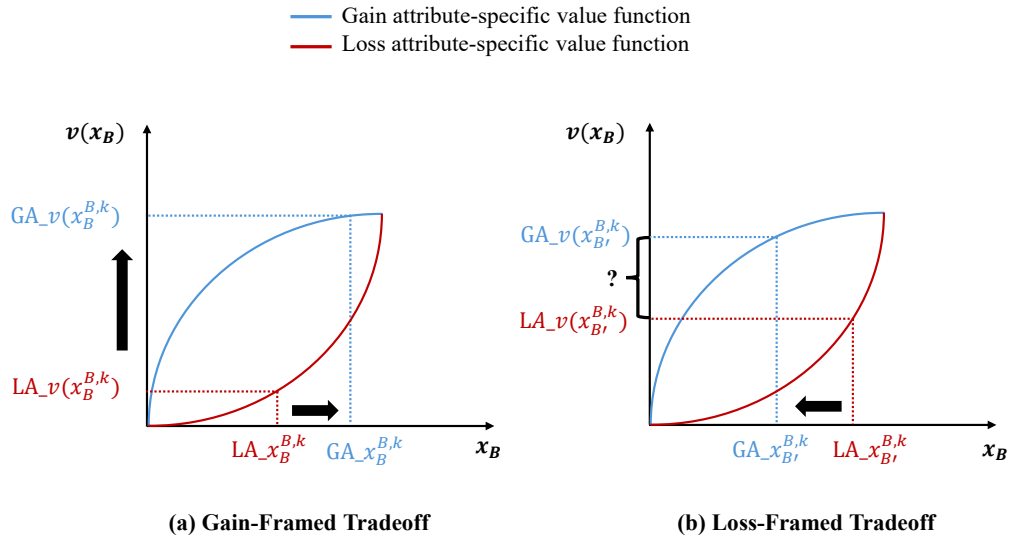
This relationship is formalized below:

Proposition 2. If, in a gain-framed tradeoff procedure, framing the attribute as a gain leads to a larger tradeoff score GA_$x_B^{B,k}$, and a value function $v(\cdot)$ with a larger AUC value compared to loss framing, then:

$$w_B^{gain} < w_B^{loss}. \tag{4.12}$$

Proof : Given a larger tradeoff score GA_$x_B^{B,j}$ and a value function with a larger AUC value under gain framing compared to loss framing, the resulting value GA_$v(x_B^{B,j})$ is higher compared to the loss framing. From equation (4.2), where $w_B^{gain} = \frac{1}{1+\sum_{j\neq B} v(x_B^{B,j})}$, a larger numerator in the denominator results in a smaller weight w_B^{gain}.

□



(a) Gain-Framed Tradeoff **(b) Loss-Framed Tradeoff**

Figure 4.4: Cross-phase and cross-bias interactions in the tradeoff procedure

In contrast, the effects are more nuanced in the loss-framed tradeoff, where the adjustment begins from the highest attribute level and moves downward. In this case, framing the attribute as a gain will lead to larger downward adjustments and thus a smaller tradeoff score (LA_$x_{B'}^{B,k} >$ GA_$x_{B'}^{B,k}$). Since the gain framing is associated with a value function with a larger AUC value, the smaller tradeoff score might still yield a similar tradeoff value $v(x_{B'}^{B,k})$. As a result, the overall weight computed from equation (4.4) remains comparable between the two framings.

Proposition 3. If, in a loss-framed tradeoff procedure, framing the attribute as a gain leads to a smaller tradeoff score $x_{B'}^{B,j}$ and a more concave value function $v(\cdot)$ compared to loss framing, then:

$$w_B^{gain} \approx w_B^{loss}. \tag{4.13}$$

Proof : The gain framing leads to a smaller tradeoff score GA_$x_{B'}^{B,j}$ and a value function with a higher AUC value. The loss framing leads to a larger tradeoff score GA_$v(x_{B'}^{B,j})$ and a value function with a lower AUC value. Because a higher AUC value function assigns a larger value to the same tradeoff score, the smaller tradeoff score under gain framing is offset by a higher value, while the larger score under loss framing is offset by a lower value. As a result, the tradeoff score values under the two framings are expected to be similar. Therefore, according to equation (4.4), where $w_B^{loss} = \frac{1}{n-\sum_{j\neq B} v(x_{B'}^{B,j})}$, there is no strong theoretical reason to expect the resulting weights to differ substantially between the two framings.

□

Based on these propositions, we hypothesize that:

H3a: In a gain-framed tradeoff procedure, framing the most important attribute in terms of gain (rather than loss) will lead to a smaller weight for that attribute.

H3b: In a loss-framed tradeoff procedure, framing the most important attribute in terms of gain (rather than loss) will not significantly affect the weight for that attribute.

4.5 Experiment Design

This study employed a structured questionnaire based on the MAVT procedure to test the three hypotheses related to gain-loss framing. The questionnaire was developed using Qualtrics and consisted of six parts: (i) informed consent, (ii) decision problem presentation, (iii) additivity assumption check, (iv) tradeoff-based weight elicitation, (v) value function elicitation, and (vi) demographic and control variables. This section outlines the content and purpose of each part and describes the overall experimental design.

In the first part, participants were informed of the experiment's purpose, procedures, potential risks, and benefits. They were then asked to voluntarily provide informed consent to participate in the study.

In the second part, participants were presented with a hypothetical decision problem involving the purchase of a reverse osmosis water filter. The alternatives were described as identical on all aspects, such as installation cost, maintenance, filtration speed, and energy use, except for two attributes: water output and mineral content. This problem allows for controlled manipulation of attribute framing, a key component of Hypotheses 2 and 3.

Each attribute was framed either in gain or loss terms, with objectively equivalent descriptions. For the water attribute (Table 4.3), participants were presented with either the clean water recovery rate (gain frame) or the wastewater generation rate (loss frame). For example, a recovery of 3 liters of clean water corresponds to 7 liters of wastewater per 10 liters of raw water. The mineral attribute (Table 4.4) was similarly framed as either the retention (gain) or removal (loss) rate of essential minerals such as calcium and magnesium. By manipulating these frames, we test whether DMs respond differently to objectively equivalent but differently framed attribute descriptions.

Table 4.3: Water attribute in gain-loss frames

Attribute	Frame	Description	Unit	Range
Clean Water Recovery Rate	Gain	The amount of clean water produced from 10 liters of raw water. A higher value means greater efficiency.	liters	[3, 8]
Wastewater Generation Rate	Loss	The amount of wastewater discarded from 10 liters of raw water. A lower value means greater efficiency.	liters	[2, 7]

Note. The two framings are objectively equivalent because clean water recovery and wastewater generation sum to 10 liters (e.g., 3+7 and 8+2).

To investigate the framing effect (Hypotheses 2 and 3), a between-subjects design is adopted. Participants were randomly assigned to one of four framing conditions based on the combination of gain/loss descriptions for the water and mineral attributes (gain-gain, gain-loss, loss-gain, loss-loss). A between-subject design is widely used in framing effect experiments because exposure to multiple forms within the same participant can trigger cross-frame comparisons and affect the investigation of the framing effect (Tversky & Kahneman, 1981; Levin et al., 1998). Random assignment to conditions allows for clean identification of treatment effects while reducing carryover and learning effects that often remain challenges in within-subjects designs (Charness et al., 2012; Greenwald, 1976).

Table 4.4: Mineral attribute in gain-loss frames

Attribute	Frame	Description	Unit	Range
Essential Mineral Retention Rate	Gain	The percentage of beneficial minerals (e.g., calcium, magnesium) remaining in the filtered water. A higher value indicates better water quality.	%	[50, 95]
Essential Mineral Removal Rate	Loss	The percentage of beneficial minerals (e.g., calcium, magnesium) removed from the water during filtration. A lower value indicates better water quality.	%	[5, 50]

Note. The two framings are objectively equivalent because mineral retention and mineral removal sum to 100% (e.g., 50+50 and 95+5).

In the third part, the additivity assumptions were verified by checking if mutual preference independence and difference independence are satisfied. It can be done using a set of tradeoff comparisons, see (Keeney & Raiffa, 1976, pp. 118–119) for an example.

In the fourth part, attribute weights were elicited using the tradeoff procedure. To test Hypothesis 1 (gain-loss bias in the tradeoff procedure), we employed a within-subjects design, where all participants completed both gain and loss tradeoff tasks. Since the decision problem involves only two attributes, each tradeoff procedure required participants to evaluate a single indifference pair. A within-subjects design is particularly suitable in this context because it allows us to directly compare how the same individual's preferences change under the two framing conditions, while controlling for between-subject variability (Keppel, 1991). Moreover, the low task complexity ensures that completing both tradeoff tasks does not impose excessive cognitive demands on participants (Charness et al., 2012; Greenwald, 1976).

In addition, combining the within-subject design for weight elicitation with the between-subjects design for attribute framing allows us to compare the weights produced under homogeneous (gain–gain, loss–loss) and heterogeneous (gain–loss, loss–gain) framings. This comparison provides insights into how biases interact across elicitation steps and their potential for bias mitigation.

In the fifth part, the mid-value splitting procedure was conducted to elicit the two attribute-specific value functions. Take eliciting the attribute-specific value function for the clean water recovery rate in the gain-gain group as an example. To obtain the first midvalue point, we ask, "*If the clean water recovery rate of a water filter decreases as the essential mineral retention rate increases, consider the following scenarios: 1. The mineral retention rate increases from 50% to* M_1*%. 2. The mineral retention rate increases from* M_1*% to 95%. Assume the decrease in clean water recovery rate is the same in both scenarios. What value of* M_1 *would make you feel indifferent between the two increases in the mineral retention rate?*" M_1 is then identified as the first midvalue point. The subsequent mid-value points can be obtained following similar questions, and the attribute-specific value function can be plotted using the midvalue points.

In the final part, participants provided demographic information and responses to control variables, including concern about water quality, prioritization of water-related criteria, and whether they currently use a water filter at home. These variables were collected to characterize the sample and assess potential covariates in the analysis. Collecting such information can help identify factors that might confound the gain-loss bias and, by controlling for them, improve

the precision in investigating the gain-loss bias (Carneiro et al., 2020; Shadish et al., 2002).

Participants were recruited through the Prolific online platform, which offers pre-screening options and response verification tools to enhance data quality. A total of 283 individuals from European countries participated. Since the questionnaire was in English, we limited participation to individuals fluent in English using the pre-screening functions. Moreover, the response verification function enables us to reject incomplete answers or those that provided answers outside of the value ranges. The detailed demographics of the sample are presented in Table 4.5.

Table 4.5: Demographic characteristics of participants ($n = 283$)

Characteristics	Levels	Percent
Gender	Female	43.8%
	Male	54.1%
	Other	2.1%
Age	[18,24]	29.3%
	[25,34]	37.1%
	[35,44]	29%
	> 44	10.9%
Education	High School	29.7%
	Bachelor's degree	42%
	Master's degree	17.3%
	Other	11%

4.6 Results and Discussion

This section presents the results related to the main hypotheses and their implications. In addition to the primary analyses, follow-up ANCOVA and post hoc tests were conducted to assess the potential influence of control variables. The results indicated that none of the control variables had a significant effect on the outcomes across all hypotheses (all $p > 0.05$).

4.6.1 The Loss Aversion within the Tradeoff Procedure

The Main Effect

Hypothesis 1 posited that using the gain-framed tradeoff procedure would result in a higher weight for the more important attribute than using the loss-framed tradeoff procedure. To test this, a paired sample *t*-test was conducted to compare the weights of the more important attribute between the two tradeoff groups, regardless of whether the attribute is water or mineral. The results, as shown in Table 4.6, show that the weight was significantly higher in the gain-framed tradeoff condition ($Mean = 0.65$, $SD = 0.12$) than in the loss-framed tradeoff condition ($Mean = 0.63$, $SD = 0.12$), with $t(df) = 1.84$, $p = 0.033$ (one-tailed). This finding supports the hypothesis and suggests that loss aversion can systematically influence the weights elicited in the tradeoff procedure.

Table 4.6: Paired samples t-test of the most important attribute weight under gain- and loss-framed tradeoffs

Comparison	Mean Difference	Std. Deviation	*t*	One-Sided *p*	Maximum Difference
Gain-Tradeoff vs. Loss-Tradeoff	0.0149	0.1299	1.844	0.033	0.4304

This result aligns with prior findings of loss aversion in the tradeoff tasks, where participants tend to assign significantly larger gains to compensate for the losses (Delquié, 1993; Bleichrodt & Pinto, 2002). These early studies primarily examined this effect in the tradeoff scores, our results extend this line of research by showing that the bias also affects derived weights. From a prescriptive perspective, this finding raises concerns about the use of the tradeoff procedure in multi-attribute decision-making and highlights the need for mitigation strategies in practice.

The Mitigation

First, since two tradeoff procedures lead to different weights, one mitigation strategy is to ask DMs to conduct both procedures and take the averages of the derived weights. Exposing to different conditions and averaging has been recognized as a simple and effective bias mitigation strategy in prescriptive decision analysis, as it balances out distortions that might arise under different elicitation methods (Montibeller & von Winterfeldt, 2015; Almashat et al., 2008).

Besides, to explore the interaction between loss aversion and framing effect and to see whether they can cancel each other out, four experimental groups were examined: gain-gain, gain-loss, loss-gain, and loss-loss. The gain-gain and loss-loss groups (both attributes framed in the same direction) were categorized as homogeneous framing conditions, while the gain-loss and loss-gain groups (attributes framed in opposite directions) were classified as heterogeneous framing conditions.

We compared the weights of the more important attribute across gain and loss tradeoff procedures within each framing type. As shown in Table 4.7, the gain-framed tradeoff yielded significantly higher weights than the loss-framed tradeoff under heterogeneous framing ($p = 0.040$), while no significant difference was observed under homogeneous framing ($p = 0.193$). This pattern suggests that framing consistency across attributes plays a critical role in the elicited weights. When all attributes are framed in the same direction, loss aversion affects all attributes equally in the same direction, but they will cancel each other out during the normalization of weights. In contrast, when attributes are framed heterogeneously, loss aversion can disproportionately distort the weights of minority-framed attributes than the majority-framed attributes, as weights redistribute to satisfy the constraint that the sum of weights is 1.

Table 4.7: Paired samples t-test for attribute weight under different framing conditions

Comparison	Gain Tradeoff Mean	Loss Tradeoff Mean	*t*	One-Sided *p*
Heterogeneous Frame	0.66	0.64	1.769	0.040
Homogeneous Frame	0.64	0.63	0.869	0.193

This finding shows that the interaction of biases can sometimes counterbalance each other, consistent with prior work demonstrating that certain cognitive biases can cancel each other

when elicitation processes are planned (Lahtinen & Hämäläinen, 2016). Practically, this result also underscores the importance of framing consistency in applying decision analysis methods.

The Effect on Ranking

To assess the practical implications of loss aversion, we tested whether it affects the final decision outcomes in MAVT. Two analyses were conducted.

First, we examined whether the selection of the best alternative, based on the overall value function in MAVT, differed between the gain- and loss-framed tradeoff procedures. A Wilcoxon signed-rank test indicated a statistically significant difference in the selection of the best alternative, $z = -2.669, p = 0.008$. This suggests that the effect on weights can be carried to affect the final decision outcome.

Second, we assessed the similarity of full alternative rankings generated under gain and loss frames using Kendall's Tau-b. Kendall's τ_b is a nonparametric correlation coefficient that measures the ordinal association between two rankings, correcting for ties (Kendall, 1938). It is defined as $\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$, where n_c and n_d denote the number of concordant and discordant pairs, respectively, and the denominator adjusts for ties in both rankings. This formulation ensures that τ_b ranges from -1 (perfect disagreement) to +1 (perfect agreement). Because it explicitly handles ties, Kendall's Tau-b is widely applied in decision analysis to evaluate the similarity of rank orders of alternatives (Shekhovtsov, 2021; Sałabun & Urbaniak, 2020).

The mean correlation between rankings ($Mean = 0.77, SD = 0.26$) shows that the rankings produced from gain- and loss-framed tradeoff procedures are far from identical. As shown in Table 5.4, this similarity was significantly lower than perfect agreement ($\tau_b = 1$). To further assess the extent of deviation, we iteratively tested descending benchmark values. The results showed that $\tau_b = 0.81$ was the lowest threshold at which the difference remained statistically significant. These results indicate that while the rankings from gain- and loss-framed tradeoffs are broadly aligned, systematic deviations persist due to the gain-loss bias, and these deviations are strong enough to influence the final decisions.

Table 4.8: Deviation of ranking similarity (Kendall's τ_b) from benchmarks

Kendall's Tau-b	**Mean Difference**	***t***	**One-Sided *p***
Gain - Loss Tradeoff (Test Value= 1)	-0.22260	-14.467	< 0.001
Gain - Loss Tradeoff (Test Value= 0.81)	-0.03260	-2.119	0.017

4.6.2 The Framing Effect on Attribute-Specific Value Function

The Main Effect

Hypothesis 2 posits that the shape of the attribute-specific value function is influenced by how the attribute is framed, as a gain or a loss. To test this, we used the AUC as a summary measure of the elicited value functions.

We first tested whether the AUC for gain-framed attributes is different from the loss-framed ones, regardless of whether it is water or mineral. The independent-samples *t*-test result shows that gain-framed AUC ($Mean = 0.50$, $SD = 0.08$) is significantly larger than the loss-framed AUC ($Mean = 0.43$, $SD = 0.09$), with $p < 0.001$, supporting the hypothesis. This average AUC value indicates that a gain-framed value function is approximately linear in shape, while a loss-framed value function is convex. Although the gain side does not display the concavity predicted by the prospect theory (Kahneman & Tversky, 1979), the loss side deviates strongly from linearity and shows convexity is consistent with the prospect theory's central prediction of asymmetric sensitivity between the gain and loss domains.

To examine whether this framing effect is consistent across attributes, independent-samples *t*-tests were also conducted for the two attributes, water and mineral. As shown in Table 4.9, the results indicate that this pattern held for both water and mineral; the AUC is significantly larger under gain framing than under loss framing ($p < 0.001$ for both), suggesting a robust framing effect at the attribute level.

Table 4.9: Independent samples t-test: AUC comparison between gain and loss frames

Attribute	**AUC Gain Mean**	**AUC Loss Mean**	***t***	**One-Sided *p***
Water	0.50	0.44	5.663	< 0.001
Mineral	0.50	0.42	7.413	< 0.001

To further assess whether the framing effect on value functions is independent of the framing of the other attribute, we conducted post hoc comparisons across all framing combinations (gain–gain, gain–loss, loss–gain, and loss–loss). For each attribute, we compared conditions where its framing was manipulated, while controlling for the framing of the other attribute.

As shown in Table 4.10, the gain frame consistently led to a higher AUC than the loss frame, regardless of the framing of the other attribute. This pattern held for both water and mineral, suggesting that the framing effect operates at the attribute level and is not moderated by the framing of the other attribute in the decision problem.

Table 4.10: Multiple comparisons of AUC across framing combinations

Comparisons (a vs. b)	**AUC Mean Difference (a-b)**	**Sig.**
Gain–gain vs. **Loss**–gain (water)	0.0901	< 0.001
Gain–gain vs. **Loss**–loss (water)	0.0672	< 0.001
Gain–loss vs. **Loss**–gain (water)	0.0496	< 0.001
Gain–loss vs. **Loss**–loss (water)	0.0266	0.05
Gain–**gain** vs. Gain–**loss** (mineral)	0.0837	< 0.001
Gain–**gain** vs. Loss–**loss** (mineral)	0.0824	< 0.001
Loss–**gain** vs. Gain–**loss** (mineral)	0.0765	< 0.001
Loss–**gain** vs. Loss–**loss** (mineral)	0.0752	< 0.001

In summary, these results collectively provide strong evidence that attribute-specific value functions are systematically affected by framing. We find that gain-framed attributes' value functions are associated with a higher AUC value than loss-framed ones, meaning that across the entire attribute range, the values assigned under gain framing are systematically higher than

those under loss framing. This pattern is consistent with the attribute-framing literature, which shows that people give more positive evaluations when attributes are described in gain rather than loss terms (Levin et al., 1998; Kim et al., 2014). Moreover, investigating attribute framing in the multi-attribute decision-making context enables a more comprehensive analysis. Our results show that this effect is consistent across both attributes and is not contingent on the framing of the other attribute in the decision context.

The Mitigation

To mitigate the framing effect on the value function, we propose two strategies. The first strategy addresses the issue at the individual level. DMs are asked to evaluate each attribute under both gain and loss frames. The final value function is obtained by averaging the two resulting functions. This approach is commonly used in the bias mitigation literature and helps neutralize systematic framing-induced biases by combining perspectives from different frames (Montibeller & von Winterfeldt, 2015; Almashat et al., 2008; Hammond et al., 1998). It has also been discussed previously to address loss aversion in the tradeoff procedure. While we do not directly test this bias mitigation method in the current study, the finding that different groups produce statistically different value functions under gain and loss frames suggests that averaging the two functions within individuals can help neutralize the framing effect.

The second bias mitigation strategy involves group decision-making, which is also commonly used as an effective bias mitigation strategy for cognitive biases (Montibeller & von Winterfeldt, 2015; Larrick, 2004). Here, the same decision problem and attributes are formulated in both gain and loss frames, and different group members are randomly assigned to one frame. Each DM is exposed to only one frame throughout the process, which reduces task complexity and cognitive load. Their elicited value functions are then aggregated to produce a group-level value function. Group decision-making can take various forms, including consensus building, comparative discussion, or aggregation (Huang et al., 2013; Chiclana et al., 2013; Altuzarra et al., 2010; Hirokawa, 1990). In this study, we focus on the last one. Specifically, we constructed 100 artificial groups of sizes 10, 5, 3, and 2 by randomly sampling participants from the full sample. For each group, we computed the average value function across members.

To evaluate whether this group aggregation mitigates the framing effect, we assessed whether the group-averaged value functions lie between the extreme value functions derived under pure gain and loss framings. Although real-world group formation is seldom random, the use of simulated random groups allows us to systematically test the effectiveness of the aggregation strategy under controlled conditions.

To test the statistical significance of these differences, we employed ANOVA post hoc bootstrap comparisons with 1000 resamples and bias-corrected and accelerated (BCa) 95% confidence intervals (CI). The results consistently show that the group-averaged AUC values were significantly lower than (the 95% CIs for all mean differences were entirely below zero) the gain-only framing conditions, and significantly higher than (the 95% CIs for all mean differences were entirely above zero) the loss-only framing conditions. This result held regardless of group size, suggesting that aggregating across individuals with varied frames can effectively reduce the framing-induced bias in attribute-specific value functions. Table 4.11 presents the results for the groups of sizes 2 and 3. Because groups were formed by randomly sampling from the full dataset, the chance of drawing mixed-frame groups increases for larger groups, this

greater heterogeneity ensures that the averaged value function lies between the two extremes, thereby reducing framing-induced bias. We tested groups of 5 and 10 and found consistent results, they are not presented here to avoid redundancy.

Table 4.11: Multiple comparisons for group decision-making (Bootstrap, BCa 95% CI)

Comparisons	Mean Difference	Bias	BCa 95% CI
Group of 2 vs. gain framing (water)	-0.03648	0.00021	[-0.05592, -0.01939]
Group of 2 vs. loss framing (water)	0.02031	0.00015	[0.00360, 0.03843]
Group of 2 vs. gain framing (mineral)	-0.03193	-0.00068	[-0.05037, -0.01450]
Group of 2 vs. loss framing (mineral)	0.04790	0.00014	[0.02675,0.06924]
Group of 3 vs. gain framing (water)	-0.02290	0.00013	[-0.04045, -0.00455]
Group of 3 vs. loss framing (water)	0.03390	-0.00010	[0.01772,0.05091]
Group of 3 vs. gain framing (mineral)	-0.05103	0.000003	[-0.06908,-0.03296]
Group of 3 vs. loss framing (mineral)	0.02880	0.000003	[0.01015,0.04743]

Note. The column "Bias" refers to the bootstrap estimate of bias (i.e., the difference between the bootstrap mean and the observed sample mean).

4.6.3 The Framing Effect on Weight

Hypothesis 3 investigates whether framing an attribute in terms of gains versus losses influences the weights derived from the tradeoff procedure, particularly when the attribute is the most important one. We further hypothesize that this framing effect occurs only in the gain-framed tradeoff procedure, not in the loss-framed tradeoff procedure. This is because we theoretically expect that, in the gain-framed tradeoff, both the change in tradeoff scores (due to loss aversion) and the change in the value functions (due to the framing effect) that evaluate these scores act in the same direction, amplifying the distortion; whereas in the loss-framed tradeoff, they are in opposite directions, thereby offsetting each other. This hypothesis captures two levels of interaction: the interaction between two biases, framing effect and loss aversion, and the interaction between two key steps in MAVT, the value function and weight elicitation steps.

To test hypothesis 3, we performed independent samples *t*-tests comparing the weights of the most important attribute under gain and loss framing conditions, regardless of whether the attribute was water or mineral. As shown in Table 4.12, the results support our hypothesis: within the gain tradeoff procedure, the weight for the most important attribute is significantly lower when framed as a gain compared to a loss ($p = 0.044$). In contrast, no significant difference in weights emerged between framing conditions in the loss tradeoff procedure ($p = 0.390$).

Table 4.12: Independent samples t-test for framing effect on weights

Comparisons	Gain Frame Mean (SD)	Loss Frame Mean (SD)	*t*	One-Sided *p*
Gain Tradeoff Weight	0.64 (0.11)	0.66 (0.13)	-1.716	0.044
Loss Tradeoff Weight	0.64 (0.12)	0.63 (0.13)	0.279	0.390

These findings, integrated with the framing effect on attribute-specific value functions discussed in Section 4.6.2, are consistent with theoretical expectations and underscore the pervasive influence of framing throughout the MAVT process. They suggest that framing influences

both components of the tradeoff procedure, the elicited tradeoff scores and the attribute-specific value functions used to translate the scores. When these two components are influenced in the same direction in the gain-framed tradeoffs, the bias can lead to systematic distortions in the final weights. The attribute framing literature has primarily examined single-attribute decision making (Levin & Gaeth, 1988; Levin et al., 1998), and thus focused on attribute evaluation, rather than on the relative weight of attributes. Kim et al. (2014) examined the framing effect in an unaided multi-attribute setting and found that attribute-framing does not significantly affect the attribute weights. Yet their study relied on direct judgments of attributes and alternatives, rather than applying a formal multi-attribute decision-making method, and therefore did not capture how framing might influence weights derived from structured elicitation procedures. Our study extends this literature by being, to our knowledge, the first to demonstrate how attribute framing influences attribute weights in a MAVT setting.

To mitigate the framing effect in weight elicitation, one effective approach is to use the loss-framed tradeoff procedure, as the two framing-induced distortions, on the value function and on the tradeoff elicitation, tend to operate in opposite directions, thus the effects are canceled out. Alternatively, weights can be elicited under both gain and loss frames and then averaged, a bias mitigation technique often recommended in behavioral decision analysis (Hammond et al., 1998; Almashat et al., 2008; Montibeller & von Winterfeldt, 2015).

4.7 Conclusion

The influence of gain-loss bias in decisions with multiple objectives is more complex and less understood than in single-objective contexts. This study investigated gain-loss bias within a formal multi-attribute decision-making method, aiming to understand how and where biases enter the decision-making process and to develop practical bias mitigation strategies.

We tested three hypotheses addressing different stages of the decision process. First, results showed that loss aversion influences the tradeoff procedure: the weight of the most important attribute was higher under gain-framed tradeoffs than loss-framed tradeoffs. This bias can be mitigated by conducting tradeoffs under both frames and averaging the results or by using a homogeneous frame. Second, the study found that attribute framing affects the shape of the elicited value function: attributes framed as gains produce value functions with a larger AUC than when framed as losses. This framing effect can be reduced by eliciting value functions under both frames and averaging them, or by using group-based decision-making where different DMs evaluate attributes under different frames. Third, the interaction between biases and elicitation steps influenced the weights assigned to the attributes. Specifically, when the most important attribute is framed as a gain, it receives a smaller weight compared to when framed as a loss. This effect was only significant when tradeoffs were framed as gains, suggesting that loss-framed tradeoffs can reduce this bias. Bias mitigation strategies include using loss-framed tradeoffs or averaging weights across different framing conditions. In summary, to mitigate framing effects throughout the MAVT process, applying both gain and loss frames for the attributes and tradeoff procedures and averaging the outputs is recommended. This approach is also broadly used in the bias mitigation literature (Montibeller & von Winterfeldt, 2015; Almashat et al., 2008; Hammond et al., 1998; Larrick, 2004).

The findings advance the understanding of gain-loss bias in decisions with multiple objec-

tives by showing that biases can emerge from different sources and interact across elicitation steps. The results confirm that gain-loss framing systematically distorts preference elicitation, consistent with earlier findings on attribute framing and loss aversion (Kim et al., 2014; Delquié, 1993; Bleichrodt & Pinto, 2002). It also extends these insights to attribute-level evaluation (e.g., value functions and weights), rather than general evaluation of attributes and alternatives. Moreover, the proposed mitigation strategies, such as using both frames and averaging outputs, and in group settings (Almashat et al., 2008; Yaniv, 2011), offer actionable guidance for analysts seeking to reduce bias in real-world decisions. These mitigation strategies can be understood within the choice architecture literature (Soll et al., 2015; Thaler & Sunstein, 2021), as they operate by redesigning the elicitation and aggregation procedure, rather than targeting the decision-maker's cognitive processes directly, to reduce the influence of framing on preferences. This highlights the potential of integrating bias mitigation mechanisms into the procedural design of decision analysis methods. These contributions highlight the complex nature of decisions with multiple objectives and emphasize the need to jointly consider methodological procedures and cognitive mechanisms when examining biases in multi-attribute decision-making contexts.

Several limitations should be acknowledged. First, the experimental design included only two attributes, which limits the generalizability of the findings to more complex decision problems. The hypotheses regarding weight elicitation primarily focused on the most important attribute, and the effect on the second attribute was relatively straightforward to interpret. However, real-world decisions often involve more than two attributes, where interactions among attributes and potential tradeoff inconsistencies may be more complex. Future research should explore how gain-loss bias affects the weights of all attributes when there are more than two attributes involved. Second, the proposed bias mitigation strategies, such as averaging frames or using group-based responses, were tested using the same sample of participants rather than independent validation samples. As such, the efficacy of these strategies might be partly influenced by sample-specific characteristics or learning effects. Additionally, the group decision-making scenarios used in this study were artificially constructed by aggregating individual responses, which differs from natural group deliberation processes involving communication, negotiation, and influence dynamics (Huang et al., 2013; Chiclana et al., 2013; Altuzarra et al., 2010; Hirokawa, 1990). Future studies should test these strategies in more ecologically valid group settings to assess their robustness in practice.

Bibliography

Allen, E. J., P. M. Dechow, D. G. Pope, G. Wu (2017) Reference-dependent preferences: Evidence from marathon runners, *Management Science*, 63(6), pp. 1657–1672.

Almashat, S., B. Ayotte, B. Edelstein, J. Margrett (2008) Framing effect debiasing in medical decision making, *Patient Education and Counseling*, 71(1), pp. 102–107.

Altuzarra, A., J. M. Moreno-Jiménez, M. Salvador (2010) Consensus building in ahp-group decision making: A bayesian approach, *Operations Research*, 58(6), pp. 1755–1773.

Barberis, N. C. (2013) Thirty years of prospect theory in economics: A review and assessment, *Journal of Economic Perspectives*, 27(1), pp. 173–196.

Barnes, O. G., S. Hess, T. O. Hancock (2025) Revisiting framing effects: integrating multiple valence frames in choice modeling, preprint.

Beinat, E. (1997) *Value functions for environmental management*, Springer, Dordrecht.

Benartzi, S., R. H. Thaler (1995) Myopic loss aversion and the equity premium puzzle, *The Quarterly Journal of Economics*, 110(1), pp. 73–92.

Beratšová, A., K. Krchová, N. Gažová, M. Jirásek (2016) Framing and bias: A literature review of recent findings, *Central European Journal of Management*, 3(2), pp. 23–32.

Bibby, P. A., E. Ferguson (2011) The ability to process emotional information predicts loss aversion, *Personality and Individual Differences*, 51(3), pp. 263–266.

Bier, V. M., B. L. Connell (1994) Ambiguity seeking in multi-attribute decisions: Effects of optimism and message framing, *Journal of Behavioral Decision Making*, 7(3), pp. 169–182.

Bleichrodt, H., J. L. Pinto (2002) Loss aversion and scale compatibility in two-attribute trade-offs, *Journal of Mathematical Psychology*, 46(3), pp. 315–337.

Carmon, Z., K. Wertenbroch, M. Zeelenberg (2003) Option attachment: When deliberating makes choosing feel like losing, *Journal of Consumer Research*, 30(1), pp. 15–29.

Carneiro, P., S. Lee, D. Wilhelm (2020) Optimal data collection for randomized control trials, *The Econometrics Journal*, 23(1), pp. 1–31.

Cassotti, M., M. Habib, N. Poirel, A. Aïte, O. Houdé, S. Moutier (2012) Positive emotional context eliminates the framing effect in decision-making, *Emotion*, 12(5), pp. 926–931.

Charness, G., U. Gneezy, M. A. Kuhn (2012) Experimental methods: Between-subject and within-subject design, *Journal of Economic Behavior & Organization*, 81(1), pp. 1–8.

Cheng, F.-F., C.-S. Wu (2010) Debiasing the framing effect: The effect of warning and involvement, *Decision Support Systems*, 49(3), pp. 328–334.

Chiclana, F., J. T. García, M. J. del Moral, E. Herrera-Viedma (2013) A statistical comparative study of different similarity measures of consensus in group decision making, *Information Sciences*, 221, pp. 110–123.

Delquié, P. (1993) Inconsistent trade-offs between attributes: New evidence in preference assessment biases, *Management Science*, 39(11), pp. 1382–1395.

Delquié, P. (1997) "bi-matching": A new preference assessment method to reduce compatibility effects, *Management Science*, 43(5), pp. 640–658.

Dyer, J. S., R. K. Sarin (1979) Measurable multiattribute value functions, *Operations Research*, 27(4), pp. 810–822.

Fishburn, P. C. (1967) Methods of estimating additive utilities, *Management Science*, 13(7), pp. 435–453.

Fu, L., J. Yu, S. Ni, H. Li (2018) Reduced framing effect: Experience adjusts affective forecasting with losses, *Journal of Experimental Social Psychology*, 76, pp. 231–238.

Gächter, S., E. J. Johnson, A. Herrmann (2022) Individual-level loss aversion in riskless and risky choices, *Theory and Decision*, 92(3), pp. 599–624.

Ganzach, Y., N. Karsahi (1995) Message framing and buying behavior: A field experiment, *Journal of Business Research*, 32(1), pp. 11–17.

Garcia-Retamero, R., M. K. Dhami (2013) On avoiding framing effects in experienced decision makers, *Quarterly Journal of Experimental Psychology*, 66(4), pp. 829–842.

Genesove, D., C. Mayer (2001) Loss aversion and seller behavior: Evidence from the housing market, *The Quarterly Journal of Economics*, 116(4), pp. 1233–1260.

Gong, J., Y. Zhang, Z. Yang, Y. Huang, J. Feng, W. Zhang (2013) The framing effect in medical decision-making: a review of the literature, *Psychology, Health & Medicine*, 18(6), pp. 645–653.

Greenwald, A. G. (1976) Within-subjects designs: To use or not to use?, *Psychological Bulletin*, 83(2), pp. 314–320.

Hammond, J. S., R. L. Keeney, H. Raiffa (1998) The hidden traps in decision making, *Harvard Business Review*, 76(5), pp. 47–58.

Hardie, B. G., E. J. Johnson, P. S. Fader (1993) Modeling loss aversion and reference dependence effects on brand choice, *Marketing Science*, 12(4), pp. 378–394.

Heath, C., R. P. Larrick, G. Wu (1999) Goals as reference points, *Cognitive Psychology*, 38(1), pp. 79–109.

Hirokawa, R. Y. (1990) The role of communication in group decision-making efficacy: A task-contingency perspective, *Small Group Research*, 21(2), pp. 190–204.

Huang, Y.-S., W.-C. Chang, W.-H. Li, Z.-L. Lin (2013) Aggregation of utility-based individual preferences for group decision-making, *European Journal of Operational Research*, 229(2), pp. 462–469.

Kahneman, D., J. L. Knetsch, R. H. Thaler (1991) Anomalies: The endowment effect, loss aversion, and status quo bias, *Journal of Economic Perspectives*, 5(1), pp. 193–206.

Kahneman, D., A. Tversky (1979) Prospect theory: An analysis of decision under risk, *Econometrica*, 47(2), pp. 263–292.

Keeney, R. L. (1974) Multiplicative utility functions, *Operations Research*, 22(1), pp. 22–34.

Keeney, R. L. (2009) *Value-focused thinking: A path to creative decisionmaking*, Harvard University Press, Cambridge.

Keeney, R. L., H. Raiffa (1976) *Decisions with multiple objectives: Preferences and value trade-offs*, Cambridge University Press, Cambridge.

Kendall, M. G. (1938) A new measure of rank correlation, *Biometrika*, 30(1-2), pp. 81–93.

Keppel, G. (1991) *Design and analysis: A researcher's handbook*, Pearson Prentice Hall, New Jersey.

Kim, J., J.-E. Kim, R. Marshall (2014) Search for the underlying mechanism of framing effects in multi-alternative and multi-attribute decision situations, *Journal of Business Research*, 67(3), pp. 378–385.

Klapper, D., C. Ebling, J. Temme (2005) Another look at loss aversion in brand choice data: can we characterize the loss averse consumer?, *International Journal of Research in Marketing*, 22(3), pp. 239–254.

Knetsch, J. L. (1989) The endowment effect and evidence of nonreversible indifference curves, *The American Economic Review*, 79(5), pp. 1277–1284.

Lahtinen, T. J., R. P. Hämäläinen (2016) Path dependence and biases in the even swaps decision analysis method, *European Journal of Operational Research*, 249(3), pp. 890–898.

Larrick, R. P. (2004) Debiasing, in: *Blackwell Handbook of Judgment and Decision Making*, Blackwell Publishing, Malden, pp. 316–338.

Levin, I. P., G. J. Gaeth (1988) How consumers are affected by the framing of attribute information before and after consuming the product, *Journal of Consumer Research*, 15(3), pp. 374–378.

Levin, I. P., G. J. Gaeth, J. Schreiber, M. Lauriola (2002) A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of effects, *Organizational Behavior and Human Decision Processes*, 88(1), pp. 411–429.

Levin, I. P., S. L. Schneider, G. J. Gaeth (1998) All frames are not created equal: A typology and critical analysis of framing effects, *Organizational Behavior and Human Decision Processes*, 76(2), pp. 149–188.

Levy, J. S. (1992) An introduction to prospect theory, *Political Psychology*, 13(2), pp. 171–186.

McNeil, B. J., S. G. Pauker, H. C. Sox, A. Tversky (1982) On the elicitation of preferences for alternative therapies, *New England Journal of Medicine*, 306(21), pp. 1259–1262.

Meyerowitz, B. E., S. Chaiken (1987) The effect of message framing on breast self-examination attitudes, intentions, and behavior, *Journal of Personality and Social Psychology*, 52(3), pp. 500–510.

Miu, A. C., L. G. Crişan (2011) Cognitive reappraisal reduces the susceptibility to the framing effect in economic decision making, *Personality and Individual Differences*, 51(4), pp. 478–482.

Montibeller, G., D. von Winterfeldt (2015) Cognitive and motivational biases in decision and risk analysis, *Risk Analysis*, 35(7), pp. 1230–1251.

Mrkva, K., E. J. Johnson, S. Gächter, A. Herrmann (2020) Moderating loss aversion: Loss aversion has moderators, but reports of its death are greatly exaggerated, *Journal of Consumer Psychology*, 30(3), pp. 407–428.

Payne, J. W., N. Sagara, S. B. Shu, K. C. Appelt, E. J. Johnson (2013) Life expectancy as a constructed belief: Evidence of a live-to or die-by framing effect, *Journal of Risk and Uncertainty*, 46(1), pp. 27–50.

Sałabun, W., K. Urbaniak (2020) A new coefficient of rankings similarity in decision-making problems, in: *International Conference on Computational Science*, Springer, pp. 632–645.

Shadish, W., T. D. Cook, D. T. Campbell (2002) *Experimental and quasi-experimental designs for generalized causal inference*, Houghton Mifflin, Boston.

Shekhovtsov, A. (2021) How strongly do rank similarity coefficients differ used in decision making problems?, *Procedia Computer Science*, 192, pp. 4570–4577.

Smith, J. E., J. S. Dyer (2021) On (measurable) multiattribute value functions: An expository argument, *Decision Analysis*, 18(4), pp. 247–256.

Sokol-Hessner, P., M. Hsu, N. G. Curley, M. R. Delgado, C. F. Camerer, E. A. Phelps (2009) Thinking like a trader selectively reduces individuals' loss aversion, *Proceedings of the National Academy of Sciences*, 106(13), pp. 5035–5040.

Soll, J. B., K. L. Milkman, J. W. Payne (2015) A user's guide to debiasing, in: *The Wiley Blackwell handbook of judgment and decision making*, Wiley Online Library, pp. 924–951.

Sun, G., M. Kroesen, J. Rezaei (2025) Anchoring bias in value function elicitation within multi-attribute value theory, *Decision Analysis*, 22(4), pp. 284–304.

Thaler, R. (1980) Toward a positive theory of consumer choice, *Journal of Economic Behavior & Organization*, 1(1), pp. 39–60.

Thaler, R. H., C. R. Sunstein (2021) *Nudge: The final edition*, Penguin, London.

Thomas, A. K., P. R. Millar (2012) Reducing the framing effect in older and younger adults by encouraging analytic processing, *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 67(2), pp. 139–149.

Tom, S. M., C. R. Fox, C. Trepel, R. A. Poldrack (2007) The neural basis of loss aversion in decision-making under risk, *Science*, 315(5811), pp. 515–518.

Tversky, A., D. Kahneman (1981) The framing of decisions and the psychology of choice, *Science*, 211(4481), pp. 453–458.

Tversky, A., D. Kahneman (1991) Loss aversion in risaess choice: A reference-dependent model, *The Quarterly Journal of Economics*, 106(4), pp. 1039–1061.

Watson, S. R., D. M. Buede (1987) *Decision synthesis: The principles and practice of decision analysis*, Cambridge University Press, Cambridge.

Yaniv, I. (2011) Group diversity and decision quality: amplification and attenuation of the framing effect, *International Journal of Forecasting*, 27(1), pp. 41–49.

Chapter 5

The Mitigation Role of Multi-Attribute Value Theory on Status Quo Bias

---

Abstract: Status quo bias refers to individuals' preference for their current options while objectively equivalent or superior options are available. Although this bias has been widely studied in laboratory experiments and field studies, its operation within formal decision analysis methods remains largely unexplored. Existing theoretical work suggests that status quo bias is a strategy-based bias and should therefore be mitigable through structured decision analysis. However, empirical evidence for this claim is limited. This study examines the mechanism through which the status quo bias enters the elicitation process within a decision analysis method and examines whether the use of a decision analysis method mitigates this bias in multi-attribute decision-making. We designed a smartphone purchase problem under three status quo conditions and collected data from 312 participants. The results show strong status quo effects in participants' unaided rankings and choices, and a substantial reduction in these effects after applying the multi-attribute value theory (MAVT). Further analyses reveal that the bias operates primarily through the weight elicitation rather than the value function elicitation. These findings provide new insights into when and how decision analysis methods can mitigate status quo bias and offer implications for bias mitigation in multi-attribute decision-making.

Keyword: Status quo bias; multi-attribute decision-making; decision analysis method; bias mitigation

This chapter is based on the following manuscript:

Sun, G., Kroesen, M., Rezaei, J. The Mitigation Role of Multi-Attribute Value Theory on Status Quo Bias (*manuscript under review*)

5.1 Introduction

Real-world decisions are influenced by time pressure, limited information, and various cognitive and motivational factors (Simon, 1955; Payne et al., 1993; Kahneman, 2002). A substantial body of literature in behavioral decision research has documented that decision-makers (DMs) systematically deviate from normative decision models, and that these deviations are often driven by cognitive biases. Cognitive biases refer to the systematic error in judgment that occurs when people process and interpret information (Tversky & Kahneman, 1974). Such biases can lead to flawed judgment and, consequently, suboptimal choices.

Status quo bias refers to the tendency of DMs to prefer the current or default option over available alternatives, even when alternative options are objectively equivalent or superior (Samuelson & Zeckhauser, 1988; Kahneman et al., 1991). Status quo bias has been robustly demonstrated in controlled laboratory experiments and in a variety of field settings, including pension plan enrollment, insurance decisions, and consumer choice (Madrian & Shea, 2001; Johnson et al., 1993; Hartman et al., 1991). These findings suggest that the way in which options are framed relative to an existing state or default can substantially distort preferences and lead to inertia, under-adjustment, or resistance to welfare-improving changes. Research on status quo bias mostly concern unaided decision-making, where individuals make choices without the support of a formal method or a decision analyst (Montibeller & von Winterfeldt, 2024).

Despite this extensive literature, far less is known about how status quo bias operates within aided decision-making, that is, when individuals rely on a formal decision analysis method to elicit preferences and make decisions. Classical decision analysis argues that structured elicitation should help reduce decision-making biases by promoting deeper, more reflective thinking (Keeney, 1977, 2004). However, recent work in behavioral decision-making suggests that biases can still arise within the application of decision analysis methods. For example, Jacobi & Hobbs (2007) document value-tree induced biases in several weight elicitation methods, and Sun et al. (2025) show that anchoring bias can substantially affect elicited value functions when a starting point is introduced in the elicitation process. Importantly, recent comparative evidence indicates that not all decision analysis methods are equally susceptible to such effects. Rezaei et al. (2024) show that anchoring systematically affects weight elicitation in SMART and Swing but is substantially mitigated in the Best-Worst Method, suggesting that certain structured procedures have greater potential to reduce bias than others. Montibeller & von Winterfeldt (2015) reviewed the cognitive and motivational biases within decision analysis, and argue that association-based and psychophysically-based biases may persist within decision analysis, whereas strategy-based biases are more amenable to mitigation (Arkes, 1991; Harkness et al., 1985; Tetlock & Kim, 1987). Yet these arguments are largely conceptual, and there is limited empirical evidence on whether, and to what extent, structured decision analysis methods actually reduce status quo bias in multi-attribute decision-making contexts.

The present study addresses this gap by investigating the impact of status quo bias in a multi-attribute decision problem and examining whether a structured decision analysis method can reduce its influence. We focus on a smartphone purchase context with two attributes (cost and memory) and distinguish between two sources of status quo: (i) a real status quo, corresponding to the participant's current condition, and (ii) an experimentally provided status quo, embedded in the decision scenario.

To this end, we designed an experiment in which participants first ranked a set of smartphone alternatives under different status quo conditions, and subsequently evaluated the alternatives with the MAVT procedure. By comparing choices and rankings before and after the MAVT procedure, we assess (i) the extent to which framing an option as the status quo shifts preferences toward that option and nearby alternatives, and (ii) whether the decision analysis method mitigates these shifts. In doing so, the study provides empirical evidence on the bias mitigation potential of decision analysis methods, specifically MAVT.

The remainder of this paper is organized as follows. Section 5.2 reviews the literature on status quo bias, reference-dependent preferences, and bias mitigation in decision analysis, and develops our hypotheses. Section 5.3 introduces the decision analysis method used in this study. Section 5.4 describes the experimental design and implementation. Section 5.5 presents the empirical results and discusses the implications for decision analysis and behavioral decision research, and Section 5.6 concludes with limitations and directions for future work.

5.2 Theoretical Background and Hypotheses

The status quo bias refers to the tendency for decision-makers (DMs) to prefer the current or default option over available alternatives (Samuelson & Zeckhauser, 1988). Seminal laboratory experiments have repeatedly demonstrated that individuals exhibit a disproportionate preference for existing states, even when switching yields objectively equivalent or superior outcomes. In the seminal study by Samuelson & Zeckhauser (1988), the authors investigated status quo bias in various hypothetical decision problems, including budget allocation, wagon color choice, investment portfolios, and college job scenarios. Participants were presented with neutral and status quo versions of the same decision problem, and the proportion of participants choosing an option was substantially higher when it was framed as the status quo than when it was not. Subsequent work showed that the impact of the status quo becomes even stronger as the number of available alternatives increases (Tversky & Shafir, 1992; Redelmeier & Shafir, 1995).

Field studies have further demonstrated that status quo bias is pervasive in real-world decision-making. Defaults and pre-existing options have been shown to influence behavior in domains such as pension plan enrollment (Madrian & Shea, 2001), insurance decisions (Johnson et al., 1993), public policy making (Lang et al., 2021), energy-related choice (Blasch & Daminato, 2020), retirement savings decisions (Choi et al., 2004), and consumer product selection (Dhar, 1997). Collectively, this body of work shows that status quo bias is a pervasive phenomenon in human choice, observable in both controlled laboratory environments and complex real-world settings.

Status quo bias is often examined in the broader context of reference-dependent preferences, a concept rooted in Prospect Theory. Kahneman & Tversky (1979) demonstrated that individuals evaluate outcomes relative to a reference point rather than in absolute terms. This reference point serves as a psychological benchmark that determines whether an outcome is perceived as a gain or a loss. The current or default option is a natural candidate for such a reference point. Departures from the reference point are typically experienced as losses, and because losses weigh more heavily than gains, individuals tend to remain with the status quo. This explanation follows directly from the principles of reference dependence and loss aversion developed in the Prospect Theory literature (Tversky & Kahneman, 1991; Kahneman et al., 1991; Thaler, 1980)

and is consistent with later empirical evidence (Moshinsky & Bar-Hillel, 2010). Research in this tradition has further shown that reference points systematically shape preferences between alternatives. When two options, A and B, are evaluated relative to a reference point that lies closer to A, individuals disproportionately favor A; when the reference shifts closer to B, the preference reverses (Kahneman et al., 1991). Thus, even when the objective attributes of A and B remain constant, their perceived desirability depends on their relative position to the reference point, illustrating how reference dependence can generate strong and predictable shifts in choice behavior.

Importantly, research has shown that the reference point does not always coincide with an individual's current state; it can also arise from expectations, norms, or externally provided cues (Kőszegi & Rabin, 2006; Masatlioglu & Uler, 2013; Bleichrodt, 2007). In expectations-based models, for instance, overall utility combines standard "consumption utility" with gain-loss utility defined over deviations from a reference outcome given by the DM's rational expectations (Kőszegi & Rabin, 2006). This implies that status quo bias may more broadly reflect a reference effect (Masatlioglu & Uler, 2013; Munro & Sugden, 2003; Herne, 1998), whereby evaluations of all options are distorted by their proximity to a reference point, of which the status quo is a prominent special case.

Several additional mechanisms have been proposed to explain why DMs exhibit a preference for the status quo. First, a rational explanation is that switching away from the current option may involve transition costs. Such costs can be broadly grouped into two categories: uncertainty about the consequences of switching (Anderson, 2003; Ritov & Baron, 1992; Nebel, 2015) and the effort or disutility associated with adapting to a new state (Nebel, 2015). Beyond these rational considerations, a range of psychological mechanisms can also account for status quo bias. For instance, individuals may anchor on the current situation and insufficiently adjust their evaluations when considering alternatives (Tversky & Kahneman, 1974; Beshears et al., 2009), which can reinforce the attractiveness of the status quo. Other accounts emphasize existence and longevity biases (Eidelman & Crandall, 2012), mere exposure effects (Eidelman & Crandall, 2014), and emotional reactions to change (Ritov & Baron, 1992; Shevchenko et al., 2014; Loewenstein et al., 2001). Taken together, these explanations suggest that status quo bias can arise from a combination of both rational transition costs and non-rational psychological factors, such as reference-dependent evaluation, heuristic anchoring, and affective responses.

Although much of the literature has focused on documenting status quo bias, several strands of research suggest that it can be reduced. Interventions such as providing more information about both the status quo and the change (Wiedmann et al., 2011; Kim, 2010), manipulating or reversing the status quo (Labrecque et al., 2017; Bostrom & Ord, 2006), training in decision-making principles (Morewedge et al., 2015), and relying on more experienced decision-makers (Bellé et al., 2018) have shown some promise in reducing the tendency to stick with the current option. These approaches highlight that status quo bias can be influenced by how the decision context and decision process are structured.

In parallel, the literature on decision analysis suggests that formal, structured elicitation procedures can reduce certain cognitive and motivational biases (Montibeller & von Winterfeldt, 2015; Keeney, 2004). The effectiveness of bias mitigation depends on the psychological origin of the bias. Biases can be classified as strategy-based, association-based, or psychophysically-based (Arkes, 1991; Montibeller & von Winterfeldt, 2015). Strategy-based biases arise from the use of suboptimal heuristics or simplification strategies. Status quo bias falls into this class

because default choices can serve as a cognitively efficient heuristic (Montibeller & von Winterfeldt, 2015). Association-based biases arise from the automatic activation of mental associations stored in memory, which becomes a liability when the associations triggered are judgmentally irrelevant or misleading. Psychophysically-based biases are rooted in perceptual limitations and the nonlinear mapping of physical stimuli to psychological responses. Decision analysis methods can be particularly effective in correcting strategy-based biases, since adopting the method encourages deeper thinking and replaces intuitive heuristics with structured reasoning. In contrast, association-based and psychophysically-based biases are embedded in cognitive and perceptual processes, making them much harder to mitigate. However, despite these theoretical arguments, the extent to which formal decision analysis methods can reduce strategy-based biases such as status quo bias remains largely untested empirically.

Most existing experimental studies of reference-dependent evaluation remove the status quo option from the choice set or make it dominated by the alternatives, in order to isolate the psychological reference effect. While such designs are useful for theoretical identification, they limit the applicability of findings to real-world contexts, where the status quo option is typically included among viable alternatives and not necessarily inferior (Bleichrodt, 2007). Realistic multi-attribute decisions, such as purchasing a product, selecting an investment, or choosing a service, often involve comparing the status quo to similar alternatives rather than rejecting it outright. Understanding how reference dependence operates in this more ecologically valid setting, and how it interacts with structured decision analysis procedures, remains an open question.

This study addresses these gaps by investigating whether a structured decision analysis method can reduce the influence of status quo bias, and whether its effectiveness depends on the source of the reference point. We distinguish between (i) a real status quo, corresponding to the DM's current state, and (ii) an experimentally provided status quo, embedded within the decision context. By investigating the bias in a multi-attribute decision problem and systematically analyzing its impact on preference construction, this research provides novel empirical evidence on when and how formal decision processes can mitigate reference-dependent biases.

Given the well-documented presence of status quo bias and reference-dependent preference in the literature, our first hypothesis is not intended as a novel contribution but serves to formally provide a baseline for the analysis. The second hypothesis concerns our main contribution and examines whether the decision analysis method reduces the bias.

H1: When an option is framed as the status quo, decision-makers will exhibit a tendency to favor the status-quo-aligned option and nearby alternatives.

H2: The tendency to favor the status-quo-aligned option and nearby alternatives when an option is framed as the status quo will be mitigated through the use of MAVT.

5.3 Multi-Attribute Value Theory

Multi-Attribute Value Theory (MAVT) is a foundational and widely acknowledged method in the multi-attribute decision-making (MADM) field (Keeney & Raiffa, 1976; Dyer & Sarin, 1979; Farquhar, 1984; Golabi et al., 1981). It quantifies the decision-maker's (DM's) preference through an overall value function that combines the alternative's performance on the attributes with the corresponding attribute weights. The associated assessment procedures not only yield a

formal preference model for evaluating and comparing alternatives, but also encourage the DM to reflect systematically and in depth on their objectives and tradeoffs (Keeney, 1977), which is particularly valuable in complex, high-stakes decision contexts.

The method typically starts with defining the decision context, clarifying the objectives, and identifying the attributes through which these objectives are measured. Alternatives are then described in terms of their performance levels on each attribute.

For each attribute, MAVT requires an attribute-specific value function that describes how desirable different performance levels are to the DM. These attribute-specific values are usually normalized between 0 and 1, where 0 corresponds to the least preferred level in the attribute range and 1 to the most preferred. Different elicitation methods have been developed for constructing such value functions, such as the midvalue splitting procedure, direct rating, the standard difference procedure, and the curve fitting approach (Farquhar, 1984; Beinat, 1997; Fishburn, 1967; Keeney & Raiffa, 1976; Smith & Dyer, 2021). In our study, we adopted two procedures: the midvalue splitting procedure for the continuous attribute and the direct rating for the discrete attribute. In the midvalue splitting procedure, the DM identifies several indifference points (midvalue points) in the attribute range such that the improvements in each sub-interval are equally desirable for the DM. By plotting the midvalue points, the value function can be obtained. In the direct rating method, the DM directly assigns a value to each performance level, and is thus more suitable for discrete attributes.

The original weight elicitation method within MAVT is the tradeoff procedure (Keeney & Raiffa, 1976; Keeney, 1977), although other weight elicitation methods such as the simple multi-attribute rating technique (SMART) (Edwards, 1977), Swing (von Winterfeldt & Edwards, 1993), are also widely used as alternatives in practice due to their simplicity. This study used the tradeoff procedure to elicit the weights. In the tradeoff procedure, indifference pairs are constructed by eliciting the attribute levels that make the DM indifferent between consequences. Each indifference pair captures the tradeoff the DM is willing to make between the most important attribute and the remaining attributes. The attribute weights (or scaling constants) are then calculated based on these indifferent relationships between attributes.

Under conditions of mutual preferential independence and difference independence for the DM's preference structure over the attributes (Keeney & Raiffa, 1976; Dyer & Sarin, 1979), the overall value of an alternative can be represented by an additive value function that aggregates the attribute-specific value functions and the weights:

$$v(a_i) = \sum_{j=1}^{N} w_j v_j(a_{ij}) \tag{5.1}$$

where $v(a_i)$ is the overall value of alternative i. $v_j(a_{ij})$ is the value of the performance of a_i on attribute j, and w_j is the corresponding weight (or scaling constant). The additive form is widely used in practice due to its transparency and interpretability.

Based on the overall value of each alternative, the alternatives can be compared. The alternative with the highest value is identified as the most preferred. This comparison establishes the preference ordering implied by the DM's elicited preferences in the process. For more detailed information, please read Keeney & Raiffa (1976).

5.4 Experiment Design

To test the hypotheses, we developed a questionnaire based on the MAVT procedure using the online platform Qualtrics. The questionnaire has a total of seven parts: (i) informed consent, (ii) decision problem presentation, (iii) holistic ranking, (iv) preference structure check, (v) weight elicitation, (vi) value function elicitation, and (vii) demographic and current status quo collection. This section describes the content and purpose of each part.

In the first part, participants were informed of the experiment's purpose, procedures, potential risks, and benefits. They were then asked to voluntarily provide informed consent to participate in the study.

In the second part, participants were presented with a hypothetical smartphone purchase problem involving two attributes, cost and memory (see Table 5.1). We chose this context because smartphone choice is a familiar and widely experienced decision problem, particularly for participants recruited via an online panel. Most participants are likely to own a smartphone, which allows us to elicit their actual current device as a real status quo and use this information in our analysis. We used a between-subject design where each participant was randomly assigned to one of the three status quo groups: (i) SQ-A, (ii) No-SQ, and (iii) SQ-C. In SQ-A, (64 GB, €380), a status quo with both low price and low memory was provided; in No-SQ, there is no experimentally provided status quo; in SQ-C, (1024 GB, €1180), a status quo with both high price and high memory was provided. The between-subjects design and random assignment to conditions allow for a clean identification of treatment effects, which is commonly used in studies on status quo bias (Samuelson & Zeckhauser, 1988; Moleman et al., 2025; Blanchar et al., 2024). It also helps reduce carryover and learning effects that often remain challenges in within-subjects designs (Charness et al., 2012; Maxwell et al., 2017).

Table 5.1: Attributes used in the decision problem

Attribute	**Unit**	**Range**
Memory	Gigabyte	[64, 1024]
Cost	Euro	[380, 1180]

Participants were asked to imagine that they were making a realistic smartphone purchase decision. For instance, in the SQ-A condition, the scenario was described as: "*Imagine your current smartphone, which has 64 GB of memory and originally cost 380 euros, was recently lost. Now you're considering buying a new phone. After narrowing down your choices, you've decided on a specific model. The only decision left is choosing the memory size. Larger memory sizes allow you to store more apps, photos, and videos, but they come at a higher cost. All other features, such as brand, camera quality, design, and processor, are exactly the same.*" In the SQ-C condition, the same description was used except that the status quo option was changed to (1024 GB, €1180). In the NO-SQ condition, no status quo information was provided; instead, participants were simply instructed:"*Imagine you are considering buying a new smartphone. After narrowing down your choices, you have decided on a specific model. The only decision left is choosing the memory size. Larger memory sizes allow you to store more apps, photos, and videos, but they come at a higher cost. All other features, such as brand, camera quality, design, and processor, are exactly the same.*"

In all conditions, the status quo option and the other alternatives belong to the same smart-

phone model and differ only in memory and price. As a result, any possible non-monetary transition costs of switching from the participant's current phone to this new model (e.g., learning a new interface, transferring data, adapting to a new design) are essentially the same for all alternatives. The status quo option does not offer a lower transition cost than the other options. This makes it unlikely that, in our analysis later, the observed preference for the status quo option is driven by different transition cost (Nebel, 2015) but the status quo bias.

After presenting the problem, participants were asked to directly rank five alternatives: (64 GB, €380), (128 GB, €510), (256 GB, €680), (512 GB, €930), (1024 GB, €1180). This allowed us to examine the status quo bias directly, without relying on a decision analysis method.

Parts four, five, and six are steps of MAVT. The additivity assumptions were verified by checking if mutual preference independence and difference independence are satisfied. Then, the tradeoff procedure was used to elicit the attribute weights. The midvalue splitting procedure was used to elicit the value function for the price attribute, and the direct rating procedure was used to elicit the value function for the memory attribute.

In the final part, participants provided demographic information and the memory and price of their current smartphones.

Participants were recruited through the Prolific online platform, which offers pre-screening options and response verification tools to enhance data quality. Since the questionnaire was in English and the price was measured in euros, we limited participation to European citizens and fluent in English using the pre-screening functions. Moreover, the response verification function enables us to reject incomplete answers or those that provide answers outside of the value ranges. After data collection and cleaning, a total of 312 participants were included. The detailed demographics of the sample are presented in Table 5.2.

Table 5.2: Demographic characteristics of participants ($n = 312$)

Characteristics	Levels	Percent
Gender	Female	43.9%
	Male	54.5%
	Other	1.6%
Age	[18,24]	30.4%
	[25,34]	36.5%
	[35,44]	28.2%
	> 44	4.9%
Education	High School	28.5%
	Bachelor's degree	42.6%
	Master's degree	16.7%
	Other	12.2%

5.5 Results and Discussion

This section presents evidence of status quo bias in the decision problem and examines the extent to which the Multi-Attribute Value Theory (MAVT) reduces its influence. We first compare the direct rankings elicited prior to MAVT with the MAVT-implied rankings to evaluate the bias

mitigation effectiveness of the method. We then analyze the overall values of a set of hypothetical alternatives to trace the underlying reference-dependent pattern within the MAVT model. Finally, we investigate how the status quo affects the elicited attribute weights and value functions, and characterize the mechanism through which reference dependence enters the MAVT structure.

5.5.1 Status Quo Bias in Ranking

In our experiment, participants were randomly assigned to one of the three conditions: SQ-A (64 GB, €380), NO-SQ (no experimentally provided status quo), and SQ-C (1024 GB, €1180). Before implementing the tradeoff procedure, it is necessary in MAVT to identify the most important attribute, which then serves as the reference attribute in the construction of indifference pairs. To this end, participants were first asked to choose between the two extreme alternatives in the attribute space, corresponding to the combinations (64 GB, €380) and (1024 GB, €1180). These alternatives coincide with the status-quo-aligned options in the SQ-A and SQ-C conditions, respectively, and therefore also provide a direct comparison of participants' preferences over the two status quo options. Table 5.3 shows that, when a status quo is experimentally provided, participants disproportionately select the alternative aligned with that status quo, whereas in the No-SQ condition, choices are approximately balanced. A Chi-square test confirmed that the distribution of choices differed significantly across conditions, $\chi^2(4, N = 312) = 50.162$, $p < 0.001$. This finding is consistent with the literature that framing an option as the status quo option will result in a significantly higher number of people selecting it compared to when it is not (Samuelson & Zeckhauser, 1988; Kahneman et al., 1990).

Table 5.3: Participants' selection across provided status quo conditions

Provided Status Quo	**Preference** $A \prec C$	**Preference** $A \succ C$	**Preference** $A \sim C$	**Total**
SQ-A: (64 GB, €380)	19 (17.6%)	82 (75.9%)	7 (6.5%)	108
B: No status quo	47 (43.1%)	53 (48.6%)	9 (8.3%)	109
SQ-C: (1024 GB, €1180)	62 (65.3%)	28 (29.5%)	5 (5.2%)	95

Before the MAVT procedures, participants were also asked to rank the five smartphone alternatives directly. Consistent with the reference-dependence literature (Tversky & Kahneman, 1991; Masatlioglu & Uler, 2013), the resulting rank data exhibited a clear status quo bias and reference-dependent pattern across conditions. The distribution of the ranking is different in the three status quo conditions. In SQ-A, participants tended to favor the status quo option and the alternative closest to it: many participants assigned first and second rank to (64 GB, €380) and (128 GB, €510), respectively, whereas the most expensive option, (1024 GB, €1180), was ranked last by the majority of participants. In contrast, in SQ-C, the high-end alternatives received the highest ranks, and the low-end option (64 GB, €380) was most frequently placed last. This reversal across conditions indicates that participants evaluated the same set of alternatives relative to the reference point established by the experimentally provided status quo.

After implementing the MAVT procedures, we also obtained a ranking of the alternatives by comparing their overall values. The resulting ranking patterns for each alternative were markedly more similar across the three status quo conditions. To statistically assess differences

in rank distributions across conditions, we conducted Chi-square tests for each alternative, separately for the self-reported rankings and the MAVT rankings (Table 5.4). For the self-reported rankings, all alternatives showed significant differences across conditions (all $p < 0.05$), indicating strong status quo effects in the initial, unstructured evaluations. In contrast, for the MAVT rankings, three out of five alternatives did not exhibit significant differences across conditions ($p > 0.05$), suggesting that the structured MAVT procedure reduced, though did not completely eliminate, status quo bias. This pattern is inline with the view that decision analysis methods can reduce strategy-based biases by helping to structure the decision problem and encouraging DMs to think deeply (Montibeller & von Winterfeldt, 2015; Keeney, 2004).

Table 5.4: Chi-square test for participants' ranking across provided status quo conditions

Alternative	**Self-reported ranking p (two-sided)**	**MAVT ranking p (two-sided)**
(64 GB, €380)	< 0.001	0.242
(128 GB, €510)	< 0.001	0.006
(256 GB, €680)	0.023	0.351
(512 GB, €930)	< 0.001	0.141
(1024 GB, €1180)	< 0.001	< 0.001

We observe a similar pattern when considering participants' real status quo (their current smartphone) in the NO-SQ condition. The NO-SQ group is split into subgroups based on whether their actual device is closer to the low-end or high-end reference: *closer to A* (storage $\leq$ 128 GB and price $\leq$ €780), *closer to C* (storage $\geq$ 256 GB and price $>$ €780), or *mixed* (in between). The pre-MAVT rankings again show clear reference-dependent shifts, whereas post-MAVT rankings converge across subgroups. Chi-square tests (Table 5.5) confirm that real status quo effects are strong in the direct rankings (all $p < 0.05$ except for (256 GB, €680)) but largely reduced after MAVT, with three out of five alternatives not exhibiting significant differences across conditions.

Table 5.5: Chi-square test for participants' ranking in the real status quo group

Alternative	**Self-reported ranking p (two-sided)**	**MAVT ranking p (two-sided)**
(64 GB, €380)	0.002	0.017
(128 GB, €510)	0.004	0.745
(256 GB, €680)	0.286	0.678
(512 GB, €930)	< 0.001	0.961
(1024 GB, €1180)	0.002	0.020

Taken together, these results suggest that framing an option as the status quo results in a significantly higher number of people preferring it and the nearby alternatives, and that using a decision analysis method reduces this bias, supporting hypotheses 1 and 2, aligning with the literature (Samuelson & Zeckhauser, 1988; Tversky & Kahneman, 1991; Masatlioglu & Uler, 2013; Montibeller & von Winterfeldt, 2015).

5.5.2 Status Quo Bias in MAVT

The preceding results demonstrate that experimentally framing an option as the status quo produces significant reference-dependent distortions in participants' initial rankings, and that the application of MAVT substantially reduces, though does not fully eliminate, these effects. To understand the source and persistence of the bias, the following analysis therefore investigates how the status quo affects the evaluation of alternatives within MAVT and which elements of the elicitation procedure are most susceptible to such biases.

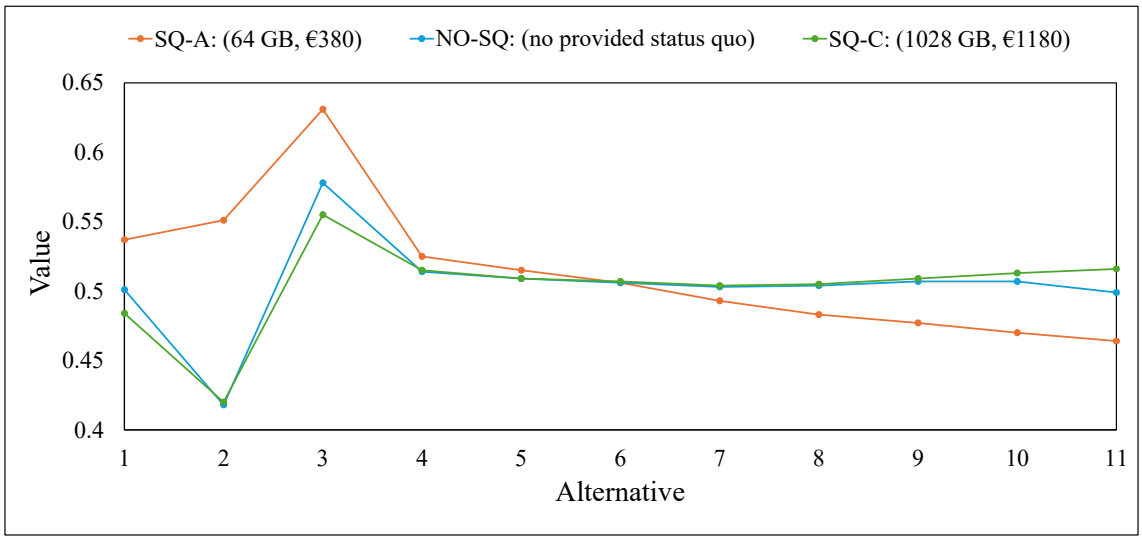
Provided Status Quo Effect

To investigate how the status quo bias influenced the evaluation of status quo and nearby alternatives, nine intermediate hypothetical options were constructed, each representing a 10% increment along the attribute range between the two extreme alternatives (64 GB, €380) and (1024 GB, €1180). These hypothetical options were introduced to create a continuous sequence of attribute combinations that allows us to examine how MAVT evaluations change as alternatives become closer to or farther from the status quo. This provides a cleaner comparison across conditions than relying solely on the original discrete alternatives, as it enables us to trace the gradient of the status quo effect along the attribute space. Although such combinations do not correspond to actual products on the market (since memory sizes are only offered in discrete steps), they are technically feasible configurations and serve as systematic interpolation points for the analysis. Moreover, this more detailed, diagnostic analysis of the bias is possible because MAVT elicits attribute-specific value functions over the entire attribute range, allowing the construction and evaluation of any technically feasible alternatives. Figure 5.1 reports the mean overall values derived from MAVT across the three experimental conditions.

From A1 to A5 (the lower-storage and lower-price range), participants in the SQ-A condition consistently assigned higher overall values than those in the NO-SQ and SQ-C conditions. Conversely, from A7 to A11 (the higher-storage and higher-price range), participants in the SQ-C condition tended to assign higher values, followed by NO-SQ and then SQ-A. For the midpoint alternative (A6), all groups showed nearly identical mean values, indicating convergence around the central attribute level. This pattern suggests that participants generally favored alternatives that were closer to their provided status quo and that this preference diminished progressively as the alternatives diverged from the reference point.

Statistical analyses further confirmed this reference-dependent pattern (Table 5.6). In the lower range (below 10% of the attribute scale), alternatives A1–A2 exhibited significant differences across conditions. Independent t-tests showed that these alternatives received significantly higher overall values in the SQ-A condition than in the SQ-C condition. The Jonckheere–Terpstra test also revealed a significant monotonic trend, with overall values following the ordered pattern SQ-A $>$ NO-SQ $>$ SQ-C.

In the higher range (above 70% of the scale), alternatives A8–A11 showed the reverse effect. Participants in the SQ-C condition assigned significantly higher values than those in SQ-A, and the Jonckheere–Terpstra test confirmed the opposite monotonic trend, SQ-C $>$ NO-SQ $>$ SQ-A. The remaining mid-range alternatives (A4–A7) exhibited no significant group differences, indicating that the bias was strongest near the status-quo reference and weakened for more neutral attribute levels.



	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
SQ-A	0.537	0.551	0.631	0.525	0.515	0.506	0.493	0.483	0.477	0.470	0.464
No-SQ	0.501	0.418	0.578	0.514	0.509	0.506	0.503	0.504	0.507	0.507	0.499
SQ-C	0.484	0.420	0.555	0.515	0.509	0.507	0.504	0.505	0.509	0.513	0.516

Note. Each intermediate alternative represents a 10% increment in both attributes between the two extremes: A1: (64 GB, €380); A2: (160 GB, €460); A3: (256 GB, €540); A4: (352 GB, €620); A5: (448 GB, €700); A6: (544 GB, €780); A7: (640 GB, €860); A8: (736 GB, €940); A9: (832 GB, €1020); A10: (928 GB, €1100); A11: (1024 GB, €1180).

Figure 5.1: Mean overall value of the eleven alternatives under different provided status quo conditions

Table 5.6: Statistical test results for the provided status quo effect

Test	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**	**A11**
Independent t-test (p, 1-sided)	0.006	< 0.001	0.157	0.242	0.327	0.479	0.185	0.045	0.014	0.007	0.006
Jonckheere–Terpstra test (p, 2-sided)	< 0.001	< 0.001	0.240	0.344	0.553	0.910	0.195	0.017	0.002	< 0.001	< 0.001

To more precisely identify the points at which the status quo effect disappeared, additional hypothetical alternatives were generated at 1% increments within the 10–20% and 60–70% ranges of the attribute continuum. Each newly generated alternative was evaluated using the elicited value functions and weights to obtain its overall MAVT value. The analysis revealed that the cutoff points occurred at 11% and 70% of the attribute range. At 11% (corresponding to the hypothetical alternative of 169.6 GB and €468), participants in the SQ-A condition assigned significantly higher overall values than those in the SQ-C condition, $t(201) = 1.723$, $p = 0.043$. For all subsequent alternatives, no significant differences were observed across the status quo conditions (all $p > 0.05$), indicating that the influence of the provided status quo diminished beyond this point. These findings on overall values provide different evidence of the effectiveness of the decision analysis method in reducing the status quo bias.

These findings show that while MAVT reduces cross-condition differences in the ranking of the main alternatives, the underlying overall values still exhibit a clear reference-dependent pattern around the experimentally provided status quo. There is still an underlying reference-dependent distortion embedded in the preference elicitation procedures that continues to shape

the overall values of the alternatives.

Real Status Quo Effect

We also investigate the impact of participants' real status quo. Table 5.7 reports the mean overall values for each subgroup in the NO-SQ condition across the eleven alternatives. Descriptively, the results suggest a weak pattern of reference dependence. Participants whose real devices resembled the low-end reference (Closer to SQ-A) tended to assign slightly higher overall values to the lower-storage and lower-price alternatives (A1–A2), whereas those whose devices were closer to the high-end reference (Closer to SQ-C) tended to favor the higher-end alternatives (A4–A11). For A3, all three groups yielded somewhat similar mean values.

Table 5.7: Mean overall value of the eleven alternatives in No-SQ by real status quo condition

Condition	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
Closer to SQ-A	0.517	0.425	0.517	0.510	0.504	0.501	0.498	0.498	0.502	0.502	0.483
In between A and C	0.503	0.421	0.525	0.512	0.504	0.499	0.494	0.495	0.495	0.496	0.497
Closer to SQ-C	0.461	0.394	0.521	0.527	0.531	0.536	0.537	0.540	0.546	0.548	0.540

However, the statistical tests indicate that these differences are not significant. As shown in Table 5.8, the independent t-tests and Jonckheere–Terpstra tests revealed no significant group differences across most alternatives (all $p > 0.05$), except for the two status quo-aligned options (A1 and A11). Specifically, these two alternatives showed marginally significant differences across real status quo groups (Jonckheere–Terpstra: both $p = 0.043$).

Table 5.8: Statistical test results for the real status quo effect

Test	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
Independent t-test (p, 1-sided)	0.101	0.208	0.462	0.318	0.215	0.138	0.104	0.098	0.110	0.129	0.101
Jonckheere–Terpstra test (p, 2-sided)	0.043	0.514	0.739	0.802	0.628	0.509	0.364	0.291	0.276	0.273	0.043

The results indicate that while the descriptive pattern suggests a mild reference-dependent tendency, the statistical evidence does not support a systematic status quo bias or reference-dependent preference when the status quo is self-owned. This suggests that the source of the status quo may influence the extent to which the status quo bias can be mitigated. The real (self-owned) status quo represents the individual's current option in real life and is thus more likely to be treated as a decision strategy, corresponding to the strategy-based bias (Montibeller & von Winterfeldt, 2015; Arkes, 1991). In contrast, an experimentally provided status quo is introduced directly within the decision context. Rather than representing a pre-existing preference, it becomes part of the preference construction process itself and can be more difficult to mitigate.

5.5.3 Status Quo Bias in Weights

To understand how an experimentally provided status quo affects the preference elicitation in MAVT, we further examined its influence on attribute weights and attribute-specific value functions. This subsection focuses on the effect on weights.

We understand that under additive MAVT with normalized single-attribute value functions ($v_m(64) = 0$, $v_m(1024) = 1$; $v_c(1180) = 0$, $v_c(380) = 1$), the two extreme alternatives directly reveal the attribute weights: $v(64, 380) = w_m\, v_m(64) + w_c\, v_c(380) = w_c$, $v(1024, 1180) = w_m\, v_m(1024) + w_c\, v_c(1180) = w_m$. Hence, differences in the overall values of (64 GB, €380) and (1024 GB, €1180) across SQ conditions map one-to-one to differences in w_c and w_m. Here, we present them again to clearly understand the effect on weights.

As shown in Table 5.9, participants in group SQ-A placed greater emphasis on *price* ($w_c = 0.537$ vs. $w_m = 0.464$), whereas those in group SQ-C placed greater emphasis on *storage* ($w_m = 0.516$ vs. $w_c = 0.484$). Participants in the NO-SQ condition exhibited nearly equal weights ($w_c = 0.501$ and $w_m = 0.499$), suggesting a balanced tradeoff between attributes in the absence of an explicit status quo.

Table 5.9: Attribute weights under different provided status quo conditions

Condition	**Weight of Cost** (w_c)	**Weight of Memory** (w_m)
SQ-A: (64 GB, €380)	0.537	0.464
No-SQ (no provided status quo)	0.501	0.499
SQ-C: (1024 GB, €1180)	0.484	0.516

Statistical analyses confirmed that these differences were significant (Table 5.10). The independent t-tests showed that both w_c and w_m differed significantly between the two status-quo–aligned groups ($p = 0.006$ for both attributes). The Jonckheere–Terpstra test further revealed a significant monotonic trend across the three conditions ($p < 0.001$), with the relative importance shifting systematically from *cost* (SQ-A) to *memory* (SQ-C) as the provided status quo increased in performance and price.

Table 5.10: Statistical test results for the weights

Test	**Weight of Cost** (w_c)	**Weight of Memory** (w_m)
Independent t-test (p, 1-sided)	0.006	0.006
Jonckheere–Terpstra test (p, 2-sided)	< 0.001	< 0.001

5.5.4 Status Quo Bias in Attribute-Specific Value Functions

We next tested whether the status quo affected the attribute-specific value functions. For each attribute, we summarized the elicited function by its area under the curve (AUC) on the normalized attribute scale. We use the AUC as a summary index of the function's shape, because a simple categorical classification into concave or convex may fail to capture more subtle changes in curvature, such as shifts from moderately to strongly concave value functions (Sun et al., 2025). We compared AUCs between SQ-A and SQ-C, between real-SQ groups, and within matched real/provided SQ subsets using independent t-tests and Mann-Whitney U tests. All comparisons were non-significant, indicating that the elicited value functions were statistically indistinguishable across conditions. Figure 5.2 plots the two average value functions.

Taken together, these findings suggest that the influence of the status quo bias on the decision outcome operates primarily through the weight elicitation process rather than through

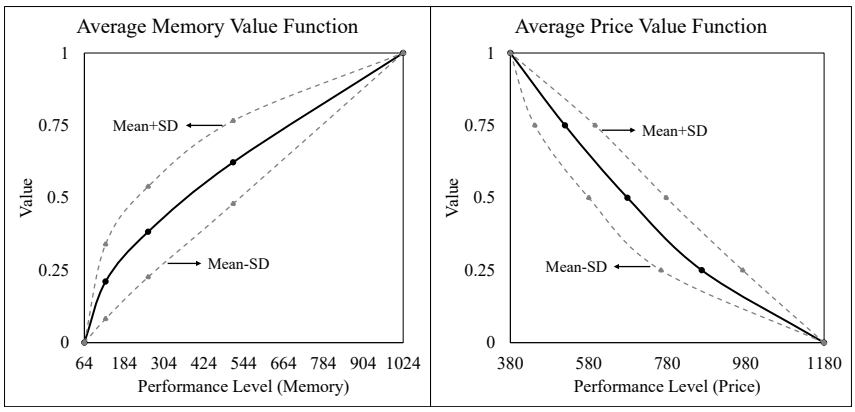


Figure 5.2: Average value functions

the elicitation of the attribute-specific value functions. Because the elicited value functions are statistically indistinguishable across conditions, any observed differences in overall evaluations must arise from differences in the attribute weights rather than from distortions in the value functions themselves. This pattern is consistent with the structure of MAVT: the elicitation of value functions is attribute-specific and reference-neutral, the evaluation is based on the whole attribute range rather than relative to the status quo. As a result, the status quo bias and reference-dependent mechanism have limited opportunity to affect the curvature of the value functions.

In contrast, the weight elicitation procedure provides a natural entry point for reference-dependent distortions. The tradeoff method begins by asking participants to compare the two alternatives at the extreme points of the attribute space (64 GB, €380) and (1024 GB, €1180) to determine which attribute is more important and to construct the indifference relations. Because these two alternatives coincide with the provided status quo options in the experiment, any preference advantage due to the status quo bias is directly expressed during these tradeoff judgments. This makes the weight elicitation step particularly susceptible to status quo bias, whereas the value function elicitation step remains unaffected.

5.5.5 The Mechanism of Status Quo Bias in MAVT

Our results show that a status quo located at one extreme of the attribute space systematically shifts the tradeoff weights. When the status quo performs strongly on one attribute and weakly on another, participants tend to place greater emphasis on the dimension on which the status quo performs well and de-emphasize the dimension on which it performs poorly. For example, when the provided status quo is memory-intensive (high storage, high price), participants assign a higher weight to memory (w_m) and a lower weight to price (w_c). Conversely, when the status quo is cost-efficient (low storage, low price), the weighting pattern reverses.

For any alternative a_i with attribute levels (a_{im}, a_{ic}), its overall value under MAVT is

$$v(a_i) = w_m v_m(a_{im}) + w_c v_c(a_{ic}).$$

Our experimental results showed that the attribute-specific value functions were statistically indistinguishable across conditions, indicating that status quo bias does not distort the curvature

of these functions. Thus, the influence of the status quo must operate through the weights. The key question, then, is how changes in the weight affect the evaluation of alternatives across status-quo conditions.

To formalize this mechanism, consider an alternative a_k with fixed attribute-specific values $v_m = v_m(a_{km})$ and $v_c = v_c(a_{kc})$, which do not differ across conditions. Under the memory-intensive status quo (SQ-A), participants assign a higher weight to memory and a lower weight to price:

$$w_m^{SQ-A} > w_c^{SQ-A}$$

Under the cost-efficient status quo (SQ-C), the pattern reverses:

$$w_m^{SQ-C} < w_c^{SQ-C}$$

Because weights in MAVT are normalized to 1,

$$w_m^{SQ-A} + w_c^{SQ-A} = w_m^{SQ-C} + w_c^{SQ-C}$$

which implies

$$w_m^{SQ-A} - w_m^{SQ-C} = -(w_c^{SQ-A} - w_c^{SQ-C})$$

The overall value of the *same* alternative under the two conditions is

$$v^{SQ-A} = w_m^{SQ-A} v_m + w_c^{SQ-A} v_c, \qquad v^{SQ-C} = w_m^{SQ-C} v_m + w_c^{SQ-C} v_c$$

The difference simplifies to

$$v^{SQ-A} - v^{SQ-C} = (w_m^{SQ-A} - w_m^{SQ-C})(v_m - v_c)$$

This expression reveals the key condition. If an alternative A is closer to SQ-A in attribute space, then it satisfies $v_m > v_c$. Combined with $w_m^{SQ-A} > w_m^{SQ-C}$, we obtain:

$$v^{SQ-A} > v^{SQ-C} \quad \Longleftrightarrow \quad v_m > v_c$$

Thus, alternatives that resemble the status quo receive higher overall values because the attribute on which they perform relatively well is precisely the attribute that receives greater weight. The attractiveness of the nearby alternatives to the status quo option, therefore, arises from the alignment between (i) the weighting pattern distorted by the status quo and (ii) the alternative's position in the attribute space.

A similar logic explains why changes in the weight vector have little impact on *middle*, *underperforming*, and *overperforming* alternatives. All three types share the property that their attribute-specific values are roughly equal across dimensions ($v_m \approx v_c$): middle alternatives have moderate values on both attributes, underperforming alternatives have low values on both, and overperforming alternatives have high values on both. When $v_m \approx v_c$, the overall value simplifies to

$$v(a_k) = w_m v_m(a_{km}) + w_c v_c(a_{kc}) \approx v_m \approx v_c,$$

no matter how the weights change across conditions, w_m and w_c always sum to one. Since $v_m \approx v_c$, the increase in one weight is almost exactly offset by the decrease in the other,

resulting in similar overall values for the same alternative across conditions.

Finally, this mechanism helps explain why the MAVT procedure reduces the overall influence of the status quo in the ranking. Status quo bias refers to the tendency to remain with the current option even when objectively superior alternatives exist. In MAVT, each alternative is evaluated systematically based on its attribute performance, and clearly overperforming alternatives achieve high overall values regardless of their similarity to the status quo. The influence of reference-dependent distortions in weights is therefore largely confined to alternatives located very close to the status quo in attribute space, which limits the overall effect of the status quo on the ranking of the main alternatives.

The findings of this study indicate that structured and decomposed preference elicitation procedures can substantially reduce the influence of status quo bias on the final decision outcome. MAVT decomposed the preference elicitation into two main steps. It elicits attribute-specific value functions and then aggregates them. This decomposition appears to shield the process from reference-dependent distortions. However, the tradeoff weight elicitation step requires a direct comparison of alternatives at the extreme attribute levels. When these alternatives coincide with the status quo options, the weight judgments become systematically biased. This pattern suggests that MADM methods relying on holistic judgments or alternative-to-alternative comparisons may be more vulnerable to status quo bias, as the full presentation of alternatives provides room for the status quo to operate as a salient reference point.

Conversely, decomposed procedures break the decision into smaller tasks and remove the reference point (status quo) from the judgment context. By doing that, it replaces intuitive and simplified strategies with more effortful reasoning and reflective evaluation. The results, therefore, point to a broader implication that structured and decomposed MADM approaches may help mitigate not only status quo bias but also other strategy-based biases.

5.6 Conclusion

This study examined how status quo bias operates in a multi-attribute decision environment and whether a structured decision analysis method can reduce its influence. We developed two hypotheses: that decision-makers would exhibit systematic preference for the status quo option and its nearby alternatives (H1), and that a formal elicitation procedure, the Multi-Attribute Value Theory (MAVT), would mitigate this tendency (H2). We further distinguished between a real status quo, corresponding to participants' current states, and an experimentally provided status quo embedded in the decision frame.

We conducted an experiment in a smartphone purchase context involving two attributes (memory and cost). Participants first made direct choices and rankings, after which they completed the MAVT procedures. The results provided clear support for H1: in both choice and direct ranking tasks, participants disproportionately favored the option framed as the status quo and alternatives close to it. This pattern was observed for both experimentally induced and naturally occurring (self-owned) status quo positions. After the MAVT elicitation, however, ranking patterns became substantially more similar across conditions, and most alternatives no longer exhibited statistically significant differences in ranks across status quo groups. These findings indicate that MAVT reduces cross-condition variability in the ordering of alternatives,

consistent with the claim that structured procedures can reduce strategy-based bias. At the same time, analyses of the overall value scores for a set of hypothetical alternatives revealed persistent reference-dependent differences around the experimentally provided status quo. These differences were driven primarily by systematic shifts in attribute weights rather than by changes in the elicited value functions.

The findings are consistent and contribute to several strands of literature. They corroborate extensive evidence that status quo bias is pervasive in human decision-making (Samuelson & Zeckhauser, 1988; Kahneman et al., 1990; Madrian & Shea, 2001), and align with models of reference-dependent preferences in which outcomes near a reference point are evaluated more favorably (Tversky & Kahneman, 1991; Kahneman et al., 1991; Kőszegi & Rabin, 2006). The results also provide empirical evidence to the argument that formal decision analysis methods can mitigate certain cognitive and motivational biases (Keeney, 2004; Montibeller & von Winterfeldt, 2015; Arkes, 1991). While MAVT substantially reduced the divergence in preference orderings across status quo conditions, it did not eliminate underlying reference-dependent distortions in the quantitative evaluation of alternatives when the status quo was experimentally provided.

Several limitations indicate opportunities for future research. First, the present study compared only two types of status quo: real and experimentally provided ones. Future work could examine the interactions between the two types and additional sources of status quo, such as expectations, norms, or aspirational benchmarks (Kőszegi & Rabin, 2006; Bleichrodt, 2007). Second, other decision analysis methods (e.g., SMART and Swing) may interact differently with reference-dependent evaluation, and systematic comparisons would be valuable. Third, the decision task involved two attributes and a limited number of alternatives; real-world decisions often involve more attributes, more complex attribute structures, continuous attribute ranges, and larger sets of feasible options. Given evidence that the magnitude of status quo bias increases with the number of alternatives (Tversky & Shafir, 1992; Samuelson & Zeckhauser, 1988), it is important to test whether decision analysis methods remain effective in mitigating status quo bias in higher-dimensional contexts. Fourth, the experimental task involved a consumer decision problem; extending the analysis to other domains, such as energy, transportation, environmental planning, or public policy, would help assess the generalizability of the findings. Finally, it would be interesting to test the mitigation effectiveness of decision analysis methods in other strategy-based biases, as suggested by Montibeller & von Winterfeldt (2015).

Bibliography

Anderson, C. J. (2003) The psychology of doing nothing: forms of decision avoidance result from reason and emotion, *Psychological Bulletin*, 129(1), pp. 139–167.

Arkes, H. R. (1991) Costs and benefits of judgment errors: Implications for debiasing, *Psychological Bulletin*, 110(3), pp. 486–498.

Beinat, E. (1997) *Value functions for environmental management*, Springer, Dordrecht.

Bellé, N., P. Cantarelli, P. Belardinelli (2018) Prospect theory goes public: Experimental evi-

dence on cognitive biases in public policy and management decisions, *Public Administration Review*, 78(6), pp. 828–840.

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian (2009) The importance of default options for retirement saving outcomes: Evidence from the united states, in: *Social security policy in a changing environment*, University of Chicago Press, pp. 167–195.

Blanchar, J. C., S. Eidelman, E. Allen (2024) Social change requires more justification than maintaining the status quo, *Frontiers in Social Psychology*, 2, 1360377.

Blasch, J., C. Daminato (2020) Behavioral anomalies and energy-related individual choices: the role of status-quo bias, *The Energy Journal*, 41(6), pp. 181–214.

Bleichrodt, H. (2007) Reference-dependent utility with shifting reference points and incomplete preferences, *Journal of Mathematical Psychology*, 51(4), pp. 266–276.

Bostrom, N., T. Ord (2006) The reversal test: Eliminating status quo bias in applied ethics, *Ethics*, 116(4), pp. 656–679.

Charness, G., U. Gneezy, M. A. Kuhn (2012) Experimental methods: Between-subject and within-subject design, *Journal of Economic Behavior & Organization*, 81(1), pp. 1–8.

Choi, J. J., D. Laibson, B. C. Madrian, A. Metrick (2004) For better or for worse: Default effects and 401 (k) savings behavior, in: *Perspectives on the Economics of Aging*, University of Chicago Press, pp. 81–126.

Dhar, R. (1997) Consumer preference for a no-choice option, *Journal of Consumer Research*, 24(2), pp. 215–231.

Dyer, J. S., R. K. Sarin (1979) Measurable multiattribute value functions, *Operations Research*, 27(4), pp. 810–822.

Edwards, W. (1977) How to use multiattribute utility measurement for social decisionmaking, *IEEE Transactions on Systems, Man, and Cybernetics*, 7(5), pp. 326–340.

Eidelman, S., C. S. Crandall (2012) Bias in favor of the status quo, *Social and Personality Psychology Compass*, 6(3), pp. 270–281.

Eidelman, S., C. S. Crandall (2014) The intuitive traditionalist: How biases for existence and longevity promote the status quo, in: *Advances in experimental social psychology*, Academic Press, New York, pp. 53–104.

Farquhar, P. H. (1984) State of the art—utility assessment methods, *Management Science*, 30(11), pp. 1283–1300.

Fishburn, P. C. (1967) Methods of estimating additive utilities, *Management Science*, 13(7), pp. 435–453.

Golabi, K., C. W. Kirkwood, A. Sicherman (1981) Selecting a portfolio of solar energy projects using multiattribute preference theory, *Management Science*, 27(2), pp. 174–189.

Harkness, A. R., K. G. DeBono, E. Borgida (1985) Personal involvement and strategies for making contingency judgments: A stake in the dating game makes a difference, *Journal of Personality and Social Psychology*, 49(1), pp. 22–32.

Hartman, R. S., M. J. Doane, C.-K. Woo (1991) Consumer rationality and the status quo, *The Quarterly Journal of Economics*, 106(1), pp. 141–162.

Herne, K. (1998) Testing the reference-dependent model: an experiment on asymmetrically dominated reference points, *Journal of Behavioral Decision Making*, 11(3), pp. 181–192.

Jacobi, S. K., B. F. Hobbs (2007) Quantifying and mitigating the splitting bias and other value tree-induced weighting biases, *Decision Analysis*, 4(4), pp. 194–210.

Johnson, E. J., J. Hershey, J. Meszaros, H. Kunreuther (1993) Framing, probability distortions, and insurance decisions, *Journal of Risk and Uncertainty*, 7(1), pp. 35–51.

Kahneman, D. (2002) Maps of bounded rationality: A perspective on intuitive judgement and choice, *American Psychologist*, 58(9), pp. 679–720.

Kahneman, D., J. L. Knetsch, R. H. Thaler (1990) Experimental tests of the endowment effect and the coase theorem, *Journal of Political Economy*, 98(6), pp. 1325–1348.

Kahneman, D., J. L. Knetsch, R. H. Thaler (1991) Anomalies: The endowment effect, loss aversion, and status quo bias, *Journal of Economic perspectives*, 5(1), pp. 193–206.

Kahneman, D., A. Tversky (1979) Prospect theory: An analysis of decision under risk, *Econometrica*, 47(2), pp. 363–391.

Keeney, R. L. (1977) The art of assessing multiattribute utility functions, *Organizational Behavior and Human Performance*, 19(2), pp. 267–310.

Keeney, R. L. (2004) Making better decision makers, *Decision Analysis*, 1(4), pp. 193–204.

Keeney, R. L., H. Raiffa (1976) *Decisions with multiple objectives: Preferences and value trade-offs*, Cambridge University Press, Cambridge.

Kim, H.-W. (2010) The effects of switching costs on user resistance to enterprise systems implementation, *IEEE Transactions on Engineering Management*, 58(3), pp. 471–482.

Kőszegi, B., M. Rabin (2006) A model of reference-dependent preferences, *The Quarterly Journal of Economics*, 121(4), pp. 1133–1165.

Labrecque, J. S., W. Wood, D. T. Neal, N. Harrington (2017) Habit slips: When consumers unintentionally resist new products, *Journal of the Academy of Marketing Science*, 45(1), pp. 119–133.

Lang, C., M. Weir, S. Pearson-Merkowitz (2021) Status quo bias and public policy: evidence in the context of carbon mitigation, *Environmental Research Letters*, 16(5), 054076.

Loewenstein, G. F., E. U. Weber, C. K. Hsee, N. Welch (2001) Risk as feelings, *Psychological Bulletin*, 127(2), pp. 267–286.

Madrian, B. C., D. F. Shea (2001) The power of suggestion: Inertia in 401 (k) participation and savings behavior, *The Quarterly Journal of Economics*, 116(4), pp. 1149–1187.

Masatlioglu, Y., N. Uler (2013) Understanding the reference effect, *Games and Economic Behavior*, 82, pp. 403–423.

Maxwell, S. E., H. D. Delaney, K. Kelley (2017) *Designing experiments and analyzing data: A model comparison perspective*, Routledge, New York.

Moleman, M. L., B. van Wee, L. B. Steketee, N. van den Hurk, M. Kroesen (2025) The role of status quo bias in shaping support for controversial transport policies: The counterfactual test, *Transport Policy*, 171, pp. 453–461.

Montibeller, G., D. von Winterfeldt (2015) Cognitive and motivational biases in decision and risk analysis, *Risk Analysis*, 35(7), pp. 1230–1251.

Montibeller, G., D. von Winterfeldt (2024) Behavioral decision research: Descriptive and prescriptive perspectives, in: *Behavioral Decision Analysis*, Springer, pp. 15–40.

Morewedge, C. K., H. Yoon, I. Scopelliti, C. W. Symborski, J. H. Korris, K. S. Kassam (2015) Debiasing decisions: Improved decision making with a single training intervention, *Policy Insights from the Behavioral and Brain Sciences*, 2(1), pp. 129–140.

Moshinsky, A., M. Bar-Hillel (2010) Loss aversion and status quo label bias, *Social Cognition*, 28(2), pp. 191–204.

Munro, A., R. Sugden (2003) On the theory of reference-dependent preferences, *Journal of Economic Behavior & Organization*, 50(4), pp. 407–428.

Nebel, J. M. (2015) Status quo bias, rationality, and conservatism about value, *Ethics*, 125(2), pp. 449–476.

Payne, J. W., J. R. Bettman, E. J. Johnson (1993) *The adaptive decision maker*, Cambridge university press, Cambridge.

Redelmeier, D. A., E. Shafir (1995) Medical decision making in situations that offer multiple alternatives, *Jama*, 273(4), pp. 302–305.

Rezaei, J., A. Arab, M. Mehregan (2024) Analyzing anchoring bias in attribute weight elicitation of smart, swing, and best-worst method, *International Transactions in Operational Research*, 31(2), pp. 918–948.

Ritov, I., J. Baron (1992) Status-quo and omission biases, *Journal of Risk and Uncertainty*, 5(1), pp. 49–61.

Samuelson, W., R. Zeckhauser (1988) Status quo bias in decision making, *Journal of Risk and Uncertainty*, 1(1), pp. 7–59.

Shevchenko, Y., B. von Helversen, B. Scheibehenne (2014) Change and status quo in decisions with defaults: The effect of incidental emotions depends on the type of default, *Judgment and Decision Making*, 9(3), pp. 287–296.

Simon, H. A. (1955) A behavioral model of rational choice, *The Quarterly Journal of Economics*, 69(1), pp. 99–118.

Smith, J. E., J. S. Dyer (2021) On (measurable) multiattribute value functions: An expository argument, *Decision Analysis*, 18(4), pp. 247–256.

Sun, G., M. Kroesen, J. Rezaei (2025) Anchoring bias in value function elicitation within multiattribute value theory, *Decision Analysis*, 22(4), pp. 284–304.

Tetlock, P. E., J. I. Kim (1987) Accountability and judgment processes in a personality prediction task, *Journal of Personality and Social Psychology*, 52(4), pp. 700–709.

Thaler, R. (1980) Toward a positive theory of consumer choice, *Journal of Economic Behavior & Organization*, 1(1), pp. 39–60.

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

Tversky, A., D. Kahneman (1991) Loss aversion in riskless choice: A reference-dependent model, *The Quarterly Journal of Economics*, 106(4), pp. 1039–1061.

Tversky, A., E. Shafir (1992) Choice under conflict: The dynamics of deferred decision, *Psychological Science*, 3(6), pp. 358–361.

von Winterfeldt, D., W. Edwards (1993) *Decision analysis and behavioral research*, Cambridge University Press, Cambridge.

Wiedmann, K.-P., N. Hennigs, L. Pankalla, M. Kassubek, B. Seegebarth (2011) Adoption barriers and resistance to sustainable solutions in the automotive sector, *Journal of Business Research*, 64(11), pp. 1201–1206.

Chapter 6

Conclusion

This dissertation investigates how four cognitive biases, anchoring bias, loss aversion, the framing effect, and status quo bias, affect the decision-making process and outcomes within multi-attribute value theory (MAVT), and it develops bias mitigation strategies to reduce these effects. Chapter 1 introduced the research background and discussed cognitive bias in both aided and unaided decision-making, distinguishing among three types of bias: strategy-based, association-based, and psychophysically based. It then identified the research gaps, outlined the research focus and research questions, and presented the overall structure of the dissertation. Chapter 2 examined the influence of anchoring bias on attribute-specific value function elicitation in MAVT and developed strategies to mitigate its effect. Chapter 3 extended this investigation to the weight elicitation stage by examining anchoring bias in the traditional tradeoff procedure and demonstrating that the Best–Worst Tradeoff (BWT) method can reduce this bias. Chapter 4 moved beyond single-bias analysis by examining the interaction of two reference-dependent biases, loss aversion and the framing effect, across multiple stages of MAVT, with the aim of identifying how their combined influence manifests in the elicitation process and where targeted mitigation may be feasible. Chapter 5 evaluated the extent to which MAVT can reduce status quo bias and provided insight into the mechanisms through which the method exerts its mitigating effect. In this final chapter, I synthesize the key findings from each study, discuss their theoretical and practical implications for research on behavioral influences in multi-attribute decision analysis, and conclude by outlining the limitations of the current research and promising directions for future work.

6.1 Key Findings

The main research question of this dissertation was examined through three sub-research questions and addressed by four empirical studies. Below, the key findings of each study are presented in relation to the research questions.

RQ1: *How can anchoring bias affect the attribute-specific value function and weight elicitation within multi-attribute value theory, and how can such effects be mitigated?* (Chapter 2 & 3)

This research question investigated anchoring bias (Tversky & Kahneman, 1974; Furnham & Boo, 2011), one of the most pervasive cognitive biases, and examined how it operated within the structured elicitation procedures of MAVT. Anchoring bias arises when individuals insufficiently adjust their judgments away from an initial reference point. MAVT required decision-makers to make quantitative judgments at several stages, each of which could involve explicit or implicit starting points that served as anchors. Understanding how anchoring emerged in these procedures was therefore essential for evaluating the behavioral robustness of MAVT and for identifying opportunities to design elicitation procedures that were less vulnerable to this bias.

To address this research question, a systematic literature review on anchoring bias was conducted. Based on the identified mechanisms and targeted debiasing strategies, hypotheses were developed to examine anchoring effects in the midvalue splitting procedure for attribute-specific value function elicitation and in the tradeoff procedure for weight elicitation. In parallel, alternative elicitation procedures were formulated and tested to assess their potential to reduce anchoring. This motivated two empirical studies, each targeting a different stage of MAVT.

Study 1 (Chapter 2) focused on the attribute-specific value function elicitation step. Using a randomized controlled experiment, it demonstrated that numerical starting points used by analysts in the midvalue splitting procedure (Keeney & Raiffa, 1993) significantly biased the elicited value functions and the resulting decision outcomes. The study also evaluated two bias mitigation strategies, no-anchor and counter-anchoring, incorporated directly into the midvalue splitting procedure. Both strategies demonstrated clear potential to reduce the magnitude of anchoring effects. The chapter further discussed the implications of these findings for other value function elicitation procedures, such as the standard difference procedure, lock-step procedure, and direct rating (Keeney & Raiffa, 1993; Beinat, 1997; Fishburn, 1967). Because these procedures exhibited structural features similar to those of the midvalue splitting procedure, the insights derived from this study could be extended to a broader set of value function elicitation methods.

Study 2 (Chapter 3) extended the investigation to the weight elicitation stage. It examined whether the selection of the attribute (most or least important) used to formulate indifference pairs in the tradeoff procedure distorted judgments due to anchoring bias and led to inconsistent weights across conditions. The study also evaluated the Best–Worst Tradeoff (BWT) method (Liang et al., 2022) as an alternative procedure that could reduce this effect through its symmetric use of both attributes. The results showed that tradeoff judgments were susceptible to anchoring, resulting in biased weights and decision outcomes. In contrast, the BWT method substantially reduced this vulnerability, demonstrating that the structural design of elicitation procedures meaningfully influenced the extent of bias.

Taken together, these studies provided systematic evidence that anchoring entered MAVT at multiple stages, affected both intermediate judgments and final decision outcomes, and could be mitigated through procedural design. The findings contributed to the literature on anchoring in behavioral decision research by clarifying how anchoring bias emerged within MAVT and, more broadly, within MADM procedures.

RQ2: *How do the loss aversion and the framing effect operate across multiple stages of multi-attribute value theory and collectively influence the decision-making process and outcomes?* (Chapter 4)

This research question investigated two forms of gain-loss bias, loss aversion and framing effect, and examined how they influenced multiple elicitation stages within MAVT. Both biases originated from individuals' asymmetric sensitivity to gains and losses (Kahneman et al., 1979; Tversky & Kahneman, 1991, 1981). In the MADM context, gains and losses arose from different elicitation stages and sources. Understanding how these biases emerged, propagated, and interacted across stages of MAVT was therefore essential for assessing the behavioral robustness of the method and for designing elicitation procedures that reduced susceptibility to gain-loss distortions.

To address this question, a systematic literature review on gain-loss bias was conducted, focusing on loss aversion and the framing effect. Potential points within MAVT where gains and losses could be induced were identified, linked to the psychological mechanisms underlying each bias, and used to develop hypotheses regarding their emergence and interaction across elicitation stages. This process motivated a single empirical study.

Study 3 (Chapter 4) developed and tested three hypotheses concerning (i) the loss aversion in the tradeoff procedure, (ii) the framing effect on attribute-specific value functions, and (iii) the interaction of these biases across elicitation stages. The results demonstrated that (i) loss aversion systematically distorted the tradeoff procedure, with gain-framed tradeoff tasks yielding significantly higher weights for the most important attribute than loss-framed tasks; (ii) framing substantially altered attribute-specific value functions, such that gain-framed attributes exhibited higher areas under the curve than loss-framed attributes; and (iii) loss aversion and the framing effect interacted across MAVT stages: the framing-induced differences in the attribute-specific value functions carried over to the tradeoff procedure and influenced the elicited weights, but only in the gain-framed tradeoff. When the most important attribute was framed as a gain, the combined effect of framing and loss aversion led to a smaller weight for that attribute. In contrast, in the loss-framed tradeoff procedure, the derived weights were not significantly affected, indicating that the two biases offset one another.

Beyond using this interaction as a bias mitigation mechanism, the study evaluated additional bias mitigation strategies across the elicitation process. The results showed that exposing decision-makers to both gain- and loss-framed tradeoff tasks and averaging the derived weights, applying homogeneous attribute frames, and aggregating judgments across individuals (group decision-making) all reduced the impact of gain-loss bias.

Taken together, these findings provided the first comprehensive evidence that loss aversion and the framing effect influenced multiple elicitation stages of MAVT, interacted in systematic ways, and jointly shaped value functions, weights, and final decision outcomes. The study advanced behavioral decision research by demonstrating how multiple cognitive biases arose within and propagated through formal MADM procedures.

RQ3: *How effective is multi-attribute value theory in reducing the status quo bias?* (Chapter 5)

While the previous research questions focused on how cognitive biases affected MAVT, this research question examined how MAVT reduced a cognitive bias. This offered a comprehensive understanding of the interaction of the decision analysis method and cognitive biases. Specifically, it investigated whether MAVT mitigated status quo bias, a pervasive strategy-based bias in behavioral decision-making (Samuelson & Zeckhauser, 1988; Montibeller & von Winterfeldt, 2015). Status quo bias arises when individuals disproportionately favor the current or default option even when superior alternatives are available. Although decision analysis methods have been argued to reduce strategy-based biases (Montibeller & von Winterfeldt, 2015; Arkes, 1991), empirical evidence in the MADM context has been limited. Assessing whether, and through which mechanisms, MAVT reduced status quo bias was therefore critical for evaluating its behavioral robustness.

To address this question, Chapter 5 distinguished between a real status quo (participants' existing situation) and an experimentally provided status quo. A systematic review of reference-dependent preferences and status quo bias informed hypotheses regarding (i) the manifestation of status quo bias in unaided rankings and choices, and (ii) the extent to which MAVT mitigated these distortions and through which elicitation stage the bias operated.

Study 4 (Chapter 5) tested two hypotheses: that framing an option as the status quo shifted preferences toward that option and nearby alternatives, and that the structured MAVT procedure mitigated this tendency. The results supported both hypotheses. Unaided rankings exhibited strong status quo effects across real and experimentally provided status quo conditions, whereas the distributions of the derived ranking after applying MAVT were substantially more similar across groups. Further analyses showed that status quo bias emerged through weight elicitation rather than attribute-specific value function elicitation, clarifying the mechanism through which MAVT reduced the bias.

Taken together, these findings provided the first comprehensive empirical evidence that MAVT meaningfully reduced status quo bias in MADM, while clarifying how the bias entered the elicitation process. The results advanced behavioral decision research by demonstrating that a strategy-based bias can be mitigated through formal elicitation procedures, though not entirely eliminated.

These findings collectively addressed the main research question.

Main RQ: *How do anchoring bias, loss aversion, framing effect, and status quo bias affect the decision-making process and outcomes within multi-attribute value theory, and how can these influences be mitigated?*

This research question examined both the influence of cognitive biases on MAVT and the extent to which such influences can be mitigated. Across the four empirical studies, the dissertation demonstrated that the four biases arose from distinct psychological mechanisms, operated at different stages of MAVT, and produced systematic distortions in elicited judgments and decision outcomes. Anchoring bias entered MAVT through explicit or implicit starting points embedded in elicitation procedures and was shown to distort both attribute-specific value functions and elicited weights. Loss aversion emerged in the tradeoff procedure and led to asymmetric tradeoff judgments, thereby altering the resulting weights. The framing effect was introduced through the way attributes are described and affected the elicited value functions by modifying

perceived marginal value across the attribute range. Moreover, loss aversion and the framing effect jointly influenced MAVT: framing-induced differences in value functions were carried forward to the tradeoff procedure, amplifying distortions in the gain-framed condition but cancelling out in the loss-framed condition. Status quo bias operated differently from the other biases: it did not influence attribute-specific value functions but affected weight elicitation by shifting the perceived relative importance of attributes, whether the status quo was self-owned or experimentally-provided.

Because MAVT aggregated the attribute-specific value functions and weights into overall evaluations, distortions at either stage can propagate to the decision outcome. Anchoring bias, loss aversion, and the framing effect all produced systematic changes in final rankings and choices, demonstrating that biases introduced at intermediate stages aggregate meaningfully influence final decisions. In contrast, status quo bias was substantially reduced when applying MAVT, and no longer affected the rankings or choices among the main alternatives.

At the same time, this dissertation identified several bias mitigation strategies that effectively mitigate these influences at both intermediate stages and final outcomes. (i) Direct procedural debiasing: no-anchor and counter-anchoring strategies embedded in the midvalue splitting procedure substantially reduced anchoring bias by either removing the initial anchor or introducing opposing anchors to cancel out their influence. This type of debiasing operates directly on the inputs (human judgments) to the MADM method and mitigates bias at the point of elicitation. Unlike general debiasing approaches, such as informing decision-makers about cognitive biases, this strategy is grounded in the identification of the underlying bias mechanisms and actively leverages the structure of the MADM procedure itself. As such, it highlights a key advantage of procedural design in mitigating cognitive biases within MADM methods. (ii) Structural bias mitigation through method design: the BWT method mitigated anchoring in the tradeoff procedure by symmetrically incorporating both attributes when formulating tradeoffs and by adopting a different computational structure for deriving weights. The bias mitigation strategies tested for gain–loss bias likewise fall under this category: exposing decision-makers to both gain- and loss-framed tradeoff tasks, applying homogeneous attribute frames, and aggregating judgments across individuals all reduce bias by altering the structure of the elicitation process rather than directly modifying individual judgments. Moreover, MAVT itself is effective in reducing status quo bias in the outcome due to its decomposed elicitation procedures. Collectively, this type of bias mitigation strategy does not intervene at the level of individual inputs, but by relying on the design of the elicitation and aggregation procedures to mitigate bias within the method, resulting in less biased outputs (decision outcomes). (iii) Counteracting bias interaction: the interaction between loss aversion and the framing effect demonstrated that biases introduced at different stages can offset each other, producing less distorted weights under certain configurations. The last type of bias mitigation strategy highlights the importance of studying multiple cognitive biases within MADM methods.

Together, these findings provide a comprehensive answer to the main research question: cognitive biases can enter MAVT at multiple stages and significantly influence elicited judgments in the decision-making process and decision outcomes, but these effects can be mitigated through carefully designed procedures, method structure, and bias interaction. Besides, MAVT itself is effective in reducing strategy-based bias. The dissertation thus advances our understanding of the behavioral robustness of MAVT and offers prescriptive guidance for designing elicitation procedures that are less vulnerable to cognitive biases.

6.2 Theoretical and Practical Implications

This dissertation offers several theoretical contributions to the literature on behavioral decision research (BDR) and multi-attribute decision-making (MADM).

First, the dissertation advances theoretical understanding of how cognitive biases enter and influence MAVT. The findings demonstrate that MAVT is not behaviorally neutral: elicited judgments in the intermediate steps and resulting decision outcomes can systematically deviate from normative expectations when cognitive biases arise at different points of the elicitation process. The four biases originate from different sources, such as the decision-maker's current status, the framing of the decision problem, or the structural features of the elicitation method, and influence MAVT in accordance with its underlying psychological mechanism. This differentiation clarifies that understanding cognitive bias in MADM requires explicit attention to where within a method each bias is likely to emerge and how it manifests in the elicitation process. This dissertation thus provides a conceptual foundation for studying cognitive biases in other MADM methods: identifying the mechanism behind a bias is essential for determining at which stage it will appear and how it will distort judgments.

Second, the dissertation contributes to BDR by demonstrating the theoretical value of process-level investigation. Much of the prior literature on bias mechanisms has focused on binary choices or simple judgments (Merkle, 2009; Fischer & Budescu, 2005; Bar-Hillel et al., 2014) to isolate direct effects, while field studies often examine bias only in final outcomes (Cooper & Meterko, 2019; Saposnik et al., 2016; Bellé et al., 2018). However, addressing MADM problems within MADM methods involves multiple sequential judgments, each of which may activate different behavioral mechanisms and propagate distortions to later stages. By empirically tracing how biases affect attribute-specific value functions, weight elicitation, and ultimately decision outcomes, this dissertation expands the theoretical perspective from "bias in outcomes" to "bias in processes". This process-level analysis enables us to understand why and how biases occur at the decision outcome. For example, Study 4 (Chapter 5) showed that status quo bias influenced elicited tradeoff weights due to direct alternative comparisons, while the decomposed structure of MAVT reduced its impact on final decision outcomes. Without a process-level analysis, such distinctions would remain hidden. Weight elicitation remains a core component of many MADM methods, and procedures such as the tradeoff method are widely used both within MAVT and in combination with other MADM approaches. It therefore remains unclear whether biased weights elicited through such procedures can be mitigated at the decision-outcome level when applied in other MADM methods. This uncertainty underscores that mitigating bias at the outcome level alone is insufficient; identifying and addressing the origin of bias within the elicitation process is essential.

Moreover, while Montibeller & von Winterfeldt (2015) define cognitive bias in prescriptive BDR as a deviation from the correct answer in a judgmental task, given by normative rules, prescriptive BDR has therefore focused on whether the elicited judgments are biased. The process-level investigation of this dissertation shows that the procedures (based on the normative axioms of MAVT) used to elicit the judgments may themselves activate behavioral mechanisms that introduce bias. This insight refines the theoretical understanding of bias in prescriptive decision analysis by locating bias not only in judgments or outcomes, but also in the methods used to elicit them.

Finally, the dissertation contributes to prescriptive BDR by demonstrating how process-level insights enable targeted bias mitigation through method design. As noted by Fasolo et al. (2024), bias mitigation strategies in decision making can be divided into debiasing, choice architecture, and a dual approach that integrates both. Rather than relying on generic debiasing approaches, such as warning decision-makers about biases or educating them on how to avoid biases (Larrick, 2004), the dissertation develops targeted bias mitigation strategies that leverage the structured elicitation procedures of MAVT. The studies show that: (i) elicitation procedures can be redesigned to directly incorporate debiasing thinking strategies (e.g., no-anchor and counter-anchoring within the midvalue splitting procedure); (ii) the structure of a method can be modified to reduce susceptibility to bias (e.g., the Best–Worst Tradeoff method reduces anchoring in traditional tradeoff elicitation); and (iii) interactions between biases across stages can naturally cancel one another, revealing a previously underexplored form of structural bias mitigation. These findings advance prescriptive theory by demonstrating that bias mitigation can be achieved not only through explicit corrective measures directed at the decision-maker but also through deliberate design of the decision analysis method. This shifts the focus from relying on individuals to consciously counteract their own biases to leveraging the structure of the elicitation procedure to mitigate biases in a seamless and unobtrusive way. This expands the theoretical landscape of prescriptive BDR by showing how structured elicitation can serve dual roles: it can be a potential entry point for bias, yet, when designed appropriately, it can also function as an inherent mechanism for mitigating that bias.

Beyond theoretical contributions, the findings of this dissertation provide actionable guidance for analysts, facilitators, and practitioners who apply MAVT and related methods in real-world decision-making contexts. Because MAVT is widely used in public policy, environmental planning, healthcare, and engineering design (Hostmann et al., 2005; Ferretti, 2016; Schuwirth et al., 2012; Khan et al., 2022), understanding where biases emerge and how they can be mitigated is essential for improving the quality and credibility of decision support.

First, the dissertation clarifies where in the MAVT process analysts should be attentive to cognitive biases. Anchoring tends to occur when elicitation procedures introduce numerical starting points or sequential comparisons, as in midvalue splitting and tradeoff elicitation. Loss aversion and framing effects are likely whenever attributes can be described in gain or loss terms, particularly in environmental or energy-related applications where advantages and disadvantages are salient. Analysts should ensure homogeneous attribute framing along the MAVT process. Status quo bias should be anticipated in consumer choice, policy evaluations, or contexts where existing alternatives serve as implicit benchmarks. By identifying these vulnerability points, the dissertation provides practitioners with a diagnostic map of where bias is most likely to distort elicited inputs.

Second, the research offers concrete bias mitigation strategies that can be integrated directly into elicitation procedures or decision-support tools. No-anchor and counter-anchor are developed within the midvalue splitting procedure to help reduce anchoring in value function construction. The BWT method serves as an alternative to the traditional tradeoff procedure to reduce anchoring bias. Homogeneous attribute framing reduces unintended asymmetries in value functions, while dual-frame elicitation mitigates gain–loss distortions in tradeoff judgments. Approaches such as group aggregation or averaging across frames provide additional ways for bias mitigation. Group decision-making can be effective under many conditions. For example, individuals may complete elicitation tasks under different anchors, attribute frames,

or naturally different status quo conditions, after which aggregation mechanisms can be used to offset opposing bias effects. Compared with approaches that debias individual judgments, such aggregation-based strategies impose lower cognitive demands on decision-makers. However, it is important to note that group decision-making is itself subject to different biases, such as false consensus (Montibeller & von Winterfeldt, 2017; Jones & Roelofsma, 2000). These strategies are straightforward to implement in software or facilitated sessions and can substantially improve the robustness of elicited preferences.

Third, the goal of normative decision analysis methods is to arrive at the optimal decision following designated decision procedures. However, BDR has identified a large number of cognitive biases that can affect the decision analysis process and lead to systematically distorted outcomes. While Montibeller & von Winterfeldt (2015) define cognitive bias in prescriptive BDR as a deviation from the correct answer in a judgmental task, as determined by normative rules, this definition does not imply that the objective of prescriptive BDR is to eliminate all biases. Rather, it provides a benchmark against which the presence and impact of biases can be assessed. In practical decision-making contexts, biases can arise at different stages of the process, and multiple biases may co-occur and interact. Designing interventions for all potential biases can increase procedural complexity and cognitive load, potentially degrading decision quality and reducing the usability of the method. This suggests that bias mitigation itself constitutes a multi-criteria problem, requiring tradeoffs between bias mitigation, cognitive effort, and procedural transparency. From this perspective, producing entirely bias-free decisions may neither be feasible nor desirable in practice. Instead, the goal of prescriptive BDR is better understood as designing decision processes that are robust to biases, in the sense that they limit the impact of systematic distortions while remaining usable and cognitively manageable for decision-makers. In this context, prioritizing the mitigation of the most influential or decision-relevant biases may be more effective than attempting to eliminate all possible biases. This shifts the focus from eliminating individual biases to structuring elicitation procedures in a way that balances normative accuracy with practical applicability.

Finally, the dissertation underscores implications for training decision analysts and facilitators. Effective application of MAVT requires more than procedural familiarity; analysts must be able to recognize conditions that may elicit bias, interpret whether value functions or weights have been distorted, and select appropriate mitigation strategies. The findings therefore support the development of behaviorally informed training programs that equip analysts with the conceptual and practical tools needed to manage cognitive biases within structured decision-making processes.

6.3 Limitations and Future Research

This dissertation provides empirical and theoretical insights into how cognitive biases influence MAVT. Nonetheless, several limitations create opportunities for further research.

The empirical studies focused primarily on the attribute-specific value function and weight elicitation stages of MAVT. While these stages are central to preference construction, earlier phases, such as defining the decision context, identifying fundamental objectives, selecting attributes, and generating alternatives, also involve subjective judgments that may be susceptible to cognitive biases. Biases such as myopic problem framing, overconfidence, confirmation

bias, or attribute omission may influence which dimensions of a decision problem are ultimately modeled. Future research should investigate how these earlier steps shape the structure of the MAVT model and whether targeted interventions can improve problem structuring as well as preference elicitation.

All four studies employed decision problems involving two or three attributes, which enabled the precise identification and isolation of the specific effects of cognitive biases. However, real-world decisions often include a larger number of attributes, complex interdependencies, and non-linear preferences. As the dimensionality of a decision increases, the potential for cognitive overload and heuristic usage increases as well. Future research should examine whether the patterns identified in this dissertation generalize to higher-dimensional decision problems and whether the magnitude of certain biases increases or diminishes as complexity grows.

All empirical studies were conducted through online experiments. In-person laboratory studies and field studies can provide richer insight into how cognitive biases manifest in practice by capturing additional contextual factors that are difficult to replicate online, such as social interaction, accountability, time pressure, emotional engagement, and organizational constraints. These settings might also affect both the emergence and mitigation of biases. Future research should replicate and extend these findings in controlled laboratory environments and real-world applications to strengthen ecological validity and practical relevance.

This dissertation focused on the MAVT, but many other MADM methods may also be vulnerable to similar distortions. For example, methods based on direct rating may be sensitive to anchoring and scale effects introduced by initial reference points or attribute ordering. Pairwise-comparison methods such as AHP require repeated direct comparisons between alternatives or criteria and may therefore be susceptible to status quo bias or reference dependence, particularly when one alternative serves as an implicit benchmark. Outranking methods, which often involve threshold setting and qualitative judgments, may also be affected by loss aversion or asymmetric sensitivity to gains and losses when defining indifference or veto thresholds. While MADM methods were developed with normative considerations, few take behavioral robustness in mind; they have rarely been evaluated against systematic behavioral factors. Future work should revisit these methods on cognitive bias, identify vulnerability points in their elicitation processes.

The dissertation examined four biases, chosen because of their theoretical relevance and potential to affect multiple stages of MAVT. However, the broader BDR literature identifies many additional biases that may also shape judgments within decision analysis. Future research should extend the empirical analysis to a wider range of cognitive and motivational biases, particularly investigating whether MADM methods can reduce other strategy-based biases or whether their bias mitigation capacity is more limited.

A notable contribution of this dissertation is the demonstration that biases can interact across elicitation stages, sometimes amplifying and sometimes offsetting one another. Similarly, when mitigating biases in the even swaps method, Lahtinen & Hämäläinen (2016) focused on designing the path of the process so that the effects of biases cancel out and do not accumulate in favor of any single alternative. In contrast, current behavioral research mostly focuses on the effects of individual biases. Future research should examine interactions more systematically, mapping how combinations of biases propagate through complex decision procedures and identifying structural conditions under which natural bias mitigation effects can occur. A

more complete understanding of such interactions may inform the design of robust elicitation frameworks that intentionally exploit cancellation effects.

This dissertation proposed several bias mitigation strategies tailored to specific biases and elicitation steps. However, these represent only an initial set of interventions. Additional work is required to design and test bias mitigation techniques that can be systematically embedded into MADM methods, including adaptive elicitation procedures, dynamic feedback mechanisms, and decision-support systems that detect and respond to bias in real time. Developing a comprehensive catalog of prescriptive tools would further advance the integration of behavioral insights into decision analysis.

Bibliography

Arkes, H. R. (1991) Costs and benefits of judgment errors: Implications for debiasing, *Psychological Bulletin*, 110(3), pp. 486–498.

Bar-Hillel, M., E. Peer, A. Acquisti (2014) "heads or tails?"—a reachability bias in binary choice, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), pp. 1656–1663.

Beinat, E. (1997) *Value functions for environmental management*, Springer, Dordrecht.

Bellé, N., P. Cantarelli, P. Belardinelli (2018) Prospect theory goes public: Experimental evidence on cognitive biases in public policy and management decisions, *Public Administration Review*, 78(6), pp. 828–840.

Cooper, G. S., V. Meterko (2019) Cognitive bias research in forensic science: A systematic review, *Forensic Science International*, 297, pp. 35–46.

Fasolo, B., C. Heard, I. Scopelliti (2024) Mitigating cognitive bias to improve organizational decisions: An integrative review, framework, and research agenda, *Journal of Management*, 51(6), pp. 2182–2211.

Ferretti, V. (2016) From stakeholders analysis to cognitive mapping and multi-attribute value theory: An integrated approach for policy support, *European Journal of Operational Research*, 253(2), pp. 524–541.

Fischer, I., D. V. Budescu (2005) When do those who know more also know more about how much they know? the development of confidence and performance in categorical decision tasks, *Organizational Behavior and Human Decision Processes*, 98(1), pp. 39–53.

Fishburn, P. C. (1967) Methods of estimating additive utilities, *Management Science*, 13(7), pp. 435–453.

Furnham, A., H. C. Boo (2011) A literature review of the anchoring effect, *The Journal of Socio-Economics*, 40(1), pp. 35–42.

Hostmann, M., T. Bernauer, H.-J. Mosler, P. Reichert, B. Truffer (2005) Multi-attribute value theory as a framework for conflict resolution in river rehabilitation, *Journal of Multi-Criteria Decision Analysis*, 13(2-3), pp. 91–102.

Jones, P. E., P. H. Roelofsma (2000) The potential for social contextual and group biases in team decision-making: Biases, conditions and psychological mechanisms, *Ergonomics*, 43(8), pp. 1129–1152.

Kahneman, D., A. Tversky, et al. (1979) Prospect theory: An analysis of decision under risk, *Econometrica*, 47(2), pp. 363–391.

Keeney, R. L., H. Raiffa (1993) *Decisions with multiple objectives: preferences and value trade-offs*, Cambridge University Press, Cambridge.

Khan, I., L. Pintelon, H. Martin (2022) The application of multicriteria decision analysis methods in health care: a literature review, *Medical Decision Making*, 42(2), pp. 262–274.

Lahtinen, T. J., R. P. Hämäläinen (2016) Path dependence and biases in the even swaps decision analysis method, *European Journal of Operational Research*, 249(3), pp. 890–898.

Larrick, R. P. (2004) Debiasing, in: *Blackwell Handbook of Judgment and Decision Making*, Blackwell Publishing, Malden, pp. 316–338.

Liang, F., M. Brunelli, J. Rezaei (2022) Best-worst tradeoff method, *Information Sciences*, 610, pp. 957–976.

Merkle, E. C. (2009) The disutility of the hard-easy effect in choice confidence, *Psychonomic Bulletin & Review*, 16(1), pp. 204–213.

Montibeller, G., D. von Winterfeldt (2015) Cognitive and motivational biases in decision and risk analysis, *Risk Analysis*, 35(7), pp. 1230–1251.

Montibeller, G., D. von Winterfeldt (2017) Individual and group biases in value and uncertainty judgments, in: *Elicitation: The science and art of structuring judgement*, Springer, pp. 377–392.

Samuelson, W., R. Zeckhauser (1988) Status quo bias in decision making, *Journal of Risk and Uncertainty*, 1(1), pp. 7–59.

Saposnik, G., D. Redelmeier, C. C. Ruff, P. N. Tobler (2016) Cognitive biases associated with medical decisions: a systematic review, *BMC Medical Informatics and Decision Making*, 16(1), p. 138.

Schuwirth, N., P. Reichert, J. Lienert (2012) Methodological aspects of multi-criteria decision analysis for policy support: A case study on pharmaceutical removal from hospital wastewater, *European Journal of Operational Research*, 220(2), pp. 472–483.

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

Tversky, A., D. Kahneman (1981) The framing of decisions and the psychology of choice, *Science*, 211(4481), pp. 453–458.

Tversky, A., D. Kahneman (1991) Loss aversion in risaess choice: A reference-dependent model, *The Quarterly Journal of Economics*, 106(4), pp. 1039–1061.

Appendix

A Appendix - Questionnaire design for Chapter 2

The full questionnaire consisted of the following sections: informed consent, presentation of the decision problem, verification of the additivity assumption, weight elicitation, and value function elicitation. In this appendix, we present the part of the questionnaire related to midvalue splitting procedure for participants in the high anchor condition of the first scenario.

Q44: Suppose you can get a lower rent for the apartment by increasing the commute distance. Suppose the drop in monthly rent would be either from 1500 euros to 1340 euros or from 1340 euros to 700 euros. For which drop in price would you accept a larger increase in commute distance?

- ☐ I would accept an equal increase in the commute distance.
- ☐ For the drop from 1500 to 1340 euros I would accept a larger increase in commute distance.
- ☐ For the drop from 1340 to 700 euros I would accept a larger increase in commute distance.

Q45: Suppose you can get a lower rent for the apartment by increasing the commute distance. Suppose the drop in monthly rent would be either from 1500 euros to 1180 euros or from 1180 euros to 700 euros. For which drop in price would you accept a larger increase in commute distance?

- ☐ I would accept an equal increase in the commute distance.
- ☐ For the drop from 1500 to 1180 euros I would accept a larger increase in commute distance.
- ☐ For the drop from 1180 to 700 euros I would accept a larger increase in commute distance.

Q46: Suppose the price drop in monthly rent would be from 1500 to a certain value (R1) or from that same value (R1) to 700. Please assign a value to R1 such that you would accept the same increase in commute distance.

R1= ☐

Q47: Suppose the price drop in monthly rent would be from 1500 to a certain value (R2) or from that same value (R2) to [Piped response from Q46]. Please assign a value to R2 such that you would accept the same increase in commute distance.

R2= ☐

Q48: Suppose the price drop in monthly rent would be from [Piped response from Q46] to a certain value (R1) or from that same value (R1) to 700. Please assign a value to R1 such that

you would accept the same increase in commute distance.

R3= ☐

Q49: Would you increase the same commute distance in the two rent reductions, from [Piped response from Q47] euros to [Piped response from Q46] euros and from [Piped response from Q46 euros to [Piped response from Q48] euros?

☐ Yes

☐ No

Q50: Suppose the price drop in monthly rent would be from [Piped response from Q47] to a certain value (R4) or from that same value (R4) to [Piped response from Q48]. Please assign a value to R4 such that you would accept the same increase in commute distance.

R4= ☐

Q51: Suppose you could pay an increase in rent to reduce the commute distance, would you pay more to reduce the commute distance from 20 km to 17 km, or from 17 km to 5 km?

☐ I am willing to increase more rent to reduce the commute distance from 20 km to 17 km.

☐ I am willing to increase more rent to reduce the commute distance from 17 km to 5 km.

☐ I would increase the same amount on rent for the two commute distance reductions.

Q52: Suppose you could pay an increase in rent to reduce the commute distance, would you pay more to reduce the commute distance from 20 km to 14 km, or from 14 km to 5 km?

☐ I am willing to increase more rent to reduce the commute distance from 20 km to 14 km.

☐ I am willing to increase more rent to reduce the commute distance from 14 km to 5 km.

☐ I would increase the same amount on rent for the two commute distance reductions.

Q53: Suppose the drop in commute distance would be from 20 to a certain value (D1) or from that same value (D1) to 5. Please assign a value to D1 such that you would accept the same increase in rent.

D1= ☐

Q54: Suppose the drop in commute distance would be from 20 to a certain value (D2) or from that same value (D2) to [Piped response from Q53]. Please assign a value to D2 such that you

would accept the same increase in rent.

D2= [ ]

Q55: Suppose the drop in commute distance would be from [Piped response from Q53] to a certain value (D3) or from that same value (D3) to 5. Please assign a value to D3 such that you would accept the same increase in rent.

D3= [ ]

Q56: Would you increase the same on rent to reduce the commute distance from [Piped response from Q54] km to [Piped response from Q53] km and from [Piped response from Q53] km to [Piped response from Q55] km.

- ☐ Yes
- ☐ No

Q57: Suppose the drop in commute distance would be from [Piped response from Q54] to a certain value (D4) or from that same value (D4) to [Piped response from Q55]. Please assign a value to D4 such that you would accept the same increase in rent.

D4= [ ]

About the author

Geqie Sun was born in 1998 in Chongqing, China. She completed her undergraduate studies in Logistics Management at Beijing Jiaotong University in 2020, where she developed a strong interest in decision-making, operations, and quantitative analysis. She subsequently obtained her master's degree in Industrial Engineering and Logistics Management from The University of Hong Kong in 2021, further strengthening her analytical background and research skills.

In December 2021, she began her PhD at Delft University of Technology in the Netherlands. Her doctoral research lies at the intersection of behavioral decision-making and multi-criteria decision analysis. In particular, her work investigates how cognitive biases influence preference elicitation processes in multi-attribute value theory and related decision analysis methods, and how structured elicitation procedures can be designed to mitigate such biases.

Publications from this dissertation

1. **Sun, G.**, Kroesen, M., Rezaei, J (2025) Anchoring bias in value function elicitation within multiattribute value theory, *Decision Analysis*, 22(4), pp. 284-304.
2. **Sun, G.**, Kroesen, M., Rezaei, J (2026) Anchoring bias in the tradeoff procedure within multi-attribute value theory, *Journal of Behavioral Decision Making*, 39 (2), e70069
3. **Sun, G.**, Kroesen, M., Rezaei, J (Under review) Framing and loss aversion in decisions with multiple objectives, Manuscript submitted to *Management Science.*
4. **Sun, G.**, Kroesen, M., Rezaei, J (Under review) The mitigation role of multi-attribute value theory on status quo bias, Manuscript submitted to *Journal of Behavioral Decision Making.*

Selected conference presentations

1. **Sun, G.**, Kroesen, M., Rezaei, J (2025, October) *Framing and loss aversion in decisions with multiple objectives*. 2025 INFORMS Annual Meeting, Atlanta, USA.
2. **Sun, G.**, Kroesen, M., Rezaei, J (2025, June) *Anchoring bias in the tradeoff procedure within multi-attribute value theory*. 34th European Conference on Operational Research (EURO 2025), Leeds, United Kingdom.
3. **Sun, G.**, Kroesen, M., Rezaei, J (2024, July) *Anchoring bias in value function elicitation within multi-attribute value theory*. 33rd European Conference on Operational Research (EURO 2024), Copenhagen, Denmark.